Patrick Noll

Statistisches Matching mit Fuzzy Logic

VIEWEG+TEUBNER RESEARCH

Entwicklung und Management von Informationssystemen und intelligenter Datenauswertung

Herausgeber:

Prof. Dr. Paul Alpar, Philipps-Universität Marburg
Prof. Dr. Ulrich Hasenkamp, Philipps-Universität Marburg

Patrick Noll

Statistisches Matching mit Fuzzy Logic

Theorie und Anwendungen in Sozial- und Wirtschaftswissenschaften

Mit einem Geleitwort von Prof. Dr. Paul Alpar

VIEWEG+TEUBNER RESEARCH

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<http://dnb.d-nb.de> abrufbar.

Dissertation Philipps-Universität Marburg, 2009

1. Auflage 2009

Lektorat: Dorothee Koch | Britta Göhrisch-Radmacher

Vieweg+Teubner ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.viewegteubner.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Druck und buchbinderische Verarbeitung: STRAUSS GMBH, Mörlenbach
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.
Printed in Germany

ISBN 978-3-8348-0836-3

Geleitwort

Statistisches Matching wurde ursprünglich zur Unterstützung der Marktforschung entwickelt. Um reichhaltigere Informationen über das Verbraucherverhalten gewinnen zu können, verschmolz man in den 1960er Jahren eine Erhebung zum Konsumverhalten mit einer Erhebung über Fernsehgewohnheiten zu einer einzigen Menge von Datensätzen, die dann Informationen zum Konsum- und Fernsehverhalten gleicher Objekte beinhaltete.

Ziel des statistischen Matchings ist es, weitere Informationen über Individuen zu erlangen, indem relevante Attribute ihrer sog. statistischen Zwillinge aus anderen Mengen von Datensätzen hinzugefügt werden. Traditionelle Matchingverfahren ermitteln die statistischen Zwillinge auf Grundlage der Distanzen zwischen den Ausprägungen der Datensätze in den sog. Matchingvariablen, die allen Datensätzen gemein sein müssen.

In der vorliegenden Arbeit wird eine Methode des statistischen Matchings mit Fuzzy Logic entwickelt. Der Autor nennt diese Methode statistisches Fuzzy-Matching. Durch die Verwendung der Theorie unscharfer Mengen zur Vorverarbeitung der Daten kann erstens eine neue Alternative zur Bestimmung der Distanzen zwischen Datensätzen entwickelt werden und zweitens wird die direkte Einbeziehung nominal und ordinal skalierter Variablen in den Matchingprozess ermöglicht. Insbesondere letzteres ist bei traditionellen Methoden nicht ohne aufwändige Vorverarbeitungen der Daten möglich. Die Umwandlung der Matchingvariablen in linguistische Variablen mit zugehörigen linguistischen Termen gestattet es, Distanzen zwischen Datensätzen auf Basis ihrer Zugehörigkeitsgrade zu einer Regelbasis zu bestimmen. Die Erstellung und der Aufbau der Regelbasis werden ebenfalls in dieser Arbeit gezeigt.

Statistisches Fuzzy-Matching dürfte u. a. in solchen Situationen den traditionellen Methoden überlegen sein, wenn kategorielle Variablen eine wichtige Rolle beim Matching spielen. Das in den Werten nicht enthaltene Anwenderwissen kann dann mit Hilfe von Zugehörigkeitsfunktionen eingebracht und für die Ermittlung der statistischen Zwillinge genutzt werden.

Neben der Entwicklung des theoretischen Ansatzes hat der Autor seine Methode auch programmtechnisch umgesetzt. In ausführlich dargestellten Anwendungsbeispielen werden detaillierte Vergleiche des statistischen Fuzzy-Matchings mit traditionellen Methoden gezogen. Gleichzeitig demonstrieren die Beispiele

die Funktionsweise der Methode und verdeutlichen unterschiedliche Ansatzpunkte des statistischen Matchings. Beim Fuzzy-Matching ist zwar bei metrisch skalierten Matchingvariablen ein etwas höherer Aufwand erforderlich, um bspw. die Definitions- und Wertebereiche der linguistischen Terme festzulegen. Dafür kann die Matching-Güte besser ausfallen, für deren Bestimmung der Autor ebenfalls eine neuartige Alternative vorstellt.

In der Praxis kann das Verfahren zur Datenanreicherung von Datenbeständen im Rahmen von Business Intelligence eingesetzt werden, die zunehmend eine wichtige Rolle auch in kleineren Unternehmen spielt, oder um umfangreiche Kundendaten unter Beachtung des Datenschutzes nutzen zu können.

Paul Alpar

Vorwort

Die Idee zu dieser Arbeit entstand während eines Forschungsprojekts, als traditionelle Methoden des statistischen Matchings zum Einsatz kommen sollten, um Ausprägungen von Variablen aus mehreren Mengen von Datensätzen miteinander vergleichen zu können. Ich erkannte relativ schnell, dass Methoden, die statistische Zwillinge allein auf Grundlage der Distanzen zwischen den Ausprägungen der Matchingvariablen ermittelten, einige Nachteile hatten. Bereits zu dieser Zeit reifte in mir der Wunsch, mich mit den Methoden des statistischen Matchings intensiver auseinanderzusetzen und eine eigene, verbesserte Alternative der Identifikation statistischer Zwillinge zu entwickeln.

Um die von mir identifizierten Nachteile traditioneller Methoden des statistischen Matchings ausgleichen zu können, benötigte ich ein Verfahren, das es mir bspw. gestattete, identischen Distanzen zwischen den Ausprägungen von Datensätzen unterschiedliche Bedeutungen beimessen zu können. Ich musste also eine Möglichkeit finden, Abstände zwischen Punkten in bestimmtem Umfang selber definieren zu können und die keine gewöhnliche Transformation von Daten darstellte. Da ich mich bereits während meines Studiums recht intensiv mit Fuzzy Logic befasst und ihre Vorzüge kennengelernt hatte, lag der Schluss nahe, statistisches Matching mit der Theorie der unscharfen Mengen zu verknüpfen. Durch die Verwendung der Fuzzy Logic und der Fuzzyfizierung der Ausgangsdaten erreichte ich das von mir gewünschte Ergebnis: Die Methode des statistischen Fuzzy-Matchings erweiterte die Funktionalitäten traditioneller Methoden des statistischen Matchings und bot darüber hinaus einige weitere Funktionalitäten wie bspw. die differenzierte Betrachtung fehlender Werte oder die direkte Einbeziehung nominal skalierter Variablen in den Matchingprozess.

Die in dieser Arbeit vorgestellte Methode stellt meine Bemühungen dar, ökonomische Theorien mit Methoden der Wirtschaftsinformatik und Statistik zu verknüpfen, um sowohl den Anwendern des statistischen Matchings eine Alternative zu den bisherigen Lösungen anzubieten als auch bspw. Anstöße zum Überdenken der gewöhnlichen Anwendung statistischer Analysen zu liefern. Die beiden ausgeführten Anwendungsbeispiele sollen unterschiedliche Einsatzmöglichkeiten des statistischen Matchings aufzeigen und Ideen zur Unterstützung multivariater Analysen liefern.

Danken möchte ich an erster Stelle meinem Doktorvater Prof. Dr. Paul Alpar für die vielen Erfahrungen, die ich zusammen mit ihm in interessanten wissenschaftlichen Projekten und in der universitären Lehre sammeln durfte, für die Unterstützung während der Erstellung dieser Arbeit und für die notwendigen Freiheiten zur Umsetzung meiner Ideen. Herrn Prof. Dr. Karlheinz Fleischer danke ich für die Übernahme des Zweitgutachtens und die wertvollen Hinweise während des Entstehens der Arbeit.

Herzlich danke ich Dr. Markus Pfuhl, Dr. Steffen Blaschke und Dr. Sebastian Pickerodt für die fruchtbaren Diskussionen und Anregungen während der Erstellung dieser Arbeit. Nicht zuletzt danke ich allen Kollegen am Institut für Wirtschaftsinformatik der Philipps-Universität Marburg für die freundschaftliche Arbeitsatmosphäre.

Ganz besonders danke ich meiner Frau Sibille. Sie hat während der Entstehung dieser Arbeit zu jeder Zeit an mich geglaubt, mich in schwierigen Phasen immer wieder aufgebaut und — genauso wie meine Tochter Anastasia — oft auf mich verzichten müssen. Meinen Eltern danke ich für ihre umfangreiche Unterstützung während der gesamten Studienzeit.

Patrick Noll

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Das Treffen wichtiger Entscheidungen erfordert die Nutzung aller relevanten und zugänglichen Informationen aus allen verfügbaren Datenquellen.[1] Qualitativ hochwertige Informationen bilden die Grundlage für faktenbasierte Entscheidungsfindungen.[2] Statistisches Matching kann dabei helfen, den Bedarf nach verlässlichen und widerspruchsfreien statistischen Informationen durch Kombination mehrerer Datenquellen zu stillen und Analysen zu ermöglichen, die auf Grundlage einzelner Datenquellen allein nicht möglich wären.[3]

Beim statistischen Matching werden Ähnlichkeiten zwischen Datensätzen bestimmt.[4] Traditionell werden dabei die Distanzen zwischen den Ausprägungen bestimmter Variablen (der sog. Matchingvariablen) betrachtet, die den zu vergleichenden Datensätzen gemein sind.[5] Generelles Ziel des statistischen Matchings ist es, Datensätze als sog. statistische Zwillinge zu finden, die sich hinsichtlich der Matchingvariablen möglichst wenig voneinander unterscheiden.[6] Der Nutzen des statistischen Matchings liegt darin, weitere Informationen über einen bestimmten Datensatz (resp. ein bestimmtes Individuum) zu erlangen, indem relevante Attribute seines statistischen Zwillings hinzugefügt werden.[7]

Durch die Kombination von Informationen aus unterschiedlichen Quellen werden vorhandene Datenbestände um zusätzliche Variablen erweitert.[8] Diese angereicherte Datenbasis kann als Grundlage für umfangreiche statistische Auswertungen und Anwendungen des Data Minings dienen. Darüber hinaus kann eine angereicherte Datenbasis die Verwendung bestimmter Data-Mining-Methoden ermöglichen, die mit den ursprünglichen Daten allein eventuell nicht eingesetzt werden könnten.[9] Data Mining stellt einen Schritt im Vorgehensmodell zur Wissensentdeckung in Datenbanken (Knowledge Discovery in Databases, KDD)[10] dar und

[1] Vgl. [KSM+07], S. 2.
[2] Vgl. [Pow00], S. 2.
[3] Vgl. [IOST00], S. 746 und [DZS01], S. 433.
[4] Vgl. [HIT97], S. 606ff.
[5] Vgl. [YA99], S. 2ff.
[6] Vgl. [RF98], S. 318.
[7] Vgl. [Sap00], S. 1.
[8] Vgl. [van00].
[9] Vgl. [KGG08], S. 594ff.
[10] Nach [FPSS96b] ist KDD „the nontrivial process of identifying valid, novel, potentially useful, and

bezeichnet die Anwendung spezifischer Algorithmen zur Extraktion von Mustern aus Daten.[11]

Bei der Suche nach Lösungen von bestimmten betriebswirtschaftlichen Problemen kann es vorkommen, dass die vorhandene Datenbasis zur Durchführung der notwendigen Analysen nicht umfassend genug ist.[12] Unternehmensexterne und / oder weitere unternehmensinterne Daten müssen zusätzlich in die Analysen einbezogen werden. Diese quantitative und qualitative Anreicherung der Daten kann helfen, unbefriedigende Ergebnisse des Data Minings auf Basis der ursprünglich vorhandenen Daten interpretierbar, interessant oder anwendbar zu machen.[13] Aus diesem Grund verlangt bspw. das Prozessmodell CRISP-DM (Cross Industry Standard Process for Data Mining) die Festlegung der Data-Mining-Ziele vor der Auswahl der Daten.[14] Betrachtet man den Data-Mining-Prozess nach Fayyad et al. (1996a) in Abbildung 1.1, so lässt sich die Datenanreicherung durch statistisches Matching sowohl im Schritt der Auswahl der Daten als auch in die Vorverarbeitung einordnen.

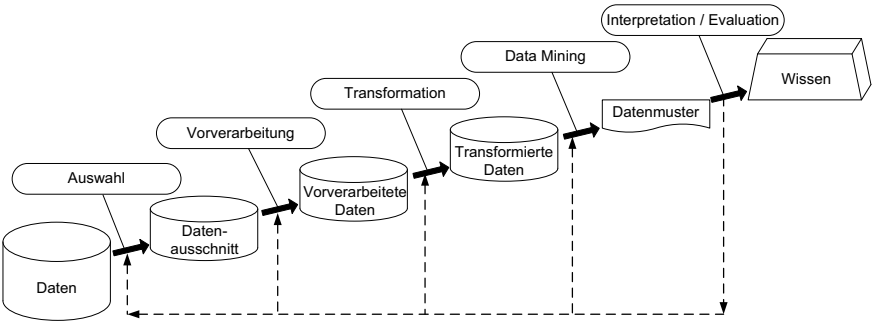

Abbildung 1.1: Schritte im Data-Mining-Prozess (vgl. [FPSS96a], S. 41.)

Im Schritt der Datenauswahl aus dem gesamten Datenbestand kann statistisches Matching eingesetzt werden, um die bestehende Datenbasis um zusätzliche Informationen zu erweitern. Durch das Hinzufügen relevanter Attribute der mit Hilfe des statistischen Matchings gefundenen statistischen Zwillinge aus anderen zugänglichen Datenquellen, kann die vorhandene Datenbasis vergrößert werden. Hir-

ultimately understandable patterns in data“.

[11] Vgl. [FPSS96a], S. 37 und [HK06], S. 5ff. Der Prozess der KDD wird auch Data-Mining-Prozess genannt. Vgl. [AN00], S. 4.

[12] Vgl. [Pet05], S. 57.

[13] Vgl. [Alp04], S. 1219f.

[14] Vgl. [CCK+00] und http://www.crisp-dm.org.

ji (2001) schlägt zur Verbesserung von Data-Mining-Ergebnissen die Erhöhung der Dimensionalität der Datenbasis durch das Hinzufügen demografischer Daten und das Anwenden zusätzlicher Data-Mining-Algorithmen im sog. *Back End Data Mining* vor, das im Zentrum des von ihm entwickelten Data-Mining-Prozesses steht.[15] Neben dem Vorteil der vergrößerten Datenbasis für Data Mining kann durch das Hinzufügen zusätzlicher Attribute die Anwendung bestimmter Data-Mining-Methoden erst möglich werden. Verfügt die angereicherte Datenbasis im Gegensatz zur ursprünglichen Datenmenge bspw. über binäre Variablen (wie *Kunde: ja/nein* oder *Zweitbankverbindung: ja/nein*), so kann es gegebenenfalls erst auf Grundlage der angereicherten Menge von Datensätzen sinnvoll sein, einen Entscheidungsbaum zu erstellen.

Nach Miller und Han (2001) umfasst der Schritt der Vorverarbeitung der ausgewählten Daten aus dem gesamten Datenbestand im Data-Mining-Prozess die Imputation fehlender Werte, das Eliminieren von Duplikaten, das Anreichern der Daten durch die Kombination mehrerer Datenquellen und weitere notwendige Aufbereitungen der Daten.[16] Statistisches Matching kann neben der Anreicherung der Daten mit zusätzlichen Informationen, die im Schritt der Vorverarbeitung analog zur Erweiterung des gesamten Datenbestands zu sehen ist, zum Ersetzen fehlender Werte eingesetzt werden. Oft werden fehlende Werte in Datenbeständen durch Lagemaße wie den arithmetischen Mittelwert, den Median oder den Modus ersetzt, die direkt aus den Daten berechnet werden.[17] Nachteile dieses Vorgehens sind, dass durch das Ersetzen aller fehlenden Werte einer Variablen mit demselben Wert die Varianz verkleinert und nur durch Zufall ein inhaltlich passender Wert für jeden Datensatz gefunden wird.[18] Durch statistisches Matching werden fehlende Werte eines Datensatzes durch die Werte seines statistischen Zwillings aus der selben Menge von Datensätzen oder aus einer anderen Quelle ersetzt. Der erste Fall wird in der Literatur zur Behandlung fehlender Werte als *Hot Deck Imputation* und der zweite als *Cold Deck Imputation* bezeichnet.[19] Durch dieses Vorgehen wird die Varianz nur minimal verkleinert, da jedem Datensatz die Werte eines anderen statistischen Zwillings zugewiesen werden und nicht die fehlenden Werte aller Datensätze in einer Variablen mit demselben Wert oder einer Zufallszahl aus einer geschätzten Verteilung ersetzt werden.[20] Darüber hinaus wird die inhaltliche

[15] Vgl. [Hir01], S. 87ff.
[16] Vgl. [MH01], S. 7f.
[17] Vgl. [TH08], S. 265.
[18] Vgl. [von04], S. 106.
[19] Vgl. [Göt07], S. 127. Eine Übersicht von Verfahren zur Behandlung fehlender Werte findet sich ebenfalls in [Göt07].
[20] Vgl. [Höf04], S. 94.

Qualität der ersetzten Werte erhöht, da jedem Datensatz die Werte eines auf Basis der Matchingvariablen sehr ähnlichen anderen Datensatzes zugewiesen werden.

Ziel der vorliegenden Arbeit ist es, eine Methode des statistischen Matchings auf Basis von Fuzzy-Logic zu entwickeln, die gängige Matching-Methoden (wie bspw. Nearest-Neighbour-Verfahren auf Grundlage der euklidischen Distanz)[21] beinhaltet und darüber hinaus weitere Funktionalitäten bietet. Die erweiterten Funktionalitäten des statistischen Fuzzy-Matchings entstehen durch die Verwendung der in der Theorie der unscharfen Mengen (Fuzzy Logic) verwendeten linguistischen Variablen mit ihren zugehörigen linguistischen Termen. In Anlehnung an den Begriff „statistisches Matching“ wird die zu entwickelnde Methode des statistischen Matchings mit Fuzzy Logic „statistisches Fuzzy-Matching“ genannt.[22]

Beim traditionellen statistischen Matching werden die Distanzen zwischen den Ausprägungen der Matchingvariablen der zu betrachtenden Datensätze bestimmt und darauf aufbauend die statistischen Zwillinge identifiziert. Die beiden Datensätze, die die geringste (Gesamt-) Distanz zueinander aufweisen, werden zu statistischen Zwillingen. Beim statistischen Fuzzy-Matching werden die Ausgangsdaten nicht direkt zur Bestimmung der Distanzen verwendet. Sie müssen zunächst *fuzzyfiziert* werden, um ihnen Zugehörigkeitsgrade zu linguistischen Termen der linguistischen Variablen zuweisen zu können. Beim statistischen Fuzzy-Matching bilden Regelbasen die Grundlage zur Bestimmung der Distanzen zwischen Datensätzen. Ihre Erstellung soll in dieser Arbeit für auf Fuzzy-Logic basierende Verfahren gezeigt werden.[23] Dabei werden nicht die Distanzen zwischen den tatsächlichen Ausprägungen der Datensätze bezüglich bestimmter Variablen berechnet, wie es bei den traditionellen Methoden des statistischen Matchings der Fall ist, sondern die Distanzen zwischen den Zugehörigkeitsgraden der Datensätze zur Regelbasis. Jedem Datensatz wird ein Zugehörigkeitsgrad zu jeder einzelnen Regel der Regelbasis zugewiesen, die gemeinsam den Vektor der Zugehörigkeitsgrade zur Regelbasis bilden. Aufbauend auf diesen Vektoren der Zugehörigkeitsgrade

[21] Vgl. [TPT01], S. 255 oder [Pd00], S. 415f.

[22] Der Aufsatz von [ACA93] impliziert eine enge inhaltliche Überschneidung mit der vorliegenden Arbeit. Abdulghafour et al. verwenden die Theorie der unscharfen Mengen, um unvollständige Bilder verschiedener Sensoren mit jeweils unterschiedlichen Inhalten (wie bspw. Schatten, Farbtiefe, 3-D-Informationen usw.) zu einem einzigen Bild zu vereinen. Ein wesentlicher Unterschied zu der vorliegenden Arbeit ist, dass sich Abdulghafour et al. sehr stark an der Fuzzy-Regelungstechnik orientieren, indem sie lediglich drei festgelegte linguistische Terme zur Klassifikation der Qualität von Bildpixeln und ausschließlich „wenn-dann-Regeln“ zur Fusion der Einzelbilder verwenden. Es handelt sich dabei nicht um statistisches Matching, wie die Verwendung des Begriffs „Data Fusion“ im Titel des Aufsatzes nahelegt, da keine Distanzen zwischen Datensätzen bestimmt und keine statistischen Zwillinge identifiziert werden.

[23] Einen guten Überblick über Theorie und Anwendungen der Fuzzy Logic geben z. B. [Zad65b], [DP80], [Zim93], [Cox94], [Zim94], [BT99] und [BB07].

wird schließlich das statistische Fuzzy-Matching durch Identifikation der statistischen Zwillinge vollzogen.

In bestimmten Situationen (z. B. bei Vorliegen identischer Abstände zwischen den ursprünglichen Ausprägungen der Datensätze) können daher Entscheidungen über statistische Zwillinge getroffen werden, in denen traditionelle Methoden keine Entscheidung finden können. Identischen Abständen zwischen den Ausprägungen einer Variablen mehrerer Datensätze kann mit statistischem Fuzzy-Matching unterschiedliche Bedeutung beigemessen werden. Sie haben nicht mehr die gleiche Bedeutung für das Finden der statistischen Zwillinge.

Betrachtet man bspw. drei Individuen im Alter von 55, 60 und 65 Jahren: Der 55-Jährige und der 65-Jährige weisen zum 60-Jährigen jeweils den selben Altersabstand von 5 Jahren auf. Auf Basis der Altersabstände kann keine Aussage darüber getroffen werden, welches der beiden Individuen ähnlicher zum 60-Jährigen ist. Es ergeben sich erst Unterschiede, sobald der 55-Jährige einen Zugehörigkeitsgrad von z. B. $0,5$ zu einem linguistischen Term (z. B. „alt“) der linguistischen Variablen „Alter“ aufweist, der 60-Jährige den Zugehörigkeitsgrad von $0,8$ und der 65-Jährige den Zugehörigkeitsgrad von 1. Basierend auf den individuellen Zugehörigkeitsgraden kann eine Entscheidung zugunsten des 65-Jährigen getroffen werden, denn der 65-Jährige weist nun eine geringere Distanz $(0,2)$ zum 60-Jährigen auf als der 55-Jährige $(0,3)$.

Neben dem Vorteil des statistischen Fuzzy-Matchings, dass identischen Abständen in den Ausgangsdaten unterschiedliche Bedeutungen beigemessen werden können, erlaubt das statistische Fuzzy-Matching das Einbeziehen fehlender Werte in den Ausgangsdaten in die Bestimmung der statistischen Zwillinge. Zusätzlich zu fehlenden Werten durch Nichtbeantwortung einzelner Items enthalten empirische Erhebungen verweigerte Antworten durch Kennzeichnung von Fragen mit „keine Angabe“ und Antworten wie „weiß ich nicht“, die durch fehlende Informationen bzw. fehlende Kompetenz zur qualifizierten Antwort entstehen.[24] Mit Hilfe des statistischen Fuzzy-Matchings können diese Angaben differenziert betrachtet und für beliebig skalierte Variablen direkt in den Matchingprozess zur Suche nach statistischen Zwillingen eingebunden werden. Bei Verwendung eines einfachen statistischen Distanzmatchings auf Basis der standardisierten Ausgangsdaten können fehlende Werte dagegen nicht direkt in die Betrachtungen integriert werden.[25] Darüber hinaus können auch nominal skalierte Variablen, wie bspw. Angaben zum Familienstand oder zum Beruf mit der Methode des statistischen Fuzzy-Matchings leicht in die Berechnungen der Distanzen zwischen Datensätzen einbezogen werden, ohne den Daten dabei eine Rangordnung zu unterstellen, wie es bspw. bei

[24] Vgl. [Cle08], S. 25.

[25] Sieht man von den Möglichkeiten der Imputation fehlender Werte ab.

ordinal oder metrisch skalierten Variablen der Fall ist.

Im Folgenden wird der Inhalt der einzelnen Kapitel und deren Funktion im Kontext der Arbeit kurz erläutert. Im Kapitel 2 werden die theoretischen Grundlagen des statistischen Matchings dargestellt. Es werden der Kerngedanke des statistischen Matchings, der traditionelle Matching-Prozess und das sog. *Propensity Score-Matching* als eine Alternative zu den traditionellen Matchingverfahren auf Grundlage der Distanzen der Ausgangsdaten vorgestellt. In Kapitel 3 wird das notwendige mathematische Fundament der Fuzzy Logic, der „Theorie der unscharfen Mengen“, geformt. In diesem Kapitel werden zunächst scharfe und unscharfe Mengen voneinander abgegrenzt und Operationen und Eigenschaften unscharfer Mengen vorgestellt. Im Anschluss daran wird das Konzept der linguistischen Ausdrücke (linguistische Variablen und linguistische Terme) kurz vorgestellt, ehe die Möglichkeiten der subjektiven und objektiven Beschaffung von Zugehörigkeitsfunktionen betrachtet werden. Kapitel 3 endet mit der grundlegenden Betrachtung der Fuzzy-Regeln, die das Fundament zur Erstellung von Regelbasen liefern.

Die Kapitel 4, 5 und 6 bilden den Hauptteil dieser Arbeit und haben die theoretische (Kapitel 4) und praktische Ausarbeitung (Kapitel 5 und 6) des statistischen Fuzzy-Matchings zum Inhalt. In Kapitel 4 wird die theoretische Basis des statistischen Fuzzy-Matchings gelegt. Nach der einleitenden Motivation zur Entwicklung der Methode des statistischen Fuzzy-Matchings wird zunächst die Anpassung der linguistischen Ausdrücke an das jeweilige Matchingproblem beschrieben. Im Anschluss daran wird die Festlegung der Zugehörigkeitsfunktionen erörtert, die jedem scharfen Ausgangswert Zugehörigkeitsgrade zu linguistischen Termen zuweisen und die Ausgangsdaten fuzzyfizieren. Die linguistischen Terme aller linguistischen (Matching-) Variablen bilden die Grundlage zur Erzeugung der Regelbasis, da jede Regel der Regelbasis aus der Verknüpfung eines linguistischen Terms jeder linguistischen Variablen besteht. Der Kern des statistischen Fuzzy-Matchings liegt in der Bestimmung der Distanzen zwischen Datensätzen durch die Ermittlung der Distanzen ihrer Zugehörigkeitsgrade zur Regelbasis. Daher werden im weiteren Verlauf des vierten Kapitels verschiedene Arten der Berechnung der Zugehörigkeitsgrade der Datensätze zur Regelbasis betrachtet, die von der verwendeten Verknüpfung der Regeln abhängt. Das Kapitel wird mit der Betrachtung der konkreten Vorgehensweise zur Identifizierung der statistischen Zwillinge beendet.

In Kapitel 5 wird die programmtechnische Umsetzung, also die Implementierung des statistischen Fuzzy-Matchings beschrieben und die Bedienung der Software kurz erläutert. Im anschließenden Kapitel 6 wird die Leistungsfähigkeit der hier entwickelten und vorgestellten Methode des statistischen Fuzzy-Matchings anhand zweier Anwendungsbeispiele überprüft und mit anderen gängigen Methoden des statistischen Matchings verglichen. Beide Anwendungsbeispiele sol-

len gleichzeitig Hinweise auf praktische Einsatzgebiete des statistischen Fuzzy-Matchings liefern und unterschiedliche Vorgehensweisen zum Erlangen weiterer Informationen bzw. weiteren Wissens aufzeigen. Im abschließenden Kapitel 7 werden die wichtigsten Ergebnisse dieser Arbeit noch einmal zusammengefasst, ein Fazit gezogen und ein Ausblick auf mögliche zukünftige Forschungsaufgaben im Gebiet des statistischen Fuzzy-Matchings gegeben.

2 Statistisches Matching

Analysen von Daten benötigen oft Informationen, die nicht in einer einzelnen Quelle enthalten, sondern über mehrere Quellen verteilt zu finden sind.[1] Die Methoden des statistischen Matchings helfen dabei, Informationen aus unterschiedlichen Quellen in einem einzigen Datenbestand zu vereinen.[2] Eine frühe Erwähnung des statistischen Matchings findet sich z. B. bei Okner (1972), der Untersuchungen aus der Mitte der 1960er Jahre beschreibt, die Beziehungen zwischen Variablen aus unterschiedlichen Quellen beinhalten. Ziel dieser Untersuchungen war die Erstellung einer Menge von Datensätzen die sowohl sozio-demografische Informationen enthielt, als auch Informationen über Einkommen und Steuerzahlungen. Das Problem war, dass es keine Untersuchung gab, die die gestellten Anforderungen komplett erfüllte. Die einzige Möglichkeit zur Durchführung einer solchen Analyse war das Verschmelzen der vorhandenen Informationen aus mehreren Quellen in eine einzige Menge von Datensätzen, weil eine neue Umfrage mit allen benötigten Attributen aus Kosten- oder Zeitmangel nicht durchgeführt werden konnte.[3] So verwendete man in der erwähnten Untersuchung die Steuerdatei aus dem Jahre 1966 und verschmolz die darin enthaltenen Informationen mit der „Survey of Economic Opportunities“ aus dem Jahre 1967 und erzielte damit das gewünschte Ergebnis. Eine Technik, die zur Durchführung solcher Analysen entwickelt wurde, ist das sog. *statistische Matching*.[4]

In den folgenden Abschnitten dieses Kapitels werden zunächst die Grundlagen des statistischen Matchings und anschließend die Annahmen und Merkmale des traditionellen Matching-Prozesses erläutert. Darüber hinaus wird in Abschnitt 2.3 das Propensity Score Matching beschrieben, das die Wahrscheinlichkeit der Zugehörigkeit eines Datensatzes zur Gruppe der Teilnehmer oder Nicht-Teilnehmer an einer Maßnahme basierend auf den Matchingvariablen bestimmt. Im Anschluss daran werden die beiden Konzepte des *constrained* und *unconstrained* Matching diskutiert. Beim constrained Matching werden bestimmte Bedingungen an die Matchingmethode gestellt, die beim unconstrained Matching abgeschwächt sind. Das Kapitel endet mit einer kurzen Beschreibung der in der Literatur zu findenden

[1] Vgl. [Bac02], S. 2.
[2] Vgl. [IOST00], S. 746.
[3] Vgl. [HIT97], S. 606.
[4] Vgl. [RD81], S. 128.

Kritik am statistischen Matching.

2.1 Grundlagen des statistischen Matchings

Statistisches Matching entstammt dem Aufgabengebiet der Marktforschung.[5] Insbesondere in Umfragen zum Medien- und Konsumverhalten von Individuen sind so viele Fragen von Interesse, dass diese nicht einer einzelnen Menge von Probanden gestellt werden können, sondern auf mehrere Fragebögen verteilt werden müssen.[6] Auf Basis gemeinsamer Attribute (z. B. demografischer Merkmale) werden die separaten Umfragen anschließend vereint, indem jedem Individuum einer bestimmten (Probanden-) Menge die fehlenden Attribute seiner statistischen Zwillinge der anderen (Probanden-) Mengen zugeordnet werden. Abbildung 2.1 zeigt eine einfache Illustration des statistischen Matchings. Es ist jeweils ein beliebiger Datensatz aus einem *consumer panel* und einem *television panel* mit den zugeordneten Attributen dargestellt. Beide Mengen von Datensätzen haben gemeinsame, aber auch spezifische Attribute. Die beiden betrachteten Datensätze werden auf Basis der gemeinsamen Variablen zusammengeführt. Dem Individuum *425* aus dem *television panel* wird das Individuum *13* aus dem *consumer panel* gegenübergestellt und dessen zusätzliche Attribute (rent cars, views daily soaps, views news und zaps advertisement) hinzugefügt.

Grundsätzlich kann die Integration von Daten aus unterschiedlichen Quellen durch drei verschiedene Methoden erreicht werden: Data Merging, record linkage und statistisches Matching.[7] Data merging bezeichnet den Prozess des Zusammenführens von Daten aus verschiedenen Quellen unter Auflösung von auftretenden Konflikten wie *Widersprüchen* (unterschiedliche Ausprägungen des selben Attributs bei dem selben Individuum in unterschiedlichen Quellen) und *Unsicherheiten* (z. B. fehlende Informationen durch Null-Werte).[8] Record linkage wird zur Eliminierung von Duplikaten in Datenbanken und zur Verknüpfung von Datensätzen zu identischen Individuen aus unterschiedlichen Quellen benutzt.[9] Statistisches Matching behandelt dagegen das Problem der Integration von Datensätzen zu unterschiedlichen Individuen in eine Tabelle auf Basis identischer Attribute.[10]

Neben der Erlangung zusätzlichen Wissens eignen sich statistische Zwillinge

[5] Vgl. [vKG02], S. 2f.
[6] Vgl. [Sap00], S. 1f.
[7] Vgl. [DZS01], S. 433.
[8] Vgl. [Ble04], S. 23f.
[9] Vgl. [Fai97], S. 428. Das Problem des record linkage wurde bereits im Jahre 1959 von [NKAJ59] beschrieben.
[10] Vgl. [Win95], S. 375.

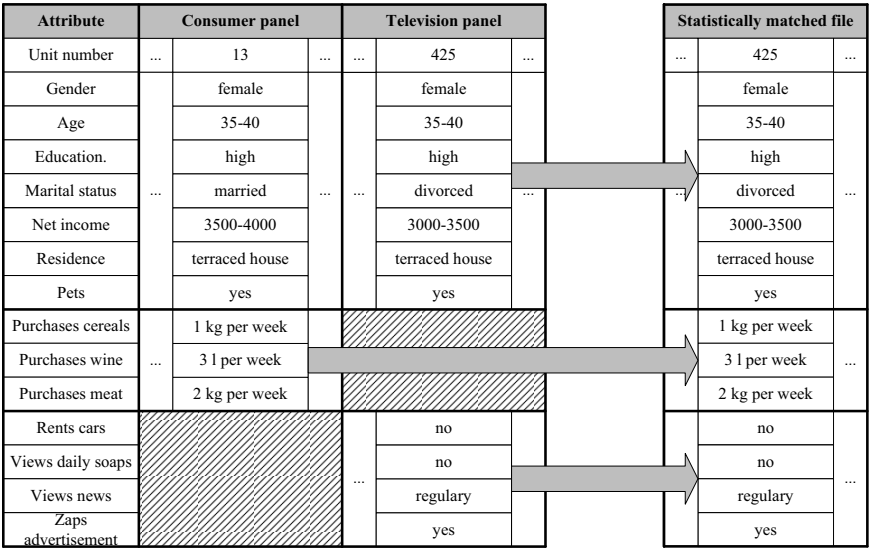

Abbildung 2.1: Illustration des statistischen Matchings (vgl. [RÖ2], S. 3.)

auch zur Schätzung bzw. Imputation fehlender Werte, zur Schätzung der Wirkung einer Untersuchungsvariablen und zur Bestimmung von Kontrollgruppen:[11]

Untersuchung der Wirkung einer Variablen auf eine andere. Man betrachte eine Menge von Datensätzen, bei denen die Wirkung einer Variablen B auf eine andere Variable Y untersucht werden soll. Die sich zum Matching gegenüberstehenden Mengen von Datensätzen ergeben sich aus den Ausprägungen B_j $(j = 1, ..., q)$ von B (z. B. $B_1 = ja$ und $B_2 = nein$ bei der Frage nach einem bestimmten Merkmal). B_k $(k \in \{1, ..., q\})$ trete eher selten auf und die Individuen mit dieser Ausprägung unterscheiden sich in den anderen beobachteten Attributen X_i, $(i = 1, ..., p)$ von den Individuen mit den Ausprägungen B_r $(r \in \{1, ..., q\},\ r \neq k)$. In den Anwendungsbeispielen zu der in dieser Arbeit im Kapitel 4 entwickelten Methode des statistischen Fuzzy-Matchings wird ein solcher Ansatz im Abschnitt 6.1 betrachtet. B stellt dabei das Merkmal „Arbeitslosigkeit“ mit den beiden Ausprägungen *ja* und *nein* dar. Die gemeinsamen Variablen X_i setzen sich aus demografischen Attributen wie Alter, Bildung und Familienstand und aus subjektiven Attributen zu gesellschaftlichen und politischen Einstellungen zusammen. Mit Hilfe des statis-

[11] Vgl. [Bac02], S. 39ff.