Die Anwendung statistischer Record-Linkage-Methoden auf selbst-generierte Codes bei Längsschnitterhebungen

von Rainer Schnell, Tobias Bachteler und Jörg Reiher [1]

Zusammenfassung

Bei wiederholten Befragungen zu sensitiven Themen werden häufig von den Befragten selbst-generierte Codes zur Zusammenführung der Daten über die Wellen hinweg verwendet. Dabei ist aufgrund fehlerhafter Daten der Verlust an Fällen in der Regel beträchtlich. Um den Prozess der Zusammenführung zu beschleunigen und die Zahl verlorener Fälle zu minimieren wird hier die Nutzung automatisierter Record-Linkage-Techniken vorgeschlagen. Bei zwei Simulationsstudien und in einer praktischen Anwendung zeigen sich Techniken des Record-Linkage der bisher verwendeten manuellen Methode überlegen. Die Experimente legen bei Verwendung dieser Techniken den Einsatz deutlich längerer Codes als bisher üblich nahe.

Abstract

Panel studies on sensitive topics usually apply respondent generated codes to link records across surveys. This implies a substantial loss of cases due to errors in the codes. In order to speed up the process of linking and minimizing the number of losses due to errors, we propose the application of automatic record-linkage procedures. In two simulations and a real-world application, the record-linkage procedures outperformed the manual procedure. The experiments suggest the use of longer self-generated codes for record-linkage applications.

1 Dr. ***Rainer Schnell*** ist Professor im Zentrum für Quantitative Methoden und Surveyforschung, Universität Konstanz, Postfach D92, 78434 Konstanz, E-Mail: rainer.schnell@uni-konstanz.de. ***Tobias Bachteler*** (M.A.) ist wissenschaftlicher Angestellter im Zentrum für Quantitative Methoden und Surveyforschung, Universität Konstanz. ***Jörg Reiher*** studiert Informatik an der Fernuniversität Hagen.

1 Einleitung

Bei wiederholten Befragungen steht man vor dem Problem, für eine sinnvolle Analyse die Antworten der gleichen Person einander zuzuordnen.[2] Bei schriftlichen Befragungen in institutionellen Kontexten versucht man die Angabe des Namens oder administrative verwendete Identifikationsnummern (Wehrpassnummer, Mitarbeiternummer, Matrikelnummer) zu vermeiden. Insbesondere bei Teilpopulationen, die für Datenschutzprobleme besonders sensibel sind (Studenten, Patienten, Angehörige devianter Subkulturen), werden anstelle der administrativen Codes häufig so genannte „selbst-generierte Codes“ verwendet. Dies sind Codes, deren Bestandteile von den Befragten selbst aus Merkmalen wie dem Vornamen der Mutter usw. hergeleitet werden. Das Problem dieser Codes besteht darin, dass Fehler bei der Erzeugung der Codes durch die Befragten dazu führen, dass nur ein Teil der vorhandenen Fragebögen über die Wellen hinweg zusammengeführt werden kann. Einerseits ist der Verlust vorhandener Daten immer unerwünscht. Andererseits besteht die Möglichkeit, dass Personen, deren Daten erfolgreich zusammengeführt werden können, sich von denjenigen Personen, deren Daten nicht zusammengeführt werden konnten, systematisch unterscheiden. Um die Zahl nicht zusammenführbarer Fragebögen zu verringern, wird hier die Anwendung zweier bislang nicht für diesen Zweck verwendeter fehlertoleranter maschineller Techniken vorgeschlagen.

2 Selbst-generierte Codes in Längsschnittstudien

Selbst-generierte Codes bestehen in der Regel aus Namensbestandteilen des eigenen Namens oder der Namen naher Verwandter sowie Teilen des Geburtsdatums oder Geburtsortes (***Kearney*** et al. 1984: 373). Die resultierenden Codes haben fast immer eine Länge von 6-10 Zeichen. In der Praxis erfolgt die Zuordnung zunächst anhand der exakt übereinstimmenden Codes, anschließend wird manuell eine Zuordnung derjenigen Fälle vorgenommen, bei denen ein oder zwei Zeichen nicht übereinstimmen (***Pöge*** 2005: 61). Typischerweise verliert man bei dieser Vorgehensweise auch bei einer konstanten Population je nach Codemerkmal, Bildung und Abstand zwischen den Datenerhebungen einen beträchtlichen Teil der Fälle, wobei Anteile von 25-40 % nicht ungewöhnlich sind (***Pöge*** 2005: 65-66). Anstelle dieser semimanuellen Vorgehensweise schlagen wir die Anwendung der statistischen Methoden des Record-Linkage für das Problem der Zusammenführung selbst-erzeugter

2 Bei Face-to-Face-Surveys ist diese Zuordnung über den Namen und die Anschrift möglich, bei telefonischen Befragungen über die Telefonnummer und eine Kombination aus Alter, Geschlecht und Vorname; bei postalischen Erhebungen verwendet man nahezu immer eine sichtbare und dem Befragten erläuterte laufende Nummer auf dem Fragebogen.

Codes vor. Interessanterweise ist die Anwendung des probabilistischen Record-Linkage auf selbst-generierte Codes bislang in der Literatur nicht vorgeschlagen worden.

3 Statistische Methoden für die Zuordnung selbst-generierter Codes

Seit Jahrzehnten wird in der amtlichen Statistik die Zusammenführung von Datenbeständen über identische Merkmalsträger anhand ihrer Namen und anderer Identifizierungsmerkmale als Record-Linkage bezeichnet. In Anlehnung an die Literatur zum statistischen Record-Linkage[3] kann man zwischen distanz-basierten und probabilistischen Record-Linkage-Techniken unterscheiden. Beide Techniken kommen als Methoden zur Zusammenführung selbst-generierter Codes in Betracht.

Bei einem **distanz-basierten Record-Linkage** werden Stringdistanzen zwischen den Codes berechnet. Für einen gegebenen Code aus einer ersten Befragungswelle wird dann derjenige Fall aus der zweiten Welle gesucht, für den die Distanz zwischen den Codes minimal ist. Eine für ähnliche Problemstellungen oft verwendete Stringdistanzfunktion ist die so genannte „***Damerau-Levenshtein***"-Distanz. Die Grundidee hierbei ist, dass sich jede Zeichenkette *a* durch eine gewisse Zahl der Operationen „Ersetzen eines Zeichens", „Einfügen eines Zeichens", „Löschen eines Zeichens" oder dem „Vertauschen zweier Zeichen" in jede andere Zeichenkette *b* überführen lässt.[4] Zum Beispiel ließe sich der Code „AM02GF" durch das Ersetzen des Zeichens A mit B und durch das Entfernen des Zeichens 0 in den Code „BM2GF" überführen. Als ***Damerau-Levenshtein***-Distanz wird die Anzahl der zu einer solchen Überführung notwendigen Operationen bezeichnet, im Beispiel beträgt sie 2.

Zerlegt man selbst-erzeugte Codes in ihre Bestandteile, so können diese als Identifizierungsmerkmale für probabilistische Methoden dienen. Bei einem **probabilistischen Record-Linkage** hängt die Klassifizierung eines gegebenen Datenzeilen-Paares als zusammengehörig oder nicht zusammengehörig vom Vergleich der Ausprägungen der Identifizierungsmerkmale ab. Dabei können die Werte eines Merkmals übereinstimmen oder sie stimmen nicht überein. In beiden Fällen wird ermittelt, wie wahrscheinlich das aktuelle Vergleichsergebnis für tatsächlich zusammengehörige Datenzeilen im Verhältnis zu in Wahrheit nicht zusammengehörigen

3 Einführungen finden sich in ***Winkler*** (1995) und ***Judson*** (2004). Aktuelle Übersichten über den Stand der Forschung geben ***Elmagarmid*** et al. (2006) und ***Winkler*** (2006).

4 Eine formale Definition der ***Damerau-Levenshtein***-Distanz findet sich in ***Ukkonen*** (1985).

Datenzeilen ist. Eine Übereinstimmung in einem Merkmal wird nun umso stärker für eine korrekte Zuordnung sprechen, je seltener sie unter den nicht zusammengehörigen im Vergleich zu den zusammengehörigen Datenzeilen auftritt. Die Summe dieser Verhältnisse stellt für die Datenzeilenpaare das Kriterium für die Zuordnung dar: Je größer die Summe, desto eher sollten zwei Datenzeilen einander zugeordnet werden.[5]

Wir vermuteten, dass beide Techniken aufgrund ihrer Fehlertoleranz gegenüber einer automatisierten exakten Verknüpfung eine höhere Zuordnungseffektivität aufweisen werden. Daraus würde dann eine deutlich erhöhte Effizienz gegenüber dem bisherigen Verfahren als Ganzem folgen, da es sich um vollständig automatisierte Verfahren handelt, die im Vergleich zu der manuellen fehlertoleranten Zuordnung eine immense Zeitersparnis erbringen können. Dabei gingen wir davon aus, dass die Überlegenheit gegenüber einem exakten Abgleich um so stärker sein wird, je größer die Wahrscheinlichkeit eines Fehlers in den selbst-generierten Codes ist.

4 Experimente

Da die exakte Verknüpfung von selbst-erzeugten Codes im bisherigen Verfahren stets den ersten und einzig automatisierten Schritt darstellt, sollte die Leistungsfähigkeit der neuen Verknüpfungsmethoden im Vergleich zu einer Zusammenführung anhand ihrer exakten Übereinstimmung bestimmt werden.

4.1 Erstes faktorielles Experiment mit simulierten Daten

Mit Hilfe eines ersten faktoriellen Experiments auf der Basis simulierter Daten sollte geprüft werden, ob und wenn ja unter welchen Bedingungen sich die von uns vermutete Überlegenheit der neuen Verknüpfungsmethoden zeigt. Dazu sollte a) die Art der Codes, b) die Länge der Codes und c) das Ausmaß der Fehler experimentell variiert werden. Die relative Leistungsfähigkeit der drei verglichenen Verknüpfungsmethoden – a) der exakten Verknüpfung, b) des distanz-basierten Record-Linkage und c) des probabilistischen Record-Linkage – sollte durch die Zahl der korrekt gefundenen Code-Paare bestimmt werden.

5 Eine knappe technische Darstellung findet sich bei ***Grannis*** et al. (2003).

4.1.1 Experimenteller Aufbau und Datenbasis

Für das erste Experiment wurde ein 3*3*3*3 Design gewählt. Zentraler Designfaktor war die verwendete Verknüpfungsmethode mit den Niveaus exakt, distanzbasiert und probabilistisch. Die exakte Verknüpfung und die Verknüpfung nach der ***Damerau-Levenshtein***-Distanz erfolgten jeweils über den kompletten Code. Für das probabilistische Record-Linkage[6] wurden die Codes jeweils in Teilcodes der Länge 2 zerlegt.[7] Diese zweistelligen Codes dienten als Verknüpfungsmerkmale für das Record-Linkage.[8] Experimentell variiert wurden daneben die Art des Codes, die Länge des Codes sowie das Ausmaß der je in den zwei Wellen vorhandenen fehlerhaften Codes. Diese Fehlerbelegung wurde für jede Zelle des Designs 30-mal wiederholt. Tabelle 1 enthält eine Übersicht der Designfaktoren mit ihren Niveaus.

Tabelle 1 Experiment 1: Designfaktoren

Experimenteller Faktor	***Ausprägungen***		
Verknüpfungsmethode	exakt	distanz-basiert	Probabilistisch
Art des Codes[9]	binär 0-1	dezimal 0-9	Buchstaben mit Umlauten und ß
Länge des Codes	6 Stellen	8 Stellen	10 Stellen
Fehlerwahrscheinlichkeit (in % fehlerhafte Zeilen) [10]	0,1 %	1 %	5 %

Je nach den drei Codearten und den drei Codelängen wurden jeweils 1 000 Codes simuliert. Diese Codes wurden dann zweimal zufällig mit Fehlern versehen, so dass ein 2-Wellen-Panel simuliert wurde. Für jede der 27 experimentellen Bedingungen (3 Codetypen, 3 Codelängen und 3 Fehlertypen) wurden 30 Replikationen erzeugt,

6 Bei dem gewählten Verfahren handelt es sich um probabilistisches Record-Linkage mit wertespezifischen Gewichten (***Newcombe*** 1988: 17-18), mit erzwungener eins-zu-eins Zuordnung (***Jaro*** 1989: 417-418) und einer Anpassung für die ***Damerau-Levenshtein***-Distanz nach dem Verfahren von ***Jaro*** (***Winkler*** 1990: 356).

7 Etwa würde der Code AM01NE12FR in die Teilcodes AM, 01, NE, 12 und FR zerlegt.

8 Alle Verknüpfungen wurden mittels des Record-Linkage Programms „Merge Toolbox (MTB) ausgeführt. Die MTB wird in ***Schnell*** et al. (2005) vorgestellt.

9 Zum Beispiel könnte „00101110“ ein binärer Code, „38129923“ ein dezimaler Code, „dläßtsgr“ ein Buchstabencode jeweils der Länge 8 sein.

10 Diese kleinen Fehlerwahrscheinlichkeiten wurden aufgrund in der von ***Kearney*** et al. (1984) berichteten hohen Erfolgsrate der Zusammenführung von 92 % bzw. von 78 % bei einem Abstand von einem Monat bzw. einem Jahr zwischen zwei Wellen gewählt.

so dass die Simulation zu 810 simulierten Panelstudien führte. Auf die Daten jedes dieser 810 Panel wurden die 3 Methoden angewandt. Die Zahl der pro simuliertem Panel korrekt zugeordneten Codes bildete das Kriterium für die Leistungsfähigkeit der Verknüpfungsmethoden.

4.1.2 Ergebnisse

Im Zuge der Analysen zeigte sich schnell, dass sich die Ergebnisse für die binäre Codeversion aufgrund vieler Duplikate in den Wellen nicht sinnvoll interpretieren lassen. Tabelle 2 zeigt die im Mittel korrekt zugeordneten Codes nach Methode und Art des Codes. Die niedrigen Werte für die binären Codes rühren dabei von vielen Duplikaten in den Codes her. Zwischen solchen Duplikaten kann hinsichtlich der Zuordnung von Codes mit keiner Methode differenziert werden. Für die weiteren Analysen wurde die binäre Codeversion daher nicht mehr berücksichtigt.

Tabelle 2 Experiment 1: Mittelwerte korrekt zugeordnete Codes nach Methode und Art des Codes

Methode	***Codeversion***		
	binär	*dezimal*	*Buchstaben*
exakt	13.3	966.3	964.4
distanz-basiert	13.3	999.1	999.9
probabilistisch	12.9	1000	1000

Tabelle 3 zeigt die Ergebnisse der Anpassung des multiplen ANOVA-Modells mit der abhängigen Variable „Zahl korrekt zugeordnete Paare“ und den Faktoren Methode, Fehlerschema und Codelänge nebst deren Interaktionen. Bei einem Determinationskoeffizienten von 0,98 erklärt der Haupteffekt für die Methode 37 % der Gesamtvarianz in den experimentellen Daten, der Faktor Fehlerschema 20,5 %. Zusammen mit der durch die Interaktion zwischen der Methode und dem Fehlerschema erklärten 39,4 % sind das bereits 96,9 % der Gesamtvarianz. Weder der Haupteffekt noch Interaktionen mit dem Faktor Codelänge erklären einen bedeutenden Anteil der Varianz.

Tabelle 3 Experiment 1: Ergebnisse ANOVA-Modell ohne binäre Codes

Quelle	***SS***	***d.f.***	MS	***F***	***Prob>F***
Modell	1129835.93	26	43455.23	3489.29	0.00
Methode	425566.85	2	212783.42	17085.68	0.00
Schema	235960.02	2	117980.01	9473.34	0.00
Codelänge	2986.26	2	1493.13	119.89	0.00
Methode*Schema	452547.96	4	113136.99	9084.46	0.00
Methode*Codelänge	4789.61	4	1197.40	96.15	0.00
Schema*Codelänge	3102.08	4	775.52	62.27	0.00
Methode*Schema*Codelänge	4883.15	8	610.39	49.01	0.00
Residual	19839.07	1593	12.45		
Total	1149675	1619	710.11	r^2	0.98

Die Mittelwerte der korrekt Zugeordneten nach Methode und Schema für die dezimale Codeversion und die Buchstabencodes sind in Tabelle 4 verzeichnet. Man sieht, dass der Effekt des Faktors Methode im Übergang von der exakten zu den beiden neuen Verfahren liegt: Die beiden neuen Techniken ordnen im Schnitt mehr Code-Paare korrekt zu. Der Effekt ist dabei durch das Fehlerschema in der Weise vermittelt, dass sich die distanz-basierte und die probabilistische Methode erst für das zweite und das dritte Schema deutlich von der exakten Methode absetzen: Bereits ab einer Fehlerwahrscheinlichkeit von 1 % der Codes zeigen sich die neuen Verfahren überlegen. Unter den experimentellen Bedingungen werden hier von der distanz-basierten Methode fast alle und von der probabilistischen Methode alle Code-Paare über die 1 620 Wellenpaare korrekt zugeordnet.

Tabelle 4 Experiment 1: Mittelwerte korrekt zugeordneter Codes nach Methode und Fehlerschema ohne binäre Codes

Methode	***Fehlerschema***		
	1	*2*	*3*
exakt	998.2	982.2	915.7
distanz-basiert	1000	1000	998.2
probabilistisch	1000	1000	1000

Aus dem ersten Experiment lässt sich als Fazit ziehen, dass sich die neuen Verknüpfungsmethoden als umso überlegener erwiesen haben, je stärker fehlerbehaftet die Codes waren, wohingegen die Länge der Codes kaum Einfluss zu haben scheint. Die Art der verwendeten Codes scheint lediglich dahingehend eine Bedeutung zu haben, dass bei rein binären Codes aufgrund der vielen Duplikate keine Methode gut abschneidet. Für die meisten Wellenpaare schneiden die neuen Methoden besser ab als die exakte Verknüpfung, in keinem Fall erreichten beide neuen Methoden weniger korrekte Zuordnungen als die exakte Methode.

4.2 Zweites faktorielles Experiment mit simulierten Daten

Das zweite faktorielle Experiment sollte Aufschluss darüber geben, ob sich die verbesserten Zuordungsraten der distanz-basierten und der probabilistischen Methode bei realistischeren Codes und höheren Fehlerwahrscheinlichkeiten als im ersten faktoriellen Experiment replizieren lassen.

4.2.1 Datenbasis

Für das Experiment wurde zunächst ein zu simulierendes Codierungsschema festgelegt. Die Codes sollten 10 Stellen aufweisen und aus den in Tabelle 5 verzeichneten Bestandteilen zusammengesetzt sein:

Tabelle 5 Experiment 2: Bestandteile der simulierten Codes

Code-Position	*Quelle*
1-2	1. und 2. Stelle des Vornamens des Vaters
3-4	Geburtstag
5-6	1. und 2. Stelle des Vornamens der Mutter
7-8	Geburtsmonat
9-10	1. und 2. Stelle des Namens der Geburtsgemeinde

Die der Simulation der Codes zugrunde liegenden Daten wurden alle realen Datenbanken entnommen. Als Basis für die Vornamen diente eine Datei, die die Vornamen von 29 Millionen Personen in der BRD enthält. Aus dieser Datei wurden jeweils 1 000 Männer und 1 000 Frauen zufällig ausgewählt. Ihre Vornamen wurden auf die ersten beiden Buchstaben abgeschnitten. Weiterhin wurden aus einer Datei mit mehr als 12 000 deutschen Gemeindenamen auf dieselbe Weise 1 000 Testzei-

len gewonnen. Als Datenbasis für die Simulation von fehlerbehafteten Angaben des Geburtstages und des Geburtsmonats wurden unter der Annahme der Gleichverteilung von Geburten 365 mögliche Kombinationen aus Geburtstag und Geburtsmonat gebildet. Jede dieser Kombinationen wurde 1 000-mal dupliziert und aus den resultierenden 365 000 Zeilen zufällig 1 000 Zeilen ausgewählt.

Die einzelnen Komponenten wurden dann zu 1 000 Kombinationen zusammengeführt. Die resultierenden 1 000 10-stelligen Codes bildeten die Grundlage für die Simulationen der fehlerbehafteten selbsterzeugten Codierungen.

4.2.2 Design

Für das Experiment wurde ein 3*3*2 Design mit jeweils 30 Replikationen der Fehlerbelegungen gewählt, so dass insgesamt 540 Datenpunkte vorlagen.

Der erste Designfaktor bezeichnet die verwendete Verknüpfungsmethode mit den Niveaus exakt, distanz-basiert und probabilistisch. Die exakte Verknüpfung und die Verknüpfung nach der ***Damerau-Levenshtein***-Distanz erfolgten über den kompletten 10-stelligen Code, das probabilistische Record-Linkage wurde mit den 5 Bestandteilen des Codes als Verknüpfungsvariablen ausgeführt.

Der zweite Design-Faktor bezieht sich auf das verwendete Schema an Fehlerwahrscheinlichkeiten, die zur Erzeugung der Daten verwendet wurden: Die 1 000 Codes wurden hierbei wiederholt mit künstlichen Fehlern versehen. Dazu wurden für jeden Code-Bestandteil in nach vorgegebenen Wahrscheinlichkeiten ausgewählten Zeilen jeweils 1 Zeichen ersetzt: Bei den Vornamen und dem Gemeindenamen durch einen zufällig ausgewählten anderen Buchstaben, bei Geburtstag und -monat durch eine zufällig ausgewählte Ziffer. Diese Fehlerwahrscheinlichkeiten wurden auf der Basis realer Fehlerquoten für die drei unterschiedlichen Schemata in folgender Weise gewählt (vgl. Tabelle 6).

Dabei sollte Schema 2 aus realistischen Fehlerwahrscheinlichkeiten bestehen. Schema 3 stellt eine extreme Variante mit hohen Fehlerquoten, Schema 1 eine eher günstige Variante mit vergleichsweise niedrigen Fehlerwahrscheinlichkeiten dar. Für jedes Schema wurden jeweils drei Verfremdungen zu einem Panel von drei Wellen zusammengefasst. Jede Welle eines Panels bestand also aus denselben Ausgangsdaten, die aber jeweils neu verfremdet wurden. Dies wurde 30-mal wiederholt, so dass für jedes Schema 30 Panels an simulierten selbst-erzeugten Codes als Datenbasis für den Vergleich der Verknüpfungsprozeduren vorlagen.

Tabelle 6 Experiment 2: Fehlerwahrscheinlichkeiten für die Schemata

Code-Bestandteil	*Fehler-schema 1*	*Fehler-schema 2*	*Fehler-schema 3*
1. und 2. Stelle Vorname Vater	3 %	4 %	5 %
1. und 2. Stelle Vorname Mutter	3 %	4 %	5 %
1. und 2. Stelle Geburtsgemeinde	4 %	5 %	6 %
Geburtstag	0,5 %	1 %	2 %
Geburtsmonat	0,5 %	1 %	2 %

Der dritte Design-Faktor besteht aus den zu verknüpfenden Wellenpaaren, also einmal Welle 1 zu Welle 2 und einmal Welle 2 zu Welle 3. Dieser Design-Faktor wurde aufgenommen, um die Effekte nicht nur über zwei, sondern auch über drei Wellen zu untersuchen. Er spielt für die Untersuchungen hier keine Rolle.

4.2.3 Ergebnisse

Tabelle 7 gibt die Ergebnisse der Berechnung des ANOVA-Modells mit der abhängigen Variable „Zahl korrekt zugeordneter Codes“ und den Faktoren Methode und Fehlerschema sowie deren Interaktion wieder. Die verwendete Verknüpfungsmethode allein erklärt 93,6 % der Gesamtvarianz in den experimentellen Daten. Die Interaktion zwischen der Methode und dem Fehlerschema mit 3,9 % und der Faktor Fehlerschema mit 2 % erklären fast die komplette Restvarianz.

Tabelle 7 Experiment 2: Ergebnisse des ANOVA-Modells

Quelle	*SS*	*d.f.*	*MS*	*F*	*Prob>F*
Modell	8211037.84	8	1026379.73	14411.51	0.00
Methode	7724019.75	2	3862009.87	54226.91	0.00
Schema	163823.35	2	81911.67	1150.13	0.00
Methode*Schema	323194.74	4	80798.68	1134.50	0.00
Residual	37817.52	531	71.22		
Total	8248855.35	539	15303.10		r^2 0.99

Tabelle 8 gibt die Mittelwerte der Zahl korrekt zugeordneter Paare von Codes für die Verknüpfungsmethoden nach dem verwendeten Fehlerschema wieder. Das Ergebnis der Varianzanalyse zeigt, dass sich mit der distanz-basierten und der probabilistischen Methode deutlich mehr Paare von Codes richtig zuordnen lassen und der Effekt für steigende Fehlerwahrscheinlichkeiten zunimmt.

Interessanterweise resultiert diese Zunahme ausschließlich aus dem Nachlassen der durch die exakte Methode gefundenen Paare, die Werte für die beiden neuen Methoden bleiben für die verschiedenen Schemata von Fehlerwahrscheinlichkeiten nahezu gleich und fast perfekt.

Tabelle 8 Experiment 2: Mittelwerte korrekt zugeordneter Codes nach Methode und Fehlerschema

Methode	***Fehlerschema***		
	1	*2*	*3*
exakt	778.7	713.7	639.2
distanz-basiert	997.6	997.3	996.9
probabilistisch	998.1	998.0	998.1

4.3 Experiment mit realen Daten

Um die Leistungsfähigkeit der neuen Methoden auch im Umgang mit realen (von Menschen erzeugten) Codes zu prüfen, wurde in einem dritten Experiment untersucht, ob sich die Überlegenheit der neuen Methoden auch bei einer solchen realen Datenbasis zeigt.

4.3.1 Datenbasis und Design

In einer Pflichtveranstaltung des zweiten Semesters einer politikwissenschaftlichen Fakultät wurde den Hörern im Abstand von einer Woche dreimal ein Lehrevaluationsfragebogen vorgelegt. Dieser Fragebogen enthielt zudem Codierungsanweisungen für einen selbst-generierten Code. Die Studenten wurden aufgefordert, nach den Anweisungen diesen Code zu bilden und auf dem Fragebogen zu vermerken.

Die Studenten sollten den Code in der folgenden Weise bilden (s. Tabelle 9):

Tabelle 9 Experiment 3: Bestandteile der realen Codes

Code-Position	*Quelle*
1	1. Buchstabe des Vornamens der Mutter
2	2. Buchstabe des Vornamens der Mutter
3	1. Stelle Geburtstag
4	2. Stelle Geburtstag
5	1. Buchstabe des Vornamens des Vaters
6	2. Buchstabe des Vornamens des Vaters

Den Studenten wurde der Experimentcharakter dieser Codierung mitgeteilt. Unter Betonung der besonderen Vertraulichkeit der Angaben und der vollständigen Konsequenzenlosigkeit jedweder Antwort wurden die Studenten zusätzlich um die Angaben ihrer Matrikelnummer gebeten.

Insgesamt gaben 152 Studenten in mindestens einer Vorlesung eine valide Matrikelnummer und mindestens einen Code an.[11] Für einen Test der Tauglichkeit der selbsterzeugten Codes sind nur solche interessant, die mindestens 2-mal gebildet wurden.

Für die 140 aus der ersten Vorlesung vorliegenden Codes existieren für 103 auch Codes aus der zweiten Vorlesung und für 105 auch Codes aus der dritten Vorlesung. Für die 111 aus der zweiten Vorlesung vorliegenden Codes existieren für 92 auch Codes aus der dritten Vorlesung. Von 87 Studenten liegen Codes aus allen drei Veranstaltungen vor.

Anhand der Matrikelnummern konnte nun untersucht werden, inwieweit anhand der selbsterzeugten Codes eine Zuordnung der Studenten zwischen den Veranstaltungen durch die drei zu vergleichenden Methoden möglich ist.

4.3.2 Ergebnisse

Tabelle 10 zeigt die Zahl der korrekt zugeordneten Code-Paare nach betrachteten Vorlesungen und Verknüpfungsmethode.

11 Bei zwei unterschiedlichen Studenten ergab sich der gleiche Code, so dass eine der beiden betreffenden Datenzeilen aus dem Datenfile entfernt wurde.

Tabelle 10 Experiment 3: Korrekt zugeordnete Codes nach Vorlesungen und Methode

Vorlesungen	***Methode***		
	exakt	*distanz-basiert*	*probabilistisch*
1 und 2	93	103	103
2 und 3	82	92	92
1 und 3	99	105	105
1, 2 und 3	76	87	87

Während bei exakter Zuordnung Verluste zwischen zwei Wellen von bis zu 10 %, über alle drei Wellen hinweg von 12,6 % auftraten, konnte sowohl die distanz-basierte Methode als auch die probabilistische Methode stets alle Studenten anhand ihrer selbst-erzeugten Codes korrekt zuordnen. Auch bei den unter realen Bedingungen erzeugten Codes zeigen sich die vorgeschlagenen neuen Zuordnungstechniken der bisherigen Verfahrensweise deutlich überlegen.

4.4 Falsch-Positive

Falsch-Positive sind Paare von Codes aus zwei Wellen, die zwar nach der jeweiligen Verknüpfungsmethode einander zugeordnet werden, bei denen es sich aber nicht um dieselben Codeträger (Befragten) handelt. Gehen die durch die distanz-basierte oder probabilistische Methode mehr erfolgreich Zugeordneten mit einer sehr großen Zahl von Falsch-Positiven einher, würde dies gegen die Verwendung dieser Techniken sprechen. Die Ergebnisse unserer Experimente lassen jedoch den Schluss zu, dass Falsch-Positive kein ernsthaftes Problem für die Verwendung der neuen Techniken darstellt. Denn bei Verwendung der distanz-basierten oder der probabilistischen Methode kam es nur sehr selten zu falsch positiven Zuordnungen. So ergab die distanz-basierte Methode beim ersten Simulationsexperiment gerade 102 Falsch-Positive von 810 000 zu findenden Paaren. 543 320 wurden korrekt verknüpft, 266 578 Duplikate konnten nicht zugeordnet werden. Die probabilistische Methode weist nur 6 Falsch-Positive bei 543 484 korrekt verknüpften Paaren und 266 510 Duplikaten auf. Alle Falsch-Positiven traten unter der binären Codeversion auf, wobei alle drei Codelängen betroffen waren. Im zweiten Simulationsexperiment traten bei der distanz-basierten Methode keine Falsch-Positiven bei 179 541 korrekten Zuordnungen und 459 Ausfällen durch Duplikate auf. Für die probabilistische Methode ergaben sich 5 Falsch-Positive bei 179 659 korrekten Zuordnun-

gen und 336 Duplikaten. Im Experiment mit den real selbst-generierten Codes traten für beide Techniken keine Falsch-Positiven auf.

5 Zusammenfassung und Diskussion

Die Ergebnisse unserer Experimente haben gezeigt, dass die von uns vorgeschlagene Anwendung statistischer Methoden für die Zuordnung von Befragten anhand selbst-generierter Codes deutlich effektiver als eine automatisierte Zuordnung nach exakter Übereinstimmung ist. Da es sich um vollständig automatisierte fehlertolerante Verfahren handelt, ist unbedingt von einer wesentlich höheren Zeiteffizienz auszugehen, als das bisherige semi-manuelle Vorgehen erreichen kann. Wir schlagen deshalb vor, in Zukunft zumindest als erste Stufe des gesamten Zuordnungsverfahrens eine exakte Zuordnung durch eine der vorgestellten fehlertoleranten Techniken zu ersetzen.

Anhand unserer Ergebnisse konnten wir überdies zeigen, dass die Überlegenheit der neuen Techniken mit dem Ausmaß an Fehlern in den selbst-generierten Codes zunimmt. Diese Überlegenheit scheint dabei bei hohen Fehlerraten allein durch das Nachlassen der exakten Methode zu Stande zu kommen, denn die neuen Techniken zeigten sich für hohe Fehlerraten wenig anfällig. Dieser Umstand legt den Schluss nahe, die Ausgestaltung von selbst-generierten Codes zu verändern, wenn zur Zusammenführung fehlertolerante Zuordnungsmethoden verwendet werden können. In tatsächlichen Anwendungen von selbst-generierten Codes sollte die Fehlerrate mit der Länge und Komplexität der Codes zunehmen. Gleichzeitig steigt aber dadurch aufgrund der sinkenden Zahl von Duplikaten die Chance, für jeden Befragten einen eindeutigen Code zu erhalten. Weil nun die fehlertoleranten Techniken in der Lage sind, das Mehr an Fehlern in den Codes abzufangen und trotzdem eine hohe Zuordnungsrate zu erreichen, schlagen wir entsprechend vor, längere und komplexere Codes als bislang üblich in Verbindung mit einer Zuordnung durch fehlertolerante Zuordnungsmethoden zu verwenden.

Literatur

Elmagarmid, Ahmed; Ipeirotis, Panagiotis G. und ***Verykios, Vassilios*** 2006: Duplicate Record Detection: A Survey. Technical Report, Stern School of Business, New York University.

Grannis, Shaun J.; Overhage, J. Marc; Hui, Siu und ***McDonald, Clement J.*** 2003: Analysis of a Probabilistic Record Linkage Technique without Human Review. In: AMIA Annual Symposium Proceedings, S. 259-263 [Online verfügbar unter http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1479910].

Jaro, Matthew A. 1989: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. In: Journal of the American Statistical Association, 84 (406): 414–420.

Judson, Dean H. 2004: Computerized Record Linkage and Statistical Matching. In: ***Kempf-Leonard, Kimberly*** (Hrsg.): Encyclopedia of Social Measurement. Amsterdam: Elsevier.

Kearney, Kathleen A.; Hopkins, Ronald H.; Mauss, Armand L. und ***Weisheit, Ralph A.*** 1984: Self-Generated Identification Codes for Anonymous Collection of Longitudinal Questionnaire Data. In: Public Opinion Quarterly, 48 (1): 370–378.

Newcombe, Howard B. 1988: Handbook of Record Linkage. Methods for Health and Statistical Studies, Administration, and Business. Oxford: Oxford University Press.

Pöge, Andreas 2005: Persönliche Codes bei Längsschnittstudien: Ein Erfahrungsbericht. In: ZA-Information, 56: 50–69.

Schnell, Rainer; Bachteler, Tobias und ***Reiher, Jörg*** 2005: MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. In: ZA-Information, 56: 93–103.

Ukkonen, Esko 1985: Algorithms for Approximate String Matching. In: Information and Control, 64 (1-3): 100–118.

Winkler, William E. 1990: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: 1990 Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association, S. 354–359.

Winkler, William E. 1995: Matching and Record Linkage. In: ***Cox, Brenda G.; Binder, David A.; Chinnappa, B. Nanjamma; Christianson, Anders; Colledge, Michael J.*** und ***Kott, Phillip S.*** (Hrsg.): Business Survey Methods. New York: Wiley, S. 355–384.

Winkler, William E. 2006: Overview of Record Linkage and Current Research Directions. Technical Report RRS 2006/2, Statistical Research Division, U.S. Census Bureau, Washington, DC.