Datenfusion und Datenintegration: 6. wissenschaftliche Tagung

König, Christian (Ed.); Stahl, Matthias (Ed.); Wiegand, Erich (Ed.)

Veröffentlichungsversion / Published Version
Konferenzband / conference proceedings

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

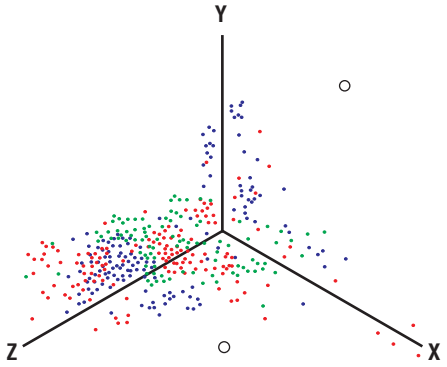
Empfohlene Zitierung / Suggested Citation:
König, C., Stahl, M., & Wiegand, E. (Hrsg.). (2005). *Datenfusion und Datenintegration: 6. wissenschaftliche Tagung* (Tagungsberichte / Informationszentrum Sozialwissenschaften, 10). Bonn: Informationszentrum Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-261147

Datenfusion und Datenintegration

6. Wissenschaftliche Tagung

Christian König, Matthias Stahl, Erich Wiegand (Hrsg.)

im Auftrag
des Statistischen Bundesamtes, Wiesbaden,
der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI)
und des ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.



Tagungsberichte Band 10

InformationsZentrum
Sozialwissenschaften

GESIS

Datenfusion und Datenintegration

Tagungsberichte

Herausgegeben vom Informationszentrum Sozialwissenschaften (IZ)
der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI), Bonn.
Band 10

Das IZ ist Mitglied der Gesellschaft Sozialwissenschaftlicher
Infrastruktureinrichtungen e.V. (GESIS).
Die GESIS ist Mitglied der Leibniz-Gemeinschaft.

Datenfusion und Datenintegration

6. Wissenschaftliche Tagung

Christian König, Matthias Stahl, Erich Wiegand (Hrsg.)

im Auftrag

des Statistischen Bundesamtes, Wiesbaden,
des ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.
und der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI)

Tagungsberichte Band 10

Informationszentrum Sozialwissenschaften, Bonn 2005

Bibliographische Information Die Deutsche Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliothek; detaillierte bibliographische Daten sind im Internet über www.ddb.de abrufbar.

ISBN 3-8206-0148-1

Herausgeber,
Druck und Vertrieb: Informationszentrum Sozialwissenschaften
Lennéstraße 30, 53113 Bonn
Tel.: 02 28 - 22 81 - 0

Printed in Germany

Inhalt

Begrüßung

Johann Hahlen
Präsident des Statistischen Bundesamtes

Sehr geehrte Kolleginnen und Kollegen aus dem ADM,
sehr geehrte Kolleginnen und Kollegen aus der ASI,
liebe Kolleginnen und Kollegen aus der amtlichen Statistik,
verehrte Gäste,

die wissenschaftliche Tagung über „Datenfusion und Datenintegration", zu der ich Sie heute hier im Statistischen Bundesamt in Wiesbaden herzlich begrüße, ist die inzwischen sechste gemeinsame wissenschaftliche Veranstaltung des ADM Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute e. V., der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V. (ASI) und des Statistischen Bundesamtes. Unsere gemeinsame Veranstaltungsreihe war von ihrem Beginn im Jahre 1995 an auf einen zweijährigen Tagungsturnus ausgerichtet. Wir nahmen an, dass sich im Zweijahresturnus über einen längeren Zeitraum stets wieder Themen finden, die gleichermaßen für die akademische Sozialforschung, die kommerzielle Marktforschung und für die amtliche Statistik von Interesse sind. Diesem Anspruch sind die Tagungen bisher gerecht geworden. Nicht zuletzt die Fortschritte in der Informationstechnik und in der statistischen Methodik zeigen immer wieder neue Themenfelder auf, die für alle Beteiligten gleichermaßen interessant sind. Die Tagungsreihe begann im Jahr 1995 mit der Tagung „Pretest und Weiterentwicklung von Fragebogen" und befasste sich 1997 mit dem Interviewereinsatz und der Interviewerqualifikation. Diese beiden ersten Tagungen behandelten klassische, originäre Bereiche der Umfrageforschung. Im Jahr 1999 standen neue Erhebungstechniken und die damit verbundenen Methodeneffekte auf der Tagesordnung, unsere Tagung richtete sich also erstmals auf neue Techniken. Im Jahr 2001 diskutierten wir Aspekte internationaler und interkultureller Umfragen und trugen der zunehmenden Internationalisierung der Umfragen Rechnung, ein Trend, der bis heute fortdauert. 2003 standen wiederum neue Erhebungstechniken auf dem Programm, nämlich die verschiedensten Facetten von Online-Erhebungen. In diesem Jahr haben wir uns mit der Datenfusion und der Datenintegration ein Thema gestellt, das ohne den immer schnelleren technischen Fortschritt auf dem Gebiet der Datenverarbeitung und Informationstechnik gar nicht vorstellbar wäre.

Der kleine Rückblick auf die Geschichte unserer gemeinsamen Tagungen zeigt, dass der zeitliche Abstand zwischen den Tagungsthemen, die neuen Tech-

niken gewidmet sind, kürzer geworden ist. Dies verwundert auch nicht, denn die Herausforderungen, vor der die kommerziellen Marktforschungsinstitute, die sozialwissenschaftliche Forschung und die amtliche Statistik stehen, sind durchaus vergleichbar. Während sich die kommerziellen Institute im Wettbewerb am Markt behaupten müssen, also Gewinne erwirtschaften bzw. Marktanteile sichern müssen, zwingen die geringeren Spielräume in den öffentlichen Haushalten die akademische Forschung wie auch die amtliche Statistik dazu, mit geringeren Ressourcen auszukommen und dennoch die Qualität der eigenen Produkte, z.B. deren Aktualität und Genauigkeit ständig zu verbessern. Wir können diesen Herausforderungen nur durch einen effizienten Einsatz modernster Technologien gerecht werden.

Die meisten von Ihnen begleiten unsere Arbeit seit vielen Jahren auch als Nutzer und Kunden und erfahren dabei, dass der time-lag bis zur Implementierung neuester Techniken deutlich geringer geworden ist. Denn Qualität ist heute mehr als nur die Genauigkeit der Ergebnisse und die Vergleichbarkeit über die Zeit. Die Aktualität der Daten und die Kosten ihrer Gewinnung sind ebenso bedeutsam. Und damit, meine Damen und Herren, sind wir wieder an einem Punkt, wo trotz unterschiedlichster Produktionsbedingungen kommerzielle Marktforschung, akademische Sozialforschung und amtliche Statistik im gleichen Boot sitzen. Wir alle müssen Kosten sparen und gleichzeitig schneller, also aktueller werden, und dies bei gleicher Ergebnis-Validität.

Das Programm der kommenden zwei Tage könnte möglicherweise zu dem Missverständnis führen, dass wir amtlichen Statistiker bei den Themen Datenfusion und Datenintegration erheblichen Nachholbedarf hätten, weil wir nur mit einem Thema und zwar aus dem Bereich des Zensustests vertreten sind. Dieser erste Eindruck täuscht aber: Die Nutzung von Sekundärdaten spielt in der amtlichen Statistik bereits in vielen Bereichen eine wichtige Rolle. In den Unternehmensstatistiken sind wir dabei, über das Unternehmensregister Daten für verschiedene Erhebungen zu nutzen und so nicht zuletzt die auskunftspflichtigen Unternehmen zu entlasten. In der Landwirtschaftsstatistik nutzen wir sekundärstatistische Daten, um gleichermaßen die auskunftspflichtigen Betriebe zu entlasten und Kosten zu sparen. Auch wenn wir schon für manche Statistiken auf Sekundärdaten zurückgreifen, ist uns bewusst, dass uns die Kolleginnen und Kollegen sowohl in der akademischen Sozialforschung, vor allem aber in den kommerziellen Marktforschungsinstituten hier zum Teil voraus sind.

Dies liegt auch daran, dass wir in vielen Statistikbereichen gesetzlich zur Erhebung von Merkmalen verpflichtet sind, die sich nur primär erheben lassen, und es für praktisch alle Statistiken vielfältige und detaillierte gesetzliche Regelungen gibt. Aus diesen Gründen könnten wir - selbst nach Klärung aller methodischen Fragen - Techniken der Datenfusion und Datenintegration kurzfristig nicht einsetzen, weil die betroffenen statistikgesetzlichen Regelungen angepasst werden müssen.

Hier sind wir in diesem Monat, was das Zusammenführen von Daten in amtlichen Statistiken angeht, ein bedeutendes Stück weitergekommen. Vor gut zwei Wochen, am 14. Juni 2005 ist das sogenannte „Gesetz zur Änderung des Statistikregistergesetzes und sonstiger Statistikgesetze" in Kraft getreten, das durch eine effizientere Nutzung der bei den Statistischen Ämtern des Bundes und der Länder bereits vorhandenen Daten neue statistische Erhebungen vermeiden und die Auskunftspflichtigen entlasten soll. Das Zusammenführen von Daten aus Wirtschafts- und Umweltstatistiken mit Daten aus dem Statistikregister und Daten, die nach dem Verwaltungsdatenverwendungsgesetz bereits in den statistischen Ämtern vorliegen, lässt uns zusätzliche Informationen gewinnen, ohne Befragungen durchführen zu müssen. Datenverknüpfungen, die bisher wegen des damit verbundenen hohen Organisations- und Arbeitsaufwandes nur selten durchgeführt worden sind, werden durch Neufassung des § 13a BStatG im Interesse einer besseren Nutzung vorhandener Daten erleichtert.

Diese Neufassung des § 13a des Bundesstatistikgesetzes lautet: „Soweit es zur Gewinnung von statistischen Informationen ohne zusätzliche statistische Erhebungen erforderlich ist, dürfen Daten aus Statistiken nach § 13 Abs. 1, Daten aus dem Statistikregister, Daten nach dem Verwaltungsdatenverwendungsgesetz und Daten, die die statistischen Ämter des Bundes und der Länder aus allgemein zugänglichen Quellen gewinnen, zusammengeführt werden."

Der neue § 13 a erfasst allerdings nur Wirtschafts- und Umweltstatistiken bei Unternehmen, Betrieben und Arbeitsstätten. Die neuen gesetzlichen Regelungen gelten also nicht bei Personenstatistiken.

Diese neue gesetzliche Regelung ermöglicht es der amtlichen Statistik künftig, in dem für uns schon von der Anzahl der Erhebungen her besonders bedeutsamen Feld der Wirtschafts- und Umweltstatistiken verstärkt sekundärstatistische Quellen zu nutzen. Einen passenderen Zeitpunkt, um eine wissenschaftliche Tagung zum Thema Datenfusion und Datenintegration hier in den Räumen des Statistischen Bundesamtes abzuhalten, hätten die Veranstalter nicht wählen können.

Das Programm der kommenden zwei Tage hat viele ausgewiesene Experten versammelt, die uns einen differenzierten Einblick in Techniken und ausgewählte Einsatzgebiete von Datenintegration und Datenfusion geben werden.

Ich freue mich sehr, dass sich Herr Hartmut Scheffler bereit erklärt hat, uns als Moderator durch die kommenden zwei Tage zu leiten. Herr Scheffler war langjähriger Geschäftsführer des Emnid-Instituts und ist seit dem Zusammenschluss von TNS Emnid und TNS Infratest zur TNS Infratest Holding dort Mitglied der Geschäftsführung mit der Verantwortung für die Bereiche Marketing, Forschung und Entwicklung, Kommunikations-, Konsum- und Mediaforschung. Darüber hinaus ist Herr Scheffler gerade in der vergangenen Woche zum neuen Vorstandsvorsitzenden des ADM gewählt worden. Deshalb gratuliere ich Ihnen, Herr Scheffler, dazu herzlich und wünsche Ihnen für die mit der

neuen Tätigkeit verbundenen Herausforderungen Erfolg und das nötige Glück. Dass Sie, Herr Scheffler, sich trotz Ihrer Verpflichtungen zwei Tage genommen haben, um uns durch die Thematik von Datenfusion und Datenintegration zu führen, ist eine Ehre für uns.

Meine Damen und Herren, ich wünsche Ihnen in diesem Jahr wieder einen interessanten Tagungsverlauf, spannende Diskussionen und anregende Gespräche am Rande der Veranstaltung.

Damit für diese Gespräche genügend Zeit und Raum zur Verfügung stehen, schließt der heutige Veranstaltungstag wieder mit einem „Get together", zu dem Sie alle direkt im Anschluss an die Veranstaltung im Innenhof unseres Hauses herzlich eingeladen sind. Mein Dank gilt dem ADM, der freundlicherweise auch in diesem Jahr wieder die Bewirtungskosten für dieses Get together übernimmt.

Ich darf nun Herrn Scheffler bitten, die Moderation zu übernehmen.

Datenfusion und Datenintegration
Machbar – wünschbar!?

Hartmut Scheffler

Befasst man sich mit Datenfusion und Datenintegration, so geht es dem Anschein nach „nur“ um technische Fragen und Softwarefragen der Verknüpfung und Verbindung unterschiedlicher Datensätze. Tatsächlich ist dies aber sozusagen die „Keimzelle“ unterschiedlicher (Forschungs-) Aktivitäten, die längst auch zu einem wichtigen Wirtschaftsfaktor geworden sind. Zu einem Wirtschaftsfaktor sui generis durch die direkt erzielten Umsätze und zu einem Wirtschaftsfaktor im Sinne der Professionalisierung und Ergebnisoptimierung der anwendenden Unternehmen.

Wer über Datenfusion und Datenintegration spricht und nachdenkt, befindet sich also automatisch inmitten einer umfassenden Wachstumsthematik. Gleichzeitig befindet man sich inmitten ganz unterschiedlicher Fragen der technischen und technologischen Möglichkeiten einerseits, der Anwendungen und Anwendungsmöglichkeiten zum Zweiten und ethischer Fragen zum Dritten.

Was ist mit der angebotenen Hardware und Software überhaupt möglich bzw. wird in wenigen Jahren möglich sein?

Zu welchem Nutzen für Anwender lassen sich diese Möglichkeiten einsetzen: Welche Geschäftsmodelle werden erfolgreich sein?

In vielen Fällen stecken hinter Datenfusion und –integration personenbezogene Daten z. B. aus Kundendatenbanken etc.: Werden dies Schritte – wenn überhaupt – hin zum gläsernen Bürger oder anders formuliert: Wie viel des Machbaren ist unter Datenschutzgesichtspunkten und letztendlich unter ethischen Gesichtspunkten überhaupt wünschbar?

Um deutlich zu machen, in welch umfangreichem Themenfeld Datenfusion und Datenintegration die zentrale „Technik“ darstellen, genügt ein Blick auf die wesentlichen Begriffe und Begriffsdefinitionen:

Business Intelligence (BI)	Sammelbegriff der Anwendungen und Technologien zur Sammlung, Sortierung, Analyse und Zurverfügungstellung von Daten zur Optimierung von Entscheidungsprozessen
OLAP (Online Analytical Processing)	Gehört wie Data-Mining zu den analytischen Informationssystemen, basiert auf Hypothesen, ermöglicht eine benutzerfreundliche Analyse der zugrunde liegenden Datenbasis (z. B. Data-Warehouse)
Data-Mining	Das systematische (automatisierte oder halbautomatisierte) Entdecken und Extrahieren unbekannter Informationen aus großen Datenmengen mit dem Ziel des Aufspürens von Regeln bzw. statistischen Auffälligkeiten
Datenfusion	Zusammenführung und Vervollständigung lückenhafter Datensätze und damit wichtiger Bestandteil der Informationsintegration
Predictive Analytics	Auf Basis anspruchsvoller statistischer Algorithmen, neuronaler Netzwerke etc. Vorhersage zukünftigen Verhaltens, zukünftiger Entwicklungen bezogen auf Kunden, Produkte, Dienstleistungen, Märkte etc.

Ein weites Feld, bei dem das Angebot die Nachfrage generieren und eine existierende Nachfrage (aufgrund der bereits jetzt vorhandenen großen und nicht vollständig gehobenen und (aus-)genutzten Datenmengen) neue Angebote schaffen wird. Hieran werden auch Misserfolgs-Zwischenphasen wie z. B. bei der Anwendung von CRM-Lösungen nichts ändern.

Ein schönes Beispiel für das Fortschreiten von Entwicklungen hin zu neuen, den Bedarfen in besonderer Weise entsprechenden Angeboten ist der Schritt von einfachem, deskriptivem Data-Warehousing über klassisches Data-Mining hin zu Predictive Analytics. Am Ende stehen neuartige Verfahren, die Anwendern wie z.B. im Bereich der Dienstleistungen tätigen Unternehmen bei der Objektivierung der Beantwortung ihrer Kernfragen helfen sollen: Nämlich der Vorhersage der Zukunft in Zeiten kontinuierlichen und auch diskontinuierlichen Wandels. Um Missverständnissen vorzubeugen: Auch die Deskription im Dataware-Housing ist bereits ein großer Schritt vorwärts und war für viele Anwender von enormer, wertschöpfender Bedeutung. Brachliegende oder zumindest nicht optimal durchanalysierte Datensätze wurden erstmals systematisch strukturiert und unter die Lupe genommen. Das Data-Mining sucht dann mit entsprechenden Verfahren die „Nuggets“ in diesen systematisierten und strukturierten Daten. Um im Bild zu bleiben: Manchmal mit großem Erfolg und hochinteressanten Erkenntnissen; manchmal nach Umwälzung riesiger Datenmengen ohne nennenswerte neue Erkenntnis. Das ganz große Versprechen ist nun aber die Hilfe bei der Gestaltung der Zukunft: Eben Predictive Analytics.

Data-Warehousing	Klassisches Data-Mining	Predictive Analytics
Frage- und Antwortfunktionen	Statistische Analyse	Neuronale Netzwerke etc.
Statische Perspektive	Kontinuierlicher Wandel	Auch diskontinuierlicher Wandel
Beschreiben von Gegenwart und Vergangenheit	Vorhersage der Vergangenheit	Vorhersage der Zukunft
Ableitung von Hypothesen	Prüfung von Hypothesen	Finden und Prüfen von Hypothesen

Die Behauptung, dass Datenfusion und Datenintegration im Mittelpunkt eines bereits bedeutenden Geschäftsfeldes und eines Wirtschaftsfaktors stehen, kann natürlich auch quantitativ belegt werden. Erstaunlich ist dabei allerdings, wie weit die Volumenschätzungen auseinander gehen. Wie groß der dahinter stehende Hardware-Software-Dienstleistungsmarkt also wirklich ist, lässt sich aus den verschiedenen Datenquellen nicht sagen. Nur eins: Sehr groß und sehr schnell wachsend.

Zunächst einmal geht man davon aus, dass die in Unternehmen anfallende Datenmenge um durchschnittlich 90 % pro Jahr wächst. Dies generiert ganz automatisch einen steigenden Bedarf. Dem entspricht ein steigendes Angebot: Z. B. bieten allein über 300 Softwareanbieter Produkte im Business Intelligence-Markt an.

Dies addiert sich dann laut IDC (International Data Corporation) zu einem Marktvolumen für Business Intelligence-Lösungen von 4,4 Mrd. $ in 2000. Es addiert sich gemäß OLAP-Report zu einem weltweiten OLAP-Markt von 4,3 Mrd. $ in 2004. Es lässt sich laut Metagroup darstellen als ein 830 Mio. € -Markt (in 2004) für Business Intelligence in Deutschland. Diese Zahlen machen unzweifelhaft deutlich, dass sehr vieles bereits machbar ist und gemacht wird. Damit stellt sich dann auch zwingend die Frage, ob all dies wünschbar ist.

Um tiefer in Fragen der Machbarkeit und damit der Nutzanwendung einerseits und der ethischen Wünschbarkeitsfragen andererseits einzusteigen, eignet sich in idealer Weise der „Klassiker“ dieser Jahre: Data-Mining.

Fast alle Dienstleistungsanbieter, sehr viele Anbieter langlebiger Gebrauchsgüter und immer mehr Anbieter klassischer Konsumgüter besitzen große Mengen an Daten über Kunden, deren Einkaufsverhalten und Kaufbiographie etc. Eine Studie des Lehrstuhls für Wirtschaftsinformatik an der Katholischen Universität Eichstätt-Ingolstadt hat jüngst ermittelt, dass innerhalb der 500 größten Unternehmen in Deutschland diese Daten nur von jedem zweiten Unternehmen genutzt werden.

Die Unternehmen wollen dies ändern: Jedes 5. möchte in nächster Zeit die vorliegenden Kundendaten mit Hilfe von Data-Mining analysieren.

Dies geschieht mit dem Ziel optimaler Kundensegmentierung, Zielgruppenanalysen und Kundenpotential-Analysen. Dies geschieht weiter mit dem Ziel, Kunden individuell „kennenzulernen“ und dann individuell und optimal ansprechen zu können (1-to-1-Marketing). Im Endeffekt sollen Data-Mining-Einsätze den gesamten Marketingprozess von der Produktoptimierung über die Preisfindung und die Kommunikation bis hin zu Kundenbindungsmaßnahmen optimierend beeinflussen. Der Reiz dieses Vorgehens und nicht zuletzt auch die Notwendigkeit im Konkurrenzwettbewerb sind evident. Ebenso eindeutig ist aber, dass aus gesammelten Daten (seien es automatisch anfallende Bewegungs- und Transaktionsdaten, seien es Befragungsdaten) Schlüsse gezogen werden, die unmittelbar in das Marketing diesen (individuellen) Kunden gegenüber einfließen (also in die angebotenen Produkte, die Bindungsmaßnahmen etc.), ohne dass die Kunden bzw. Kundenpotentiale hiervon wissen.

Dies ist nicht wirklich neu: Auch wer früher aus einfachen Kreuztabellen als Ergebnis von Umfragen Schlussfolgerungen zog im Hinblick darauf, wem welche Produkte wo zu welchen Konditionen angeboten werden, machte letztlich nichts anderes (und dies letztlich sicherlich zum Wohle von Absender und Adressat). Der Unterschied: Mit den neuen Verfahren geschieht dies systematischer, intelligenter und umfassender.

Zusammengefasst ergibt dies einen Status 2005 etwa der folgenden Art:

Die Themenfelder und Ansätze werden insgesamt klarer und strukturierter. Dem entspricht ein umfassendes Angebot an Hardware, Software, Dienstleistungen zu Business Intelligence im Allgemeinen, Themen wie Datenfusion und Datenintegration im Speziellen. Trotz sehr unterschiedlicher Markt-Volumensschätzungen handelt es sich bereits aktuell um einen relevanten Wirtschaftsfaktor, dessen Dynamik durch einen sich selbst verstärkenden Angebot-Nachfrage-Angebot-Prozess garantiert ist. Damit sind speziell Datenfusion und Datenintegration nicht nur aus der angewandten wissenschaftlichen Forschung nicht mehr wegzudenken, sondern auch aus der tagtäglichen unternehmerischen Anbieter- und Anwenderpraxis.

12 Anmerkungen - manche davon bereits „bewiesene“ Tatsache, manche eine These - mögen das Thema rund um die Stichworte Nachfrage, Machbarkeit, Wünschbarkeit vertiefen und die Diskussion anregen:

1. Der Umfang verfügbarer Daten, die Informationsmenge, nimmt „automatisch“ aufgrund technischer Entwicklungen wie auch gezielt durch entsprechende Forschungsansätze zu.
   - Beispiel GPS (Global Positioning System) = Technik
     Reichweitenforschung „Außenwerbung“ = Forschung
   - Beispiel 2: RFID (Radio Frequency Identification) = Technik
     Neue Ansätze zur Leserschaftsforschung? = Forschung

2. Immer mehr Daten steigern das Gefühl, immer weniger zu wissen bzw. viel mehr wissen zu können.
   - Wo Datenmengen zunehmen, wird Vernetzung zum Thema.
   - Wo die Datenvernetzung steigt, steigt auch die Komplexität.
   - Gesucht sind Ansätze der intelligenten Verknüpfung und Vernetzung, d.h. eine intelligente Verdichtung zur Komplexitätsreduzierung.
3. Die statistischen Verfahren zu Data-Matching, Datenfusion und Datenintegration verfeinern sich.
   - Beispiel: Just in Time-Profiling (z.B. Value Profiler von TNS Infratest)
4. Im Zeitalter der Segmentierung und Fragmentierung von Menschen, Medien, Marken steht das Marketing vor der Aufgabe, Produkte immer spitzer zu positionieren, Zielgruppen immer genauer zu definieren, Kommunikationsstrategien immer effizienter zu gestalten. Hierzu bedarf es neuer Informationen und Erkenntnisse aus der Datenfusion und Datenintegration existierender Daten.
   - Beispiel: Verknüpfte Daten zu Motiven und Wertestrukturen, Mediennutzung, Kauf- und Verwendungsdaten.
5. Der klassische Marketingprozess umfasst u. a. das Marktverständnis, die Produktentwicklung, die Kommunikation und das Kundenbindungs-Management.
   - Das „Ideal“: der 360-Grad-Ansatz als integrierter Forschungsprozess (der Single Source-Datensatz).
   - Dies ist unrealistisch: Datenfusion und Datenintegration sind der Lösungsweg hin zu ganzheitlichen Betrachtungsmöglichkeiten.
6. Aus Massenkommunikation wurde Zielgruppenkommunikation, aus Zielgruppenkommunikation immer mehr 1:1-Kommunikation. Die Kommunikationsforschung ist einer der Treiber von Datenfusion und Datenintegration.
   - Beispiel: Optimierung Direktmarketing
7. Kundenbindung ist zu *dem* Erfolgsfaktor geworden. CRM wurde vom Schlagwort zur notwendigen Marketing- und Vertriebspraxis (trotz zahlloser Flops).
   - *Der klassische Prozess*:
     Die Erkenntnisse aus kleinen Stichproben mit großer Datenmenge und hoher Datentiefe werden verknüpft mit großen Datenbanken (Kundendatenbanken) von geringer Datentiefe: Modelling und Profiling
   - CRM wird damit zu einem weiteren großen Treiber von Datenfusion und Datenintegration.

8. Appetit kommt beim Essen: Es gibt in der Praxis erfolgreiche Beispiele von Datenfusion und Datenintegration. Erfolgreiche Beispiele schaffen Verlangen nach mehr.
   - *Beispiel 1*:
     Move (Verknüpfung von Mediadaten mit Haushalts-Paneldaten)
   - *Beispiel 2:*
     Printmodell der ag.ma: Statt 26.000 Interviews mit je 180 abgefragten Titeln jetzt 39.000 Interviews mit jeweils nur 120 Titeln: Die jeweils fehlenden 60 Titel werden fusioniert.
   - *Beispiel 3:*
     Die zahllosen Beispiele aus der Mikrogeographie
9. Erfolgreiches Arbeiten mit Datenfusion und Datenintegration wird zu einem Unternehmens-Erfolgsfaktor und damit zu einem Wirtschaftsfaktor!
   - Dies erhöht den Druck auf die Anbieter.
   - Was möglich ist, wird gemacht: Gefahren für den Datenschutz!
10. Datenschutz, d. h. unter anderem Anonymisierungsgebot und Anonymisierungspflicht stehen über allem. De-Anonymisierung und (individualisierter) gläserner Bürger lassen sich nur durch entsprechende Gesetze und Standesregeln verhindern. Akzeptanz von Datenfusion und Datenintegration setzt klare Regeln und damit Verhaltenssicherheit voraus.
    - Vorbildliche Selbstregulierung in Deutschland durch BVM, ADM, ASI
11. Im Zeichen der Globalisierung werden Marketing wie auch Fusionstechniken und Integrationstechniken und nicht zuletzt der Bedarf danach immer internationaler. Dem müssen Gesetze und Regeln folgen.
    - Selbstregulierung z. B. durch ESOMAR/WAPOR
12. *Tatsache 1:*
    Wachsende Machbarkeit durch immer bessere Verfahren auf Anbieterseite.
    Tatsache 2:
    Stark wachsende Nachfrage = wünschbar auf Nachfragerseite.
    - Dies führt zwingend zur Frage der Wünschbarkeit aus gesellschaftspolitischer Sicht.
    - Die Lösung: Klare Regeln!

Datenfusion und Datenintegration: Machbar – wünschbar!?

Die 6. Wissenschaftliche Tagung des Statistischen Bundesamtes behandelt das Thema, indem Techniken der Datenfusion und Datenintegration, Anwendungsbeispiele wie auch Rahmenbedingungen (Datenschutz und Standesregeln) thematisiert und vertieft werden.

Techniken und Einsatzgebiete von Datenintegration und Datenfusion

Hans Kiesl, Susanne Rässler

1 Einleitung

Unter Datenintegration und Datenfusion versteht man Techniken, Datensätze aus mindestens zwei verschiedenen Erhebungen mit teilweise nicht identischen Variablenmengen so zu verknüpfen, dass jeder Beobachtung der einen Erhebung Daten derselben Beobachtungseinheit (bei Datenintegration) oder einer „ähnlichen“ (bei Datenfusion) aus den anderen Erhebungen hinzugefügt werden. Während es bei der Datenintegration fehlerhafte Werte in den Schlüsselvariablen sind, die den Verknüpfungsprozess schwierig gestalten, kommt bei der Datenfusion eine implizite, nicht überprüfbare Modellvorstellung über Zusammenhänge zwischen den nicht gemeinsam beobachteten Variablen hinzu. Im vorliegenden Beitrag werden diese implizite Annahme noch einmal verdeutlicht und verschiedene Gütekriterien einer Datenfusion diskutiert; außerdem wird eine alternative Vorgehensweise skizziert, die mit Hilfe multipler Ergänzung die Notwendigkeit dieser Annahme zu überwinden versucht.

2 Datenintegration

Datenintegration (auch „Record Linkage“) bezeichnet die Verknüpfung von Datensätzen aus verschiedenen Quellen, die jeweils zu denselben Objekten (Haushalte, Personen, Firmen) gehören. Ohne Vorliegen von fehlenden oder fehlerhaften Werten in den zur Zusammenführung verwendeten Schlüsseln (z.B. Name und Adresse, Sozialversicherungsnummer) ist diese Verknüpfung trivial. In der Realität hat man aber stets mit unterschiedlichen Schreibweisen (wie z.B. ä und ae) und Tipp- oder Übertragungsfehlern zu kämpfen, so dass im Allgemeinen eine Ähnlichkeitssuche zwischen den Schlüsseln verschiedener Datensätze durchgeführt werden muss. Der Ablauf eines Datenintegrationsprozesses besteht im Wesentlichen aus folgenden Schritten (Schnell et al. (2005):

- Bereitstellung der zu verknüpfenden Datensätze
- Standardisierung der Verknüpfungsschlüssel
- Berechnung der Ähnlichkeiten der potentiellen Paare
- Manuelle Verknüpfung ungeklärter Fälle

- Tatsächliche Zusammenführung der Datensätze.

Der erste und letzte Schritt sind jeweils trivial; problematisch sind die übrigen Schritte, die durch Fehler in den Schlüsselvariablen verursacht sind und im Allgemeinen einen hohen Aufwand zur Erstellung einer optimalen Verknüpfungsstrategie erfordern. Im Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit wurden in den letzten Jahren erfolgreich mehrere Datenintegrationen durchgeführt, die neue Informationsquellen für die Arbeitsmarktforschung erschließen; beispielhaft seien hier LIAB, IABS und IEB genannt.

Der Linked-Employer-Employee-Datensatz im IAB (LIAB) verknüpft Personenangaben aus der Beschäftigtenstatistik der Bundesagentur für Arbeit mit Betriebsangaben aus dem IAB-Betriebspanel. Das IAB-Betriebspanel enthält Informationen über betriebliche Strukturen und personalpolitische Entscheidungen über einen Zeitraum von 1993 bis zum aktuellen Rand. Über die Verknüpfung der Betriebsinformationen mit den Erwerbsverläufen sozialversicherungspflichtig Beschäftigter können Erwerbsbiografien im Kontext betrieblicher Variablen, die z.B. über die Beschäftigtenstruktur oder die Charakteristika der Beschäftigungsverhältnisse Auskunft geben, analysiert werden; vgl. Bellmann et al. (2002).

Die IAB-Beschäftigtenstichprobe (IABS) bietet tagesgenaue erwerbsbiografische Daten für ein Prozent (in der aktuellen Version: zwei Prozent) der sozialversicherungspflichtig Beschäftigten. In der IABS sind Daten aus zwei unterschiedlichen Quellen miteinander verknüpft: zum einen Beschäftigungsinformationen aus den Meldungen der Arbeitgeber an die Sozialversicherungsträger und zum anderen Daten über den Bezug von Arbeitslosengeld, -hilfe oder Unterhaltsgeld aus der Geschäftsstatistik der Bundesagentur für Arbeit; vgl. Bender/Haas (2002).

Die Datenbank der Integrierten Erwerbsbiografien des IAB (IEB) verknüpft Auszüge aus vier unterschiedlichen Quellen, wodurch eine entsprechend umfassende Darstellung individueller Erwerbsverläufe möglich ist. Diese vier Quellen sind die Beschäftigten-Historik des IAB, die Leistungsempfänger-Historik des IAB, die Maßnahme-Teilnehmer-Gesamtdatenbank und die BA-Geschäftsdatenbasis aus dem Bewerberangebot. Zum Aufbau dieser Datenbank vgl. Kruppe/Oertel (2003).

3 Datenfusion

Im Rahmen der Datenfusion werden Datensätze aus verschiedenen Quellen miteinander verknüpft, die Informationen über unterschiedliche Objekte (Haushalte, Personen, Firmen etc.) enthalten. Während also bei der Datenintegration ge-

trennt erhobene Informationen über identische Objekte zusammengeführt werden, müssen bei der Datenfusion Daten von ähnlichen, aber im Regelfall nicht identischen Einheiten verknüpft werden. Rodgers (1984) beschreibt diese formal analoge, inhaltlich aber entscheidend andere Vorgehensweise sehr treffend:

> „It is a relatively small step, computationally, from such procedures for exact matching of identical individuals to statistical matching of similar individuals. A small step for the computer is in this case a giant step for the statistician – a step that should only be taken with full awareness of the importance of the implicit assumptions and the potential consequences of the incorrectness of those assumptions."

3.1 Traditionelle Verfahren der Datenfusion

Der traditionelle Ansatz der Datenfusion geht zunächst von zwei Datensätzen aus, wobei bestimmte Variablen (typischerweise regionale und soziodemographische Merkmale wie Geschlecht, Alter oder Bildung, teilweise aber auch um zusätzliche Merkmale angereichert wie erinnertes Kauf- oder Fernsehverhalten) in beiden Datensätzen vorhanden sind (gemeinsame und sog. spezifische gemeinsame Variablen), während weitere Variablen wie etwa extensiv gemessene Fernseh- oder Kaufinformationen nur in jeweils einem Datensatz vorliegen (spezifische Variablen). Meist ist das weitere Vorgehen asymmetrisch in dem Sinne, dass in einem bestimmten Datensatz (sog. Empfänger- oder Rezipientenstichprobe) die nicht beobachteten Variablen ergänzt werden sollen, die nur im anderen Datensatz (sog. Spenderstichprobe) vorhanden sind. Somit kann die Datenfusion als spezielles Problem fehlender Daten beschrieben werden. Abbildung 1 zeigt dabei die Abgrenzung der Datenfusion von anderen Mustern fehlender Daten.

Bild 3 in Abbildung 1 verdeutlicht die spezifische Situation der Datenfusion. Es gibt Variablengruppen X und Y, die jeweils nur bei einem Teil der Datensätze beobachtet wurden, während die Variablengruppe Z in allen Beobachtungen vorhanden ist. Als besonderes Kennzeichen der Datenfusion soll festgehalten werden, dass die spezifischen Variablen nie zusammen beobachtet werden. In den anderen dargestellten Situationen liegen alle Paare von Variablen bei einem bestimmten Teil der Beobachtungen gemeinsam vor.[1] Es ist leicht einsichtig,

1 Bild 1 zeigt den allgemeinen Fall einer unregelmäßigen Struktur fehlender Werte; Bild 2 stellt die Situation dar, dass bestimmte Variablen bei einem Teil der Beobachtungen vollständig fehlen (z.B. designbedingt, wie etwa bei den Ergänzungsmodulen im Mikrozensus, die sich nur an eine Unterstichprobe aller Befragten wenden). Bild 4 zeigt eine Variante des Split Questionnaire Survey Design, bei dem ein Fragebogen in disjunkte Fragenblöcke aufgeteilt wird. Nur einer dieser Blöcke wird allen Befragten vorgelegt, während alle anderen Blöcke nur jeweils an bestimmte Gruppen von Befragten gehen; allerdings erfolgt die Aufteilung so, dass jedes mögliche Paar von Blöcken bei irgendeiner Gruppe gemeinsam

dass im Falle der Datenfusion ein Schätzproblem auftritt. Da die spezifischen Variablen X und Y nicht gemeinsam beobachtet werden, kann ihr Zusammenhang auch nicht direkt aus den Daten geschätzt werden, dies wird als das Identifikationsproblem der Datenfusion bezeichnet.

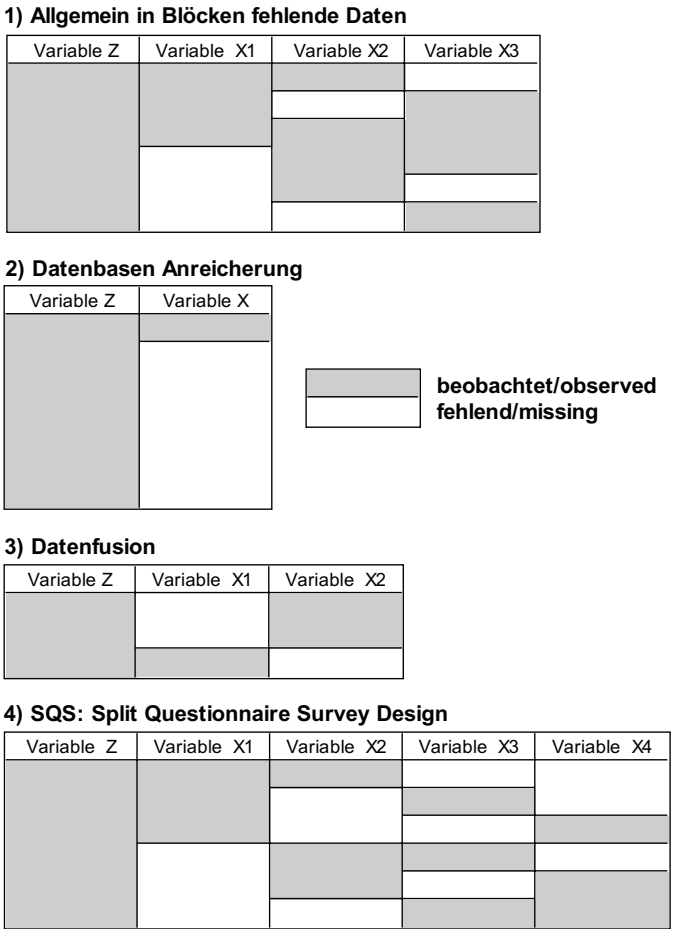


Abb. 1: Verschiedene Muster fehlender Werte in einem Datensatz

Fortsetzung Fußnote 1
erhoben wird. Auf diese Weise wird das Identifikationsproblem der Datenfusion, dass bestimmte Variablen (hier: Fragebogenitems) nie gemeinsam beobachtet werden und sich ihre Korrelation somit nicht eindeutig schätzen lässt, verhindert. Gleichzeitig kann der Aufwand für die Befragten erheblich gesenkt werden, da nicht alle den vollständigen Fragebogen beantworten müssen.

Die Algorithmen der Datenfusion, d.h. die Regeln, die jeder Beobachtung der Empfängerdatei eine Beobachtung der Spenderdatei zuordnen, unterscheiden sich im Detail, aber nicht prinzipiell. Für jede Beobachtung der Empfängerdatei wird dabei eine Beobachtung aus der Spenderdatei mit denselben Werten in den gemeinsamen Variablen Z gesucht; gibt es mehrere solche Beobachtungen, wird zufällig eine davon ausgewählt. Die Werte der spezifischen Variablen Y dieser „Spenderbeobachtung“ werden in die Empfängerdatei übertragen. Im konkreten Anwendungsfall kann es vorkommen, dass eine Beobachtung der Empfängerdatei mit keiner Beobachtung der Spenderdatei in allen gemeinsamen Variablen Z übereinstimmt; in diesen Fällen wird nach einer Spenderbeobachtung gesucht, die sich (im Sinne eines bestimmten Distanzmaßes) möglichst wenig in den gemeinsamen Variablen Z unterscheidet („Nearest-Neighbor-Matching“). Man spricht auch allgemein von der Suche nach „statistischen Zwillingen“.

Abbildung 2 zeigt ein vereinfachtes Beispiel für eine Datenfusion. Aus einer Stichprobe, mit der das Fernsehverhalten einer bestimmten Gruppe von Personen aufgezeichnet wurde, und einer Stichprobe, in der das Verbraucherverhalten einer anderen Gruppe dokumentiert ist, soll eine gemeinsame Datei erzeugt werden, in der Informationen über beide Variablen enthalten sind. Anhand der gemeinsamen Variablen Geschlecht und Alter wird für jede Zeile der Empfängerstichprobe (Fernsehpanel) eine Beobachtung der Spenderstichprobe (Verbraucherpanel) gesucht, die in den gemeinsamen Variablen Geschlecht und Alter vollständig oder näherungsweise übereinstimmt.

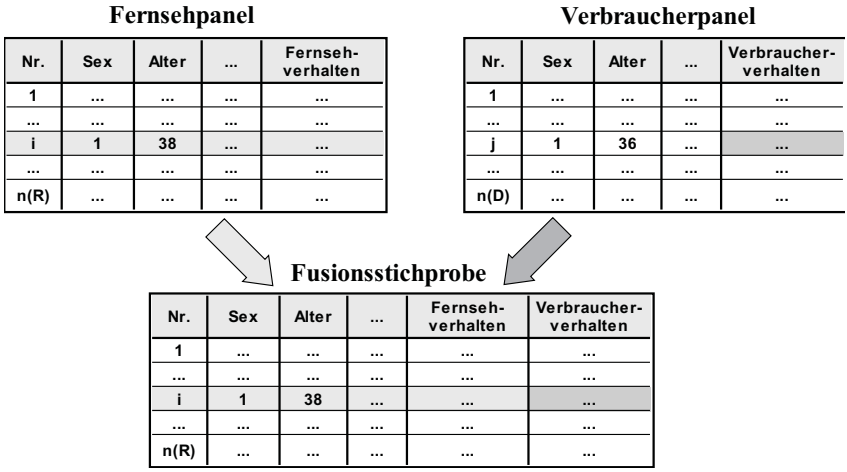


Abb. 2: Prinzip des „Nearest-Neighbor-Matching”

In der Praxis sind verschiedene Verfahren im Einsatz, die historisch unterschiedliche Wurzeln haben. In den USA und Kanada haben Verfahren der Datenfusion (unter dem Begriff „Statistical Matching“) vor allem im Rahmen der amtlichen Statistik eine lange Tradition. Die Fülle der einzelnen Algorithmen lässt sich im Wesentlichen in drei Klassen einteilen (vgl. Rodgers (1984) und Rubin (1986)). Im Rahmen des uneingeschränkten Matching (auch „Generalized Distance Method“) können Beobachtungen aus dem Spenderdatensatz beliebig oft mit einer Beobachtung im Empfängerdatensatz verknüpft werden; beim eingeschränkten Matching (auch „Generalized Rank Weight-Split Method“) ist diese Möglichkeit beschränkt (im Extremfall auf höchstens einen Empfänger je Spender), um die Randverteilungen aus dem Spenderdatensatz möglichst gut im fusionierten Datensatz zu reproduzieren. Verfahren des „Categorically Constrained Matching“ schließlich benutzen Zusatzinformationen aus externen Quellen; vgl. dazu Liu/Kovacevic (1998).

Datenfusionen in Deutschland waren lange Zeit auf den Bereich der Marktforschung und Mediaplanung beschränkt, wo immer noch der größte Bedarf an Fusionen zu verzeichnen ist. Grundlegend war dabei das so genannte topologische Konzept nach Wendt (1976, 1980, 1986), das wesentlich auf linearer Programmierung und Algorithmen der Clusteranalyse basiert. (Zu einem ausführlichen historischen Abriss vgl. Rässler (2002), Kapitel 3.)

3.2 Verteilungsanalyse der Datenfusion

Um deutlich zu machen, was bei einer Datenfusion überhaupt passiert, d.h. wie sich die Verteilungen von Variablen im Laufe eines Fusionsprozesses verändern, müssen zunächst einige Begrifflichkeiten näher definiert werden. Mit $f_{X,Y,Z}(x,y,z)$ sei die (unbeobachtete) gemeinsame Verteilung der Variablengruppen X, Y, Z bezeichnet. Im Empfängerdatensatz wird $f_{X,Z}(x,z)$, die gemeinsame Verteilung von X und Z beobachtet, im Spenderdatensatz $f_{Y,Z}(y,z)$, die gemeinsame Verteilung von Y und Z. Streng genommen sind die beiden Datensätze nur Stichproben aus der Grundgesamtheit; im Folgenden wird diese Unterscheidung, die nur bei kleinen Datensätzen eine wichtige Rolle spielt, zunächst vernachlässigt, d.h. es wird von dem Fall idealer Teilstichproben ausgegangen.

Die Fusion wird nun vorgenommen, indem jeder Empfängerbeobachtung ein Spender mit demselben Wert der gemeinsamen Variablen Z zugeordnet wird; bei mehreren möglichen Spendern erfolgt die Auswahl uneingeschränkt und zufällig. Ergebnis der Fusion ist eine ergänzte Empfängerdatei, die als Stichprobe aus einer hypothetischen Gesamtheit (Fusionsstichprobe) mit Fusionsverteilung $f_{X,Y,Z}(x, y, z)$ betrachtet werden kann.

In dieser idealen Situation lassen sich wichtige Eigenschaften der Fusionsverteilung herleiten; zum genauen Vorgehen vgl. Rässler (2002); insbesondere gilt Folgendes:

- Die Randverteilung von Y sowie die gemeinsame Verteilung von Y und Z werden in der Fusionsstichprobe exakt reproduziert.
- Für die gemeinsame Verteilung von X und Y gilt:

$$\widetilde{f}_{X,Y}(x,y) = \int f_{X|Z}(x \mid z) \cdot f_{Y|Z}(y \mid z) \cdot f_Z(z)dz$$

  Dies ist genau dann identisch mit $f_{X,Y}(x,y)$, wenn

$$f_{X|Z}(x \mid z) \cdot f_{Y|Z}(y \mid z) = f_{X,Y|Z}(x,y \mid z)$$

  für alle z gilt, d.h. wenn X und Y bedingt unabhängig sind, gegeben Z. Nur in diesem Fall stimmt also die gemeinsame Verteilung von X und Y nach der Fusion mit der tatsächlichen (unbekannten) Verteilung überein.
- Für die Korrelation zwischen X und Y gilt:

$$\widetilde{\mathrm{Cov}}(X,Y) = \mathrm{Cov}(X,Y) - \mathrm{E}(\mathrm{Cov}(X,Y \mid Z))$$

  d.h. die Kovarianz nach der Fusion entspricht der tatsächlichen (unbekannten) Kovarianz abzüglich des Erwartungswertes der bedingten Kovarianz zwischen X und Y (gegeben Z).

  Aus der bekannten Kovarianzzerlegung

$$\mathrm{Cov}(X,Y) = \mathrm{E}(\mathrm{Cov}(X,Y \mid Z)) + \mathrm{Cov}(\mathrm{E}(X \mid Z), \mathrm{E}(Y \mid Z))$$

  lässt sich folgern, dass die Kovarianz in der fusionierten Stichprobe genau dann mit der tatsächlichen Kovarianz identisch ist, wenn

$$\mathrm{E}(\mathrm{Cov}(X,Y \mid Z)) = 0$$

  d.h. wenn die bedingte Kovarianz von X und Y im Mittel verschwindet, wenn also X und Y im Durchschnitt bedingt unkorreliert sind (gegeben Z). Diese Bedingung ist schwächer als die bedingte Unabhängigkeit, d.h. die Kovarianzen können selbst dann korrekt in der Fusionsstichprobe abgebildet sein, wenn die gemeinsamen Verteilungen nicht richtig wiedergegeben werden.

Tatsächliche Stichproben weichen aufgrund des Zufallsfehlers oder des Stichprobendesigns von einer idealen Stichprobe ab, daher gelten die genannten Aussagen i.A. nur näherungsweise. Simulationsstudien haben aber gezeigt, dass die beschriebenen Beziehungen auch für die Ergebnisse von „Nearest-Neighbor-Verfahren“ gelten. Grundsätzlich reproduzieren Fusionsverfahren die gemeinsame Verteilung von X und Y also nur, wenn diese Variablengruppen in der Gesamtheit bedingt unabhängig sind (gegeben Z). Die Korrelationen zwischen X und Y werden nur dann korrekt wiedergegeben, wenn X und Y im Durchschnitt bedingt unkorreliert (gegeben Z) sind. Angesichts der Tatsache, dass ein Haupteinsatzgebiet der fusionierten Stichproben darin liegt, unbeobachtete Korrela-

tionen zwischen X und Y zu schätzen, um daraus Handlungsanweisungen abzuleiten oder zumindest zu unterstützen, wird hier ein schwerwiegendes Problem der Datenfusion offenbar, das so genannte Identifikationsproblem: die Beziehung zwischen den Variablen, die nie gemeinsam beobachtet werden, lässt sich aus der Stichprobe nicht schätzen. Herkömmliche Fusionsalgorithmen stellen immer bedingte Unabhängigkeit her, obwohl diese Annahme im besten Fall nicht überprüfbar, im schlechtesten Fall unhaltbar ist.

Abbildung 3 illustriert dies anhand eines fiktiven Beispiels von Raetzel (2000) mit nur drei Variablen. In der oberen Tabelle sind die tatsächlichen Beziehungen zwischen dem Alter (eine typische gemeinsame Variable), dem Fernsehverhalten (Konsum bestimmter Werbung) und dem Verbraucherverhalten (Kauf von Knoblauchpillen) dargestellt.

Originäre Beziehung von Kauf- und Fernsehverhalten

Kauf von Knoblauchpillen	Kinder		Ältere Leute		Summe
	Werbung: ja	Werbung: nein	Werbung: ja	Werbung: nein	
Ja	5	0	50	10	65
Nein	95	100	50	90	335
Summe	100	100	100	100	400

Verbraucherpanel

Kauf von Knoblauchpillen	Kinder	Ältere Leute	Summe
Ja	5	60	65
Nein	195	140	335
Summe	200	200	400

Fernsehpanel

Werbung gesehen	Kinder	Ältere Leute	Summe
Ja	100	100	200
Nein	100	100	200
Summe	200	200	400

Durchschnittliche künstliche Fusionsstichprobe aus 1000 Simulationen

Kauf von Knoblauchpillen	Kinder		Ältere Leute		Summe
	Werbung: ja	Werbung: nein	Werbung: ja	Werbung: nein	
Ja	2.49	2.51	29.92	30.03	64.95
Nein	97.51	97.49	70.08	69.97	335.05
Summe	100.00	100.00	100.00	100.00	400.00

Abb. 3: Bedingte Unabhängigkeit nach der Datenfusion

Um die Fernsehwerbung auf das Verbraucherverhalten abzustimmen, ist man an der gemeinsamen Verteilung der Variablen interessiert, hat aber zunächst nur die Teilstichproben und damit die Randverteilungen aus den beiden unterschiedlichen Erhebungen vorliegen, in denen nicht alle Variablen gemeinsam beobachtet wurden (mittlere Tabellen). Im dargestellten Beispiel sind dies ein Verbraucherpanel (mit den Variablen Alter und Verbraucherverhalten) und ein Fernsehpanel (mit den Variablen Alter und Fernsehverhalten).

Mit den beiden zugrunde liegenden Datensätzen wurde ein typisches „Nearest-Neighbor-Matching“ tausendmal simuliert und die durchschnittliche Randverteilung der fusionierten Stichprobe berechnet (untere Tabelle). Offensichtlich weicht die Verteilung der fusionierten Stichprobe deutlich von der Ausgangsverteilung ab. Der Matching-Algorithmus hat bedingte Unabhängigkeit hergestellt, d.h. die Variablen „Kauf von Knoblauchpillen“ und „Werbung im Fernsehen gesehen“ sind in jeder durch das Merkmal Alter definierten Teilgruppe (unter den Kindern wie unter den älteren Personen) unabhängig voneinander. Wie man der oberen Tabelle entnehmen kann, sind diese Merkmale tatsächlich aber nicht unabhängig, auch nicht innerhalb der gegebenen Altersklassen.

Wie dieses Beispiel zeigt, liefert eine Datenfusion nicht immer die Ergebnisse, die gewünscht sind. Vor einer weiteren Diskussion soll noch geklärt werden, was die unterschiedlichen Ziele einer Fusion sein können, d.h. auf welche unterschiedlichen Aspekte des fusionierten Datensatzes Wert gelegt werden kann.

3.3 Validitätsstufen der Datenfusion

3.3.1 Erste Ebene: Erhalt der individuellen Werte

Das maximal erreichbare Ziel der Datenfusion wird häufig fälschlicherweise darin gesehen, dass die in der Empfängerdatei ergänzten Werte der Variablen Y den tatsächlichen, wahren Werten der jeweiligen Objekte (Personen, Haushalte, Firmen) entsprechen. Man könnte in einem solchen Fall bei jedem reproduzierten Wert von einem Treffer sprechen und die Zahl der Treffer (d.h. der exakt geschätzten Werte von Y) im gesamten Datensatz zu maximieren versuchen.

Dieses Ziel anzustreben oder die Güte der Fusion an diesem Ziel zu messen, ist aus mehreren Gründen unangebracht. Zum einen ist bei stetigen Variablen die Trefferwahrscheinlichkeit streng genommen Null, bei diskreten Variablen mit vielen Ausprägungen zumindest sehr klein, im multivariaten Fall mit vielen Y-Variablen verschwindend gering. Andererseits ist es auch nicht sinnvoll, das Hauptaugenmerk der Fusion auf die Anzahl der Treffer zu legen, weil eine höhere Trefferzahl nicht unbedingt zu einer realistischeren Fusionsverteilung führt, wie das folgende Beispiel zeigt.[2]

Zu ergänzen sei im Empfängerdatensatz eine binäre Variable Y, die im Spenderdatensatz kaum Korrelation mit den gemeinsamen Variablen Z zeigt. Die Häufigkeit von Y=1 im Spenderdatensatz sei 60%. Eine Fusionsstrategie könnte nun darin bestehen, im Empfängerdatensatz für Y generell den Wert 1 zu setzen; man erreicht damit eine Trefferrate von ungefähr 60%, hat aber die Randverteilung von Y völlig verfälscht. Setzt man dagegen Y in jeder Empfängerbeobachtung zufällig mit einer Wahrscheinlichkeit von 60% auf den Wert 1, wird man

2 Dieses Beispiel wurde in etwas abgewandelter Form von Donald B. Rubin im Rahmen eines Vortrags auf der Data Quality Konferenz Q2004 in Mainz 2004 präsentiert.

am Ende die Randverteilung (und wohl auch die gemeinsame Verteilung) sehr gut widerspiegeln; die zu erwartende Trefferrate ist aber nur 0,6·0,6+0,4·0,4 = 52%. Offenbar führt eine Orientierung an der Trefferrate hier zu einem erkennbar suboptimalen Fusionsergebnis.

Die Irrelevanz dieser Validitätsebene ist auch leicht einzusehen, wenn man sich die Gründe vergegenwärtigt, die überhaupt zu einer Fusion führen: man ist an statistischen Analysen des fusionierten Datensatzes interessiert (Zusammenhänge, Regressionen, Clusteranalysen), aber nicht an der Rekonstruktion der individuellen Beobachtungen auf Mikroebene. Letzteres käme Zauberei gleich.

3.3.2 Zweite Ebene: Erhalt der gemeinsamen Verteilung

Könnte man diese Ebene erreichen, würde $\tilde{f}_{x,y,z}(x, y, z) = f_{x,y,z}(x, y, z)$ (für alle x,y,z) gelten, d.h. die gemeinsame Verteilung der Variablen in der fusionierten Stichprobe entspräche der wahren gemeinsamen Verteilung. In diesem Fall wären alle statistischen Analysen, die man mit der Fusionsstichprobe durchführt, valide, weil sie von der „richtigen" Verteilung ausgehen. Wie oben dargestellt, setzt die korrekte Reproduktion der gemeinsamen Verteilung die bedingte Unabhängigkeit der spezifischen Variablen, gegeben die gemeinsamen Variablen, voraus. Obgleich diese Validitätsebene im Prinzip erstrebenswert ist, lässt sich ohne Zusatzinformationen nicht testen, ob eine gegebene Fusion dieses Ziel erreicht bzw. wie weit die durchgeführte Fusion von diesem Ziel abweicht. Für ein bestimmtes Fusionsverfahren kann lediglich im Rahmen einer Simulationsstudie mit realistisch gewählten gemeinsamen Verteilungen evaluiert werden, inwieweit diese Verteilungen nach der Fusion reproduziert wurden.

Eine theoretische Eingrenzung der wahren gemeinsamen Verteilung lässt sich aus den Randverteilungen der einzelnen Variablen über die so genannten Fréchet-Hoeffding-Grenzen gewinnen; vgl. Ridder/Moffitt (2005):

$$\max\left\{0, \sum_{i=1}^{q} F_{X_i}(x_i) - (q-1)\right\} \le F_{X_1,\ldots,X_q}(x_1,\ldots,x_q) \le \min\left\{F_{X_1}(x_1),\ldots,F_{X_q}(x_q)\right\}$$

dabei ist q die gesamte Anzahl der Variablen, F_{x_i} ist die Verteilungsfunktion der Variable X_i (Randverteilung), $F_{x_1,\ldots,x_q}$ ist die gemeinsame Verteilungsfunktion der q Variablen $X_1,\ldots,X_q$. Erfahrungsgemäß schränken diese Grenzen (zumal bei sehr vielen Variablen) die mögliche gemeinsame Verteilung aber nur unzureichend ein.

3.3.3 Dritte Ebene: Erhalt der Korrelationsstruktur

Häufigster praktischer Zweck der Datenfusion sind Aussagen über Korrelationen zwischen bestimmten Variablen (also z.B. zwischen dem Fernsehverhalten und den Konsumgewohnheiten). Daher ist es nahe liegend, auf den genauen Erhalt der gemeinsamen Verteilung weniger Wert zu legen als auf die Reproduktion der Korrelationsstruktur, insbesondere auf die richtige Schätzung der Korrelation zwischen den Variablengruppen X und Y, die nie gemeinsam beobachtet werden. Wie oben begründet, wird mit den üblichen Fusionsalgorithmen die tatsächliche Korrelation genau dann exakt in der Fusionsstichprobe wiedergegeben, wenn X und Y im Durchschnitt bedingt unkorreliert (gegeben Z) sind, eine Eigenschaft, die mit den vorliegenden Daten wiederum nicht überprüft werden kann.

Die beobachteten Korrelationen zwischen X und Z sowie Y und Z enthalten aber zumindest Informationen über die möglichen Korrelationen zwischen X und Y, da die gesamte Korrelationsmatrix positiv definit sein muss. Mit dieser Nebenbedingung lässt sich der mögliche Bereich für die wahren Korrelationen erheblich einschränken; im Folgenden wird darauf noch ausführlich eingegangen (vgl. Abschnitt 3.4).

3.3.4 Vierte Ebene: Erhalt der Randverteilungen

Die Mindestanforderung an einen fusionierten Datensatz besteht in der Reproduktion der Randverteilung von Y und der gemeinsamen Verteilung von Y und Z aus dem Spenderdatensatz. Diese Bedingung kann empirisch überprüft werden und wird daher bei praktischen Anwendungen als einziges Gütekriterium für eine Datenfusion herangezogen. Zum Einsatz kommen dabei meist Signifikanztests (vor allem χ^2 - oder t-Tests), d.h. es werden Parameter der Randverteilungen im Spender- und im fusionierten Datensatz verglichen und getestet, ob etwaige Abweichungen noch vom Zufall erklärt werden können. Wenn nur sehr wenige dieser Tests signifikant sind, wird die Fusion als erfolgreich angesehen. Abgesehen davon, dass man aus dem Beibehalten (im Gegensatz zum Verwerfen) einer Nullhypothese keine belastbare Aussage ableiten kann, deutet eine in diesem Sinne erfolgreiche Fusion lediglich darauf hin, dass die beiden Teilstichproben gute Abbilder der Gesamtheit darstellen. Über die Reproduktion der gemeinsamen Verteilung von X und Y kann mit diesem Gütekriterium keine Aussage getroffen werden.

3.4 Bestimmung der möglichen Korrelationsstruktur

Um abzuschätzen, wie sehr sich die Korrelation in der Fusionsstichprobe von der tatsächlichen Korrelation zwischen X und Y unterscheiden kann, muss zunächst die Menge der möglichen Korrelationen bestimmt werden. Ausgangs-

punkt ist die Korrelationsmatrix der Variablengruppen Z (gemeinsame Variablen) und X, Y (spezifische Variablen), die sich wie folgt partitionieren lässt:

$$\mathrm{Corr}(Z,Y,X)=\begin{pmatrix}\Sigma_{ZZ} & \Sigma_{ZY} & \Sigma_{ZX}\\ \Sigma_{ZY}' & \Sigma_{YY} & \Sigma_{YX}\\ \Sigma_{ZX}' & \Sigma_{YX}' & \Sigma_{XX}\end{pmatrix}$$

Bis auf Σ_{YX} können alle Teilmatrizen konsistent geschätzt werden. Da die Variablen X und Y nie gemeinsam beobachtet werden, ergibt sich aus den Stichproben keine direkte Schätzung für ihre Korrelation. In den beiden Stichproben sind aber indirekte Informationen über Σ_{YX} enthalten. Zum einen muss die Korrelationsmatrix Corr(Z, Y, X) positiv definit sein, woraus sich Bedingungen für Σ_{YX} ableiten lassen. Zum anderen schränken die beobachtbaren Randverteilungen von X und Y über die Fréchet-Hoeffding-Grenzen die gemeinsame Verteilung und damit auch die Korrelation zwischen X und Y ein.

Während im allgemeinen Fall die Menge der Σ_{YX}, die Corr(Z, Y, X) zu einer positiv definiten Matrix machen, nur durch Monte-Carlo-Simulation näherungsweise bestimmt werden kann, ist für univariates X *oder* Y eine analytische Lösung möglich.[3] Ist z.B. X univariat, so lässt sich zeigen, dass Corr(Z, Y, X) genau dann positiv definit ist, wenn

$$\Sigma_{XX}-\begin{pmatrix}\Sigma_{ZX}' & \Sigma_{YX}'\end{pmatrix}\begin{pmatrix}\Sigma_{ZZ} & \Sigma_{ZY}\\ \Sigma_{ZY}' & \Sigma_{YY}\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{ZX}\\ \Sigma_{YX}\end{pmatrix}>0$$

Dies ist eine quadratische Form in dem Vektor Σ_{YX}, die eine geschlossene Darstellung der Lösungsmenge erlaubt. Umformungen führen zu folgender Darstellung der zulässigen Korrelationen:

$$\left(\Sigma_{YX}-\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}\right)'C\left(\Sigma_{YX}-\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}\right)<1$$

$$\text{mit } C=(1-\Sigma_{ZX}'\Sigma_{ZZ}^{-1}\Sigma_{ZX})^{-1}\cdot(\Sigma_{YY}-\Sigma_{ZY}'\Sigma_{ZZ}^{-1}\Sigma_{ZY})^{-1}$$

Das ist die so genannte Normalform eines Ellipsoids, woraus unmittelbar seine geometrische Lage abgelesen werden kann:

- Der Mittelpunkt des Ellipsoids ist gegeben durch $\Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$;
- die Halbachsen des Ellipsoids liegen entlang der Eigenvektoren von C;
- die Halbachsenlängen betragen $1/\sqrt{\lambda_i}$ (λ_i sind die Eigenwerte von C).

3 Für den Spezialfall von univariatem X *und* univariatem Y finden sich diese Überlegungen auch bei Kadane (1978) und Moriarity/Scheuren (2001).

Das folgende Zahlenbeispiel illustriert diese Überlegungen und verdeutlicht ihre Konsequenzen für die Validität einer Datenfusion. Ausgangspunkt war ein (simulierter) Datensatz, der in zwei Teilstichproben mit nicht vollständig überlappender Variablenmenge getrennt wurde. Es gibt drei gemeinsame Variablen Z, der Vektor Y besteht aus zwei Variablen, X ist univariat. Aus den beiden Teilstichproben wurde die nachstehende unvollständige Korrelationsmatrix berechnet.

$$\Sigma = \left(\begin{array}{ccc|cc|c} 1 & 0{,}2 & -0{,}5 & 0{,}5 & 0{,}39 & 0{,}6 \\ 0{,}2 & 1 & -0{,}67 & 0{,}89 & 0{,}72 & 0{,}8 \\ -0{,}5 & -0{,}67 & 1 & -0{,}77 & -0{,}56 & -0{,}7 \\ \hline 0{,}5 & 0{,}89 & -0{,}77 & 1 & 0{,}8 & \rho_{XY_1} \\ 0{,}39 & 0{,}72 & -0{,}56 & 0{,}8 & 1 & \rho_{XY_2} \\ \hline 0{,}6 & 0{,}8 & -0{,}7 & \rho_{XY_1} & \rho_{XY_2} & 1 \end{array}\right) = \begin{pmatrix} \Sigma_{ZZ} & \Sigma_{ZY} & \Sigma_{ZX} \\ \Sigma'_{ZY} & \Sigma_{YY} & \Sigma_{YX} \\ \Sigma'_{ZX} & \Sigma'_{YX} & \Sigma_{XX} \end{pmatrix}$$

Die Korrelationen zwischen X und Y_1 sowie X und Y_2 lassen sich nicht unmittelbar schätzen, weil X und Y_1 bzw. X und Y_2 nie gemeinsam beobachtet wurden. Trotzdem lässt sich der mögliche Wertebereich für die beiden Korrelationskoeffizienten ρ_{XY_1} und ρ_{XY_2}, die grundsätzlich zwischen -1 und +1 liegen können, erheblich einschränken, wenn man die notwendige Bedingung berücksichtigt, dass die Corr(Z, Y, X) positiv definit sein muss. Die vorgenannten Formeln liefern eine Ellipse, in deren Inneren alle zulässigen Korrelationen liegen (vgl. Abbildung 4).

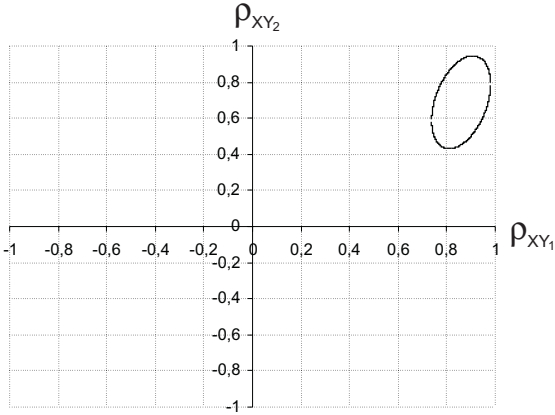


Abb. 4: Menge der möglichen Korrelationen (Ellipse)

Aus der Abbildung geht hervor, dass die Menge der möglichen Korrelationen ρ_{XY_1} und ρ_{XY_2} durch die Zusatzinformation der gemeinsamen Variablen Z erheblich eingeschränkt werden. Die Fläche der Ellipse, die die möglichen Korrelationen enthält, kann als Gütemaß für eine Datenfusion betrachtet werden. Je kleiner dieser Bereich ist, desto näher wird die Korrelation in der fusionierten Stichprobe an der tatsächlichen Korrelation zwischen X und Y liegen, da die Korrelation bei bedingter Unabhängigkeit zwischen X und Y (gegeben Z) als eine zulässige Korrelation ebenfalls im Inneren der Ellipse liegt.

Die Fréchet-Hoeffding-Grenzen können den zulässigen Bereich noch weiter einschränken. Zunächst gilt für die gemeinsame Verteilungsfunktion von X und Y_1 bzw. X und Y_2 die folgende Abschätzung:

$$\max\left\{0, F_X(x) + F_{Y_i}(y) - 1\right\} \le F_{X,Y_i}(x,y) \le \min\left\{F_X(x), F_{Y_i}(y)\right\} \quad (i = 1, 2)$$

Die beiden extremen Verteilungen lassen sich aus den beobachteten Randverteilungen leicht konstruieren. Weil der Korrelationskoeffizient nach Bravais/Pearson die von den Verteilungsfunktionen induzierte partielle Ordnung aller Verteilungen erhält, nimmt er seine extremen Werte an den beiden extremen Verteilungen an, d.h. man erhält für ρ_{XY_1} und ρ_{XY_2} aus den Fréchet-Hoeffding-Grenzen jeweils Intervalle, in denen die Korrelationskoeffizienten liegen können. In dem betrachteten Beispiel schränken diese Intervalle den zulässigen Bereich weiter ein, wie Abbildung 5 zeigt; der übrig gebliebene zulässige Bereich ist schwarz eingefärbt.

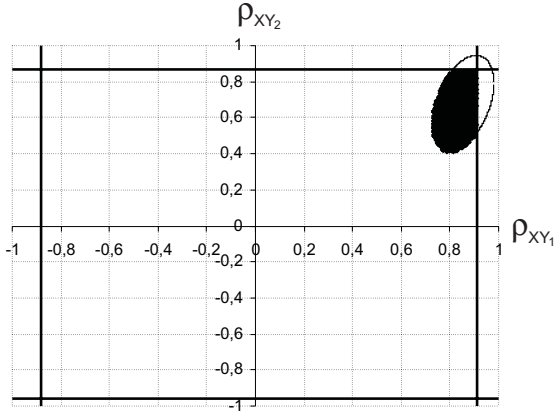


Abb. 5: Weitere Einschränkung durch Korrelationen der Fréchet-Hoeffding-Grenzen

Um die den üblichen Verfahren der Datenfusion zugrunde liegende, aber unüberprüfbare Annahme der bedingten Unabhängigkeit zu überwinden, könnte ein alternatives Verfahren wie folgt aussehen: Mit Hilfe der beschriebenen Vor-

gehensweise lässt sich der mögliche Bereich bestimmen, in dem die Korrelationen zwischen den Variablenblöcken X und Y liegen können. Aus diesem Bereich wählt man (möglichst gleichmäßig verteilt) mehrere zulässige Korrelationen aus. Mit Hilfe der Technik der Multiplen Imputation (mehrfache Ergänzung, kurz MI, siehe Rubin (1987, 2004)) lassen sich dann mehrere vervollständigte Empfängerdatensätze erzeugen, die jeweils der vierten Validitätsebene genügen und bei denen zusätzlich die Korrelationsstruktur zwischen X und Y kontrolliert wird: die unterschiedlichen Datensätze werden so erzeugt, dass sich die im Schritt vorher ausgewählten zulässigen Korrelationen ergeben. Ein entsprechender MI-Algorithmus wird bei Rässler (2002, 2004) beschrieben und durch Simulationsstudien validiert. Durch Betrachtung aller erzeugten Datensätze lässt sich der Einfluss der unbeobachteten Korrelationen auf die Ergebnisse der statistischen Analysen des fusionierten Datensatzes kontrollieren. Dieses Vorgehen ermöglicht die Durchführung von Sensitivitätsanalysen, wie schon von Rubin (1986) vorgeschlagen.

Zusammenfassung und Ausblick

Herkömmliche Methoden der Datenfusion erzeugen näherungsweise bedingte Unabhängigkeit der spezifischen Variablen, gegeben die gemeinsamen Variablen. Korrelationen zwischen nicht gemeinsam beobachteten Variablen werden in der fusionierten Datei nur dann korrekt wiedergegeben, wenn sie im Durchschnitt bedingt unkorreliert sind. Da diese Bedingungen ohne Zusatzinformation nicht überprüfbar sind, ist die Validität der Aussagen, die mit Hilfe statistischer Analysen des fusionierten Datensatzes gewonnen werden, in Frage gestellt. Eine Möglichkeit zur Überwindung dieses Problems ist die Bestimmung der zulässigen Korrelationsstrukturen, so dass mit Hilfe der Multiplen Imputation unterschiedliche Fusionsstichproben erzeugt werden können, die die Bandbreite der möglichen Zusammenhänge zwischen den einzelnen Variablen widerspiegeln. Für normalverteilte Variablen existieren bereits geeignete Algorithmen (vgl. Rässler (2002); zukünftiges Augenmerk liegt auf der Behandlung diskreter Variablen und einer Ausweitung auf nichtparametrische Korrelationsmaße.

Literatur

Bellmann, L., Bender, S., Kölling, A. (2002). Der Linked Employer-Employee-Datensatz aus IAB-Betriebspanel und Beschäftigtenstatistik der Bundesanstalt für Arbeit (LIAB), Beiträge zur Arbeitsmarkt- und Berufsforschung 250, S. 21-30.

Bender, S., Haas, A. (2002). Die IAB-Beschäftigtenstichprobe, Beiträge zur Arbeitsmarkt- und Berufsforschung 250, S. 3-12.

Kadane, J.B. (1978). Some Statistical Problems in Merging Data Files, 1978 Compendium of Tax Research, U.S. Department of the Treasury, S. 159-171. Nachdruck 2001, Journal of Official Statistics, 17, S. 423-433.

Kruppe, T., Oertel, M. (2003). Von Verwaltungsdaten zu Forschungsdaten. Die Individualdaten für die Evaluation des ESF-BA-Programms 2000 bis 2006, IAB Werkstattbericht 10/2003.

Liu, T.P., Kovacevic, M.S. (1996). Categorically Constrained Matching, Proceedings of the Survey Methods Section, Statistical Society of Canada, S. 123-133.

Moriarity, C., Scheuren, F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure, Journal of Official Statistics, 17, S. 407-422.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Lecture Notes in Statistics 168, Springer, New York.

Rässler, S. (2004). Data Fusion: Identification Problems, Validity, and Multiple Imputation, Austrian Journal of Statistics, 33, S. 153-171.

Raetzel, D.F. (2000). Werbewirkungsanalyse mit fusionierten Daten. Masters Thesis, Marburg.

Ridder, G., Moffitt, R. (2005). The Econometrics of Data Combination, in: Handbook of Econometrics Vol. 6, im Erscheinen.

Rodgers, W.L. (1984). An Evaluation of Statistical Matching, Journal of Business and Economic Statistics, 2, S. 91-102.

Rubin, D. B. (1986). Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations, Journal of Business and Economic Statistics, 4, S. 87-95.

Rubin, D. B. (1987, 2004). Multiple Imputation for Nonresponse in Surveys, Wiley, Hoboken, New Jersey.

Schnell, R., Bachteler, T., Reiher, J. (2005). MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung, Zentralarchiv-Informationen Heft 56, S. 93-103.

Wendt, F. (1976). Beschreibung einer Fusion, Gruner & Jahr, Schriftenreihe 21, Hamburg.

Wendt, F. (1980). Zusammenführung von Daten – Voraussetzungen und Grenzen, Gruner & Jahr, Schriftenreihe 28, Hamburg.

Wendt, F. (1986). Einige Gedanken zur Fusion, in: Auf dem Wege zum Partnerschaftsmodell, Arbeitsgemeinschaft Media-Analyse e.V., Media-Micro-Census GmbH, Frankfurt, S. 109-140.

Techniken der Datenfusion

Michael Wiedenbeck

1 Einführung

Die Beobachtung mehrer Variabler wird üblicherweise vollständig an jedem Element einer Stichprobe vorgenommen, wenn man die gemeinsame Verteilung der Variablen analysieren möchte. Beispielweise lassen sich Assoziationen zwischen Variablen, Interaktionseffekte etc. im Prinzip nur dann analysieren, wenn man sämtliche Variablen jeweils vollständig an den Einheiten beobachtet hat, die die Gesamtheit der Ausprägungen aller Variablen als multivariates Merkmal besitzen und daher den Zusammenhang zwischen den einzelnen Variablen vermitteln.

Dieses Prinzip schränkt natürlich den empirischen Blick auf Grundgesamtheiten ein: Nicht nur die Stichproben als Repräsentanten der jeweiligen Grundgesamtheiten sind zwangsläufig endlich und in ihrem Umfang durch begrenzte Ressourcen bestimmt, sondern auch die Umfänge der erfassten Variablen. Interviews können bestimmte Dauern nicht überschreiten, wenn man nicht einen Abbruch oder drastische Minderung der Antwortqualität in Kauf nehmen will.

Diesem Zwang zur Beschränkung von Umfrageinstrumenten muss immer wieder durch teilweise unbefriedigende Kompromisse genügt werden. Er kann allerdings auch eine unüberwindbare Barriere darstellen, wenn die zu erhebenden Variablen eine bestimmte – große – Menge nicht unterschreiten dürfen, um sinnvolle Informationen zu liefern. Ein Beispiel findet man in der Markt- und Meinungsforschung bei der Analyse von Konsumverhalten im Zusammenhang mit Mediennutzung. Beide Bereiche sind schon für sich genommen komplex und nur mit sehr umfangreichen Instrumenten zu erfassen, deren Vereinigung zu einem einzigen Instrument häufig nicht mehr handhabbar sein wird. Und dennoch ist die Analyse der gemeinsamen Verteilung einer Vielzahl von Mediennutzungs- und Konsumvariablen eine unumgängliche Voraussetzung für Empfehlungen der Markt- und Meinungsforschung, etwa hinsichtlich der Platzierung von Annoncen oder Werbesendungen in den Medien.

Auf andere Beispiele stößt man nahezu unweigerlich bei Sekundäranalysen: Die Variablen durchgeführter Studien sind zugeschnitten auf die unmittelbaren Ziele der Primärforscher, und man muss häufig in Datensätzen das Fehlen von Variablen zur Kenntnis nehmen, die zur Untersuchung einer erweiterten Fragestellung erforderlich wären. Diese befinden sich u.U. in anderen Datensätzen,

dann aber wieder nur zusammen mit Variablen, die für die Fragestellung bedeutungslos sind.

2 Das Modell der Datenfusion

Die Markt- und Meinungsforschung beispielsweise hat sich über dieses Dilemma hinweggesetzt und die Erhebung separater Variablen an verschiedenen Stichproben zur Praxis gemacht. Allerdings: völlig separat sind die Variablen dieser so genannten gesplitteten Fragebögen nicht. Wenn man den einfachsten Fall zweier Stichproben betrachtet, dann enthält jede der beiden Fragebogenversionen einen Teil mit gemeinsamen Schlüsselvariablen (overlap), die die befragten Personen bezüglich relevanter Variablen beschreiben. Liegen dann die Daten beider Stichproben vor, so kann man die befragten Personen bzgl. dieser Variablen vergleichen und feststellen, welche Ausprägungen in den Konsumvariablen einen „statistischen Doppelgänger" einer Befragungsperson aufweist, die man selbst zu ihrem Medienverhalten befragt hat. (Dies setzt natürlich voraus, dass es einen solchen Doppelgänger überhaupt gibt; andererseits kann es natürlich auch mehrere Doppelgänger für eine Person geben.)

Wenn man sich die Stichprobendaten in Form zweier üblicher Rechteckdateien vorstellt, bei denen die Zeilen die Personen der Stichproben und die Spalten die Variablen repräsentieren, dann lässt sich die Situation wie folgt schematisch darstellen:

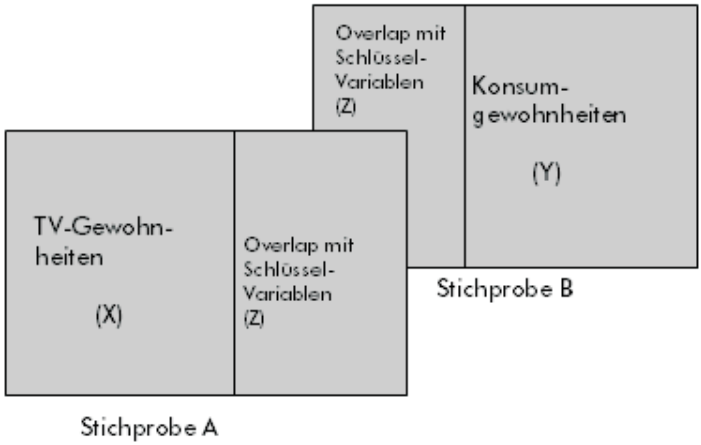


Abbildung 1

Es liegt nun sehr nahe, die Daten von Personen, die bezüglich der Schlüsselvariablen gleich sind, zusammenzuspielen und solche Daten als Ersatz für die Beobachtung der multivariaten Größe (X,Y) zu verwenden. Die Hypothese ist bei diesem Vorgehen, dass sich bezüglich der Schlüsselvariablen gleiche Personen auch hinsichtlich der Kombination von X und Y gleichen.

Dieser Ansatz lässt sich weiter in den folgenden Abbildungen 2a und 2b verdeutlichen:

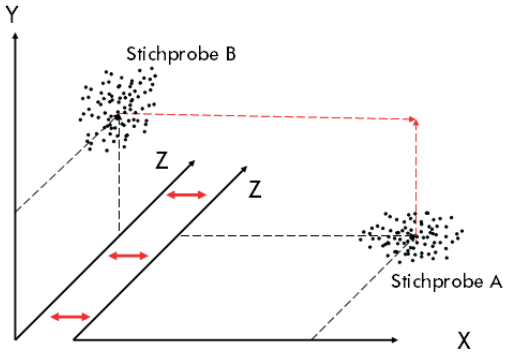


Abbildung 2a

Abbildung 2a zeigt die beobachteten Daten der Stichproben A und B als Beobachtungen der Marginalverteilungen von (X,Z) bzw. (Y,Z). Durch „Identifizierung" der Z-Achsen der beiden Räume wird der trivariate Raum von (X,Y,Z) erzeugt, wobei die (X,Z)-Beobachtungen mit (Y,Z)-Beobachtungen zusammengefügt werden, wenn sie in der Z-Komponente übereinstimmen.

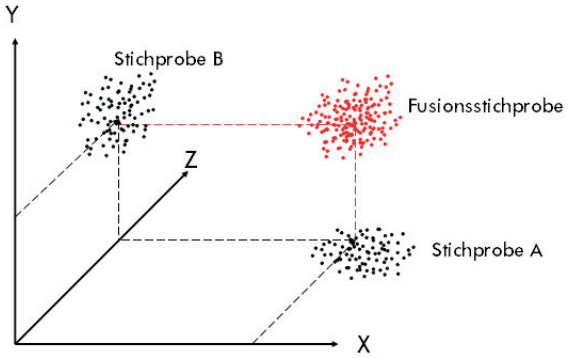


Abbildung 2b

In Abb. 2b ist nun aus Demonstrationsgründen vereinfachend angenommen, dass sich Paare von (Y,Z)- und (X,Z)-Beobachtungen perfekt zusammenfinden. Damit ist im Falle realer Datensätze natürlich nicht zu rechnen. Paarweises Matchen wird häufig nur zwischen Beobachtungen zustande kommen, wenn man Beobachtungen mit ähnlichen, aber nicht identischen Z-Profilen miteinander verbindet. Manche der Beobachtungen werden dabei möglicherweise überhaupt keinen „Partner" finden, sodass in A wie auch B ein nicht matchbarer Rest von Beobachtungen verbleiben wird. Außerdem sind i.a. die Umfänge der beiden Stichproben unterschiedlich. Die größere Stichprobe wird also mit Sicherheit nur einen Teil ihrer Beobachtungen mit der kleineren verbinden können.

Wir werden im Folgenden auf Probleme der Übereinstimmung von Beobachtungen der Stichprobe A mit denen der Stichprobe B nicht im Detail eingehen. Zu ihrer Diskussion sei auf Rodgers (1984) verwiesen. Wir unterstellen vielmehr die folgende idealisierte Situation: Stichprobe B sei so umfangreich, dass sie für jede Beobachtung in A mindestens einen Fall enthält, der in der Variablen Z mit ihr übereinstimmt. Tatsächlich werden in der Praxis der Datenfusion die Beobachtungen einer kleineren Stichprobe A um die Daten aus einer größeren Stichprobe ergänzt. Man spricht daher auch von Empfänger- und Spenderstichproben. Aus diesem Modell der Datenfusion folgt dann, wie weiter unten dargestellt wird, die wesentliche Verteilungseigenschaft von Fusionsstichproben, nämlich die bedingte Unabhängigkeit von X und Y, gegeben Z.

Man sieht also, dass durch Datenfusion multivariate Beobachtungen konstruiert werden, deren Teile von verschiedenen Beobachtungseinheiten stammen. Das Prinzip der Identität der Einheiten, die mit verschiedenen Komponenten multivariater Größen verbunden sind, wird durch ein Gleichheitsprinzip ersetzt: die Komponenten multivariater Größen stammen von nicht-identischen Einheiten, die aber bezüglich (multivariater) Schlüsselvariablen gleich sind.

3 Datenfusion und Stichprobentheorie

Zur Klärung der statistischen Eigenschaften von Stichproben, die durch ein Fusionsverfahren konstruiert worden sind (im Folgenden kurz Fusionsstichproben genannt), ist es nützlich, die Frage aufzuwerfen, ob Fusionsstichproben denn eigentlich als Stichproben aus endlichen Grundgesamtheiten, also Stichproben im Sinne der üblichen Stichprobentheorie, aufgefasst werden können. Und wenn, welchen (endlichen) Grundgesamtheiten Fusionsstichproben angehören.

Zunächst zur zweiten Frage: Offenbar liegt jeder Beobachtung in einer Fusionsstichprobe ein Paar von Beobachtungseinheiten zugrunde. Die Komponenten der (multivariaten) Variablen X gehören zu einer Einheit aus Stichprobe A, die Komponenten von Y gehören zu einer Einheit aus Stichprobe B, beide Einheiten stimmen (im Sinne unserer vereinfachenden Annahme) bezüglich der

(multivariaten) Schlüsselvariablen Z überein. Die Fusionsstichprobe ist also eine Stichprobe aus der Grundgesamtheit derjenigen Paare von Einheiten aus der Grundgesamtheit G, die bezüglich der Variablen Z übereinstimmen.

Für den Fall, dass Z eine diskrete Variable mit endlich vielen möglichen Ausprägungen ist, kann die Grundgesamtheit der Fusionsstichprobe folgendermaßen skizziert werden:

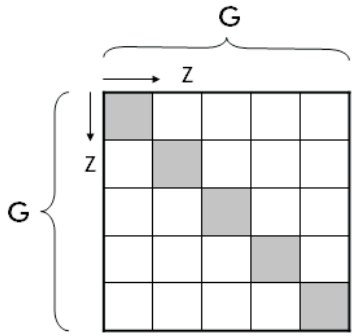


Abbildung 3

Spalten- und Zeilenbereiche der Matrix von Abbildung 3 repräsentieren die nach Z geschichteten Schichten der Grundgesamtheit G. Die gesamte Matrix repräsentiert die Menge aller Paare aus Einheiten der Grundgesamtheit; die Felder der grau unterlegten Hauptdiagonalen repräsentieren die Paare, die in der Variablen Z übereinstimmen. Diese Felder repräsentieren also die Grundgesamtheit der Fusionsstichproben.

Wenn nun eine Empfängerstichprobe A vorliegt und eine hinreichend große Spenderstichprobe B gezogen wurde, dann besteht die Konstruktion der Fusionsstichprobe darin, die beiden Stichproben nach Z zu schichten und für jede Einheit der Stichprobe A in der gleichen zugehörigen Z-Schicht zufällig eine Einheit der Stichprobe B auszuwählen und beide Einheiten zu einer neuen Einheit der Fusionsstichprobe zu verbinden.

Die erste der beiden obigen Fragen lässt sich nun präzisieren: Wenn eine Empfängerstichprobe A nach einem Design p_E und eine Spenderstichprobe B nach einem Design p_S - unabhängig von p_E - gezogen wurden, nach welchem Design wird dann die Fusionsstichprobe „gezogen“? Es ist intuitiv klar, dass sich das Design der Fusionsstichprobe aus dem Design p_E für Stichprobe A und einem bedingten Design, gegeben die Elemente der Stichprobe A als die ersten „Paarhälften“, für eine Substichprobe von B (der zweiten „Paarhälften“) konstruieren lässt.

Bezeichnet etwa F eine Teilmenge von Paaren, deren jeweilige Paarhälften in den Z-Variablen übereinstimmen, dann ist als Stichprobe A (Empfängerstichprobe) grundsätzlich jede Stichprobe denkbar, die durch Wahl von jeweils genau einer Einheit aus allen Paaren gebildet wird. Die restlichen Einheiten bilden dann die Substichprobe ($B_F|A$) einer Spenderstichprobe B, deren Einheiten mit den Einheiten von A verbunden sind. Die kombinatorischen Möglichkeiten für beide Stichproben sind für ein gegebenes F also sehr groß. Einerseits sind viele Stichproben A möglich, denen im Allgemeinen – mit Ausnahme einfacher Designs - unterschiedliche Wahrscheinlichkeitswerte durch das Design p_E zugewiesen werden. Für die Stichprobe ($B_F|A$) der ergänzenden Einheiten ist die Vielfalt noch größer: zum einen gibt es als Stichprobe des „Rests“ von F nach Abzug einer Stichprobe A die gleiche Anzahl von Möglichkeiten wie für A, zum anderen gibt es i.a. für jedes ($B_F|A$) eine Vielzahl von Oberstichproben B.

Das Design der Fusionsstichprobe ist dann darstellbar als die Summen von Produkten $p_E(A)p(B_F|A)$, summiert über alle Stichproben A und B. Leider besitzt im Allgemeinen das Fusionsdesign nicht die Form einer einfachen geschlossenen Funktion der Designs p_E und p_S.

Die Form des Fusionsdesigns soll hier allerdings nicht untersucht werden. Die Kenntnis des Fusionsdesigns würde es erlauben, die bivariate Verteilung der Kombinationsvariablen (X,Y) der endlichen Grundgesamtheit aller Paare von Einheiten mit gleichen Werten von Z zu schätzen. Das Ziel ist jedoch, die gemeinsame Verteilung von X und Y auf der Grundgesamtheit G selbst zu schätzen.

3.1 Bedingte Unabhängigkeit von X und Y, gegeben Z

Nach dem oben beschriebenen Konstruktionsprinzip wird eine Einheit der Stichprobe A, die die Ausprägung (x,z) der Kombinationsvariablen (X,Z) besitzt, mit einer Einheit der Stichprobe mit der Ausprägung (y,z) der Variablen (Y,Z) verbunden. Die Verbindung zwischen den A-Einheiten und den B-Einheiten mit den gleichen Werten von Z ist rein zufällig – außer Z gibt es keine weiteren Kriterien für die Fusion. Die gemeinsame Verteilung der durch Fusion kombinierten Variablen X und Y ist daher bedingt unabhängig, gegeben Z=z (oder kürzer: gegeben Z) (vgl. dazu auch Rodgers (1984)). Beschreibt man die gemeinsame Verteilung der Variablen X, Y und Z durch eine Dichtefunktion f(X,Y,Z), die Marginalverteilungen der in A und B gemeinsam beobachteten Variablen X und Z bzw. Y und Z durch f(X,Z) bzw. f(Y,Z) und die bedingte Verteilung von Y, gegeben (X,Z) mit f(Y|X,Z), dann ist die Dichte der durch Fusion kombinierten Variablen X, Y und Z gleich f(Y|X,Z)f(X,Z). Wegen der Unabhängigkeit des Fusionskriteriums von X ist dieser Ausdruck gleich f(Y|Z)f(X,Z) und dies ist wiederum gleich f(Y|Z) f(X|Z)f(Z), was die bedingte Unabhängigkeit von X und Y, gegeben Z, bedeutet. Diese Verteilung besitzt die gleichen

Marginalverteilungen von (X,Z) und (Y,Z), f(X,Z) und f(Y,Z) wie die gemeinsame „wahre“ Verteilung f(X,Y,Z), stimmt aber mit dieser nur dann überein, wenn diese ebenfalls die Eigenschaft der bedingten Unabhängigkeit von X und Y, gegeben Z, besitzt.

4 Fusionsbias

Die durch Fusion konstruierten Stichproben repräsentieren also nicht die gemeinsame Verteilung von X, Y und Z, sondern eine Verteilung, die die gleichen Marginalverteilungen von X und Z bzw. Y und Z wie die tatsächliche Verteilung besitzt. Die Differenz zwischen der „wahren“ und der durch Fusion konstruierten Verteilung von (X,Y,Z) wird als Fusionsbias bezeichnet. Es stellt sich daher die Frage, wie „groß“ dieser Bias ist – eine Frage, die in allgemeiner Form nicht zu beantworten ist. Im Folgenden wird der Fusionsbias am Beispiel der Kovarianz einer multivariaten Normalverteilung und eines Regressionskoeffizienten illustriert.

4.1 Gemeinsame Normalverteilung von X, Y und Z

X, Y und Z seien gemeinsam normal verteilt, wobei der Einfachheit halber die Mittelwerte gleich 0 angenommen werden. Die Varianzmatrix der Normalverteilung sei Σ:

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2 \end{pmatrix}$$

Die bedingte Varianz von (X,Y), gegeben Z, ist dann gleich

$$\Sigma_{XY|Z} = \begin{pmatrix} \sigma_X^2 - \frac{\sigma_{XZ}^2}{\sigma_Z^2} & \sigma_{XY} - \frac{\sigma_{XZ}\sigma_{ZY}}{\sigma_Z^2} \\ \sigma_{YX} - \frac{\sigma_{YZ}\sigma_{ZX}}{\sigma_Z^2} & \sigma_Y^2 - \frac{\sigma_{YZ}^2}{\sigma_Z^2} \end{pmatrix}$$

Die bedingte Unabhängigkeit von X und Y, gegeben Z ist äquivalent zu

$$\sigma_{XY} = \frac{\sigma_{XZ}\sigma_{ZY}}{\sigma_Z^2}$$

Dann repräsentiert nach Beobachtung einer Stichprobe von (X,Z) und (Y,Z) und anschließender Fusion die Fusionsstichprobe eine Normalverteilung mit Mittelwert 0 und der Varianzmatrix Σ´:

$$\Sigma' = \begin{pmatrix} \sigma_X^2 & \frac{\sigma_{XZ}\sigma_{ZY}}{\sigma_Z^2} & \sigma_{XZ} \\ \frac{\sigma_{YZ}\sigma_{ZX}}{\sigma_Z^2} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2 \end{pmatrix}$$

Die „wahre“ Verteilung und ihr Surrogat unterscheiden sich also in der Kovarianz von X und Y, also der Assoziation der nicht gemeinsam beobachteten Variablen. Die Differenz lässt sich mit Hilfe der Ungleichung von Cauchy-Schwarz für die bedingten Varianzen und Kovarianzen von X und Y, gegeben Z, abschätzen.

$$\left(\sigma_{XY} - \frac{\sigma_{XZ}\sigma_{ZY}}{\sigma_Z^2}\right)^2 \leq \left(\sigma_X^2 - \frac{\sigma_{XZ}^2}{\sigma_Z^2}\right)\left(\sigma_Y^2 - \frac{\sigma_{YZ}^2}{\sigma_Z^2}\right)$$

Die quadrierte Differenz zwischen wahrer Kovarianz und der in der Fusionsstichprobe beobachtbaren Kovarianz ist durch das Produkt auf der rechten Seite der Ungleichung beschränkt, deren Faktoren sich aus den Beobachtungen von (X,Z) und (Y,Z) schätzen lassen. Man verfügt in diesem Fall also trotz der Tatsache, dass X und Y nicht gemeinsam beobachtet wurden, über Information zur Abschätzung des Fusionsbias der Kovarianz von X und Y (vergleiche dazu auch Rässler (2001)).

Allerdings kann der Fusionsbias auch so groß ausfallen, dass die Fusion zu einem Vorzeichenwechsel in der Kovarianz führt. Sei beispielsweise das „wahre“ Σ wie folgt:

$$\Sigma = \begin{pmatrix} 1 & .7 & .49 \\ .7 & 1 & -.49 \\ .49 & -.49 & 1 \end{pmatrix}$$

Dann ergibt sich für die Kovarianz der Fusionsverteilung

$$\frac{\sigma_{XZ}\sigma_{ZY}}{\sigma_Z^2} = -.2401$$

4.2 Regression von Y auf X

Als ein weiteres Beispiel soll nun der Fusionsbias im Regressionskoeffizienten β_{YX} der Regression von Y auf X dargestellt werden. Hierfür erweist sich eine Formel von Cochran (1938) (vgl. Cox und Wermuth (2003)) als nützlich:

$$\beta_{YX} = \beta_{YX \cdot Z} + \beta_{YZ \cdot X}\beta_{ZX}$$

Der Effekt β_{YX} setzt sich also aus einem direkten Effekt $\beta_{YX.Z}$ und einem indirekten Effekt zusammen, der aus dem Produkt des Effekts β_{ZX} von X auf Z in der Marginalverteilung von (X,Z) und des direkten Effekts $\beta_{YZ.X}$ von Z auf Y besteht.

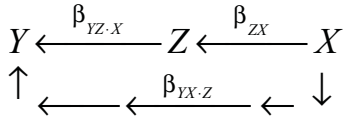


Abbildung 4

In der Fusionsstichprobe gilt nun eine analoge Beziehung, wobei die Regressionskoeffizienten der Fusionsverteilung durch eine Tilde gekennzeichnet werden:

$$\tilde{\beta}_{YX} = \tilde{\beta}_{YX \cdot Z} + \tilde{\beta}_{YZ \cdot X}\beta_{ZX}$$

Wegen der bedingten Unabhängigkeit von X und Y, gegeben Z in der Fusionsstichprobe gelten die folgenden Beziehungen

$$\tilde{\beta}_{YX\cdot Z} = \underbrace{\tilde{\sigma}_{YX\cdot Z}}_{0}\,\tilde{\sigma}^{-2}_{X\cdot Z} = 0$$

$$\tilde{\beta}_{YZ\cdot X} = \underbrace{\tilde{\sigma}_{YZ}\,\tilde{\sigma}^{-2}_{Z}}_{\tilde{\beta}_{YZ}=\beta_{YZ}} - \underbrace{\tilde{\sigma}_{YX\cdot Z}}_{0}\,\tilde{\sigma}^{-2}_{X\cdot Z}\,\tilde{\sigma}_{XZ}\,\tilde{\sigma}^{-2}_{Z} = \beta_{YZ}$$

sodass sich für die Fusionsstichprobe ergibt, dass

$$\tilde{\beta}_{YX} = \beta_{YZ}\beta_{ZX}$$

In der durch die Fusionsstichprobe repräsentierten Verteilung (Fusionsverteilung) ist der Regressionskoeffizient bzgl. der Marginalverteilung von (X,Y) gleich dem Produkt der Koeffizienten der marginalen Verteilungen von (Y,Z) und (Z,X).

Der Fusionsbias lässt sich nun durch die folgenden Beziehungen beschreiben. Vertauscht man in der Formel von Cochran X und Z, so ergibt sich

$$\beta_{YZ} = \beta_{YZ\cdot X} + \beta_{YX\cdot Z}\beta_{XZ}$$

Nach Obigem folgt daraus

$$\tilde{\beta}_{YX} = \beta_{YZ}\beta_{ZX} = (\beta_{YZ\cdot X} + \beta_{YX\cdot Z}\beta_{XZ})\beta_{ZX}$$

Sodass schließlich gilt:

$$\begin{aligned}\tilde{\beta}_{YX} &= \beta_{YX\cdot Z}(\beta_{XZ}\beta_{ZX}) + \beta_{YZ\cdot X}\beta_{ZX} \\ &= \beta_{YX\cdot Z}\rho^{2}_{XZ} + \beta_{YZ\cdot X}\beta_{ZX}\end{aligned}$$

Die letzte Gleichung lässt nun einen Vergleich mit dem „wahren“ Regressionskoeffizienten von Y auf X zu

$$\beta_{YX} = \beta_{YX\cdot Z} \quad + \beta_{YZ\cdot X}\beta_{ZX}$$

M.a.W.: Wenn X und Z perfekt korreliert sind – positiv oder negativ - , dann stimmen die beiden Korrelationskoeffizienten überein. Sind X und Z unkorreliert, dann ist der Regressionskoeffizient in der Randverteilung von (X,Y) der Fusionsstichprobe gleich dem „wahren“ indirekten Effekt, vermittelt über die

Schlüsselvariable Z. Für Werte der quadrierten Korrelation zwischen 0 und 1 liegt der der Fusionsverteilung unterliegende Regressionskoeffizient über oder unter dem entsprechenden „wahren" Wert, je nachdem, ob $\beta_{YX.Z}$ negativ oder positiv ist. Die Korrelation zwischen X und Z lässt sich aus den beobachteten Daten schätzen, der Koeffizient $\beta_{YX.Z}$ dagegen nicht. Eine Beurteilung des Fusionsbias muss daher mit Vermutungen über $\beta_{YX.Z}$ auskommen. Wenn schließlich $\beta_{YX.Z}$ und $\beta_{YZ.X}$ β_{ZX} unterschiedliche Vorzeichen haben, dann kann auch wieder ein Wechsel des Vorzeichens nach der Fusion auftreten.

5 Schlussbemerkung

Datenfusion ist gelegentlich ein plausibler Ausweg aus einer Situation des Totalausfalls von Variablen. Sie erscheint dabei als eine besondere Art des Datenmanagements, bei dem das Resultat – wie bei allen Imputationen – so aussieht wie Daten, die tatsächlich gemeinsam beobachtet wurden. Der Nutzen der Fusion ist aber nur dann zu beurteilen, wenn man eine Relation zwischen den Parametern der „wahren" gemeinsamen Verteilung von (X,Y,Z) bzw. (X,Y) und den Parametern der tatsächlich beobachteten Randverteilungen (X,Z) und (Y,Z) angeben kann. Damit bekommt allerdings das Problem der Fusionsstichproben eine andere Perspektive: An die Stelle der Konstruktion einer mit den üblichen Software-Paketen analysierbaren Rechteckdatei aus den Daten zweier verschiedener Stichproben tritt hier das Problem von Schätzungen von (Parametern von) Verteilungen, für die nur Beobachtungen aus Marginalverteilungen erfasst wurden. In dieser Perspektive entfallen grundsätzlich die Probleme infolge mangelnder Übereinstimmung von Beobachtungen verschiedener Stichproben bezüglich der Schlüsselvariablen. Außerdem würde dann die Unterscheidung zwischen Empfänger- und Spenderstichprobe obsolet, und der Informationsverlust durch Auswahl einer echten Substichprobe der Spenderstichprobe träte nicht mehr auf. Der Preis dafür wären allerdings Schätzverfahren, die auf die jeweiligen Datensituationen zugeschnitten werden müssten und nicht ohne weiteres standardisierbar sein dürften.

Literatur

D. R. Cox and N. Wermuth, *A general condition for avoiding effect reversal after marginalization*. J.R. Statist. Soc. B (2003) 65, Part 4, 937-941

S. Rässler, *Statistical Matching. A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer, New York, 2002

W. L. Rodgers, *An Evaluation of Statistical Matching. Journal of Business &Economic Statistics*. (1984) Vol. 2, No. 1, 91-102

Media-Analysen & Fusionen

Uwe Czaia

Datenfusionen sind mittlerweile Standardverfahren in allen Bereichen der Markt- und Sozialforschung sowie auch bei angewandten Naturwissenschaften, bei denen zum Entwickeln und Testen von Hypothesen oder für Prognosen fehlende Daten aus anderen Quellen per Fusionsverfahren ergänzt werden. Zu Fusionsverfahren und Beispielen listet eine Bibliographie aus dem November 2003 über 1.300 Veröffentlichungen speziell zu diesem Thema auf.

Für einige sind Fusionen ein 'Königsweg', für andere 'das kleinere Übel' und einige Puristen halten Fusionen im Bereich der Markt- und Sozialforschung für einen 'methodischen Irrweg'. Gegen letztere Einstellung spricht nicht die große Zahl 'zufrieden stellender' Fusionen aus der Praxis für die Praxis - wohl aber die Forschungsrealität:

Entwickelt wurden praxisnahe Verfahren in Europa mit einer Vorreiterrolle Deutschlands durch die ersten Fusionen innerhalb der nationalen Media-Analyse. Seit Mitte der 80er Jahre wird die Diskussion zunehmend internationalisiert. Alle Gesellschaften stehen vor dem Problem, dass immer detailliertere Zielgruppeninformationen gefordert werden, die über Single-Source-Studien - sei es ad hoc oder über Panel - nicht erhoben werden können, wenn eine Mindestqualität in Bezug auf die 'Abbildung' der Realität eingehalten werden soll.

Mit zunehmender Befragungsdauer sinkt die Antwortbereitschaft; die Ausschöpfung, das Antwortverhalten wird durch den Erhebungsumfang negativ beeinflusst, und nicht zuletzt überschreiten einige Untersuchungen tendenziell die Grenze der Zumutbarkeit.

Hauptanwendungsgebiete sind

- Ergänzung fehlender Informationen innerhalb eines Datenbestandes bei einem Teil der Stichprobe (z. B. Trennung zwischen einem Basisinterview und einem Haushaltsbuch mit unvollständigem Rücklauf).
- Ergänzung fehlender Informationen innerhalb eines Datenbestandes, wenn zur Einhaltung einer angemessenen Befragungsdauer systematisch und/oder per Zufall nur eine Auswahl der Themenbereiche erhoben wird.
- Zusammenführen von Daten aus unabhängigen Stichproben, z. B. TV-Meter-Daten mit Konsumdaten aus anderen Quellen.
- Anreicherung von Datenbanken mit Informationen aus Fremduntersuchungen oder einer Studie in einer Stichprobe dieser Datenbank und Zuschrei-

bung der Resultate über z. B. 'predictive Modeling' für alle Angehörigen der Datenbank.

- Übertragung von Medien-Reichweiten einer Referenzstudie mit Währungscharakter auf andere Markt-Media-Erhebungen.

Alle Fusionsverfahren - ob auf Basis von multidimensionaler Skalierung mit Übertragung von Antreffbarkeitswahrscheinlichkeiten (Profile Matching) oder der fallweisen Zuordnung eines Spenders (Donor) zu einem Empfänger (Rezipient) anhand von Ähnlichkeitsmaßen - gehen von der Hypothese aus, dass Personen mit ähnlichen Ausprägungen in demographischen und/oder Einstellungs-, Konsumvariablen auch ähnliche Ausprägungen bei anderen Merkmalen aufweisen, die nur in der jeweils anderen Studie oder einer Teilstichprobe vorhanden sind. Über eine Vielzahl von Experimenten konnte diese Hypothese bestätigt werden. Diese Versuche haben in aller Regel ein ähnliches Design: man nehme eine Untersuchung, teile per Zufall, bestimme eine Anzahl gemeinsamer Merkmale und fusioniere die Übrigen (z. B. Konsumverhalten, Mediennutzung etc.) kreuzweise. Durch den Vergleich mit den originären Angaben kann die Güte des Übertragungsvorganges geprüft werden.

Damit ist dann auch ein Vergleich unterschiedlicher Vorgehensweisen möglich - inklusive extensiver Berechnung von Signifikanzen über Chi-Quadrat oder anderer statistischer Tests. Diese Maßzahlen sind allerdings nicht invariant gegenüber Stichprobengröße, Aufbau der Vergleichsmatrizen, Signifikanzniveau usw., d. h. die Resultate sind 'manipulierbar'.

Und: bei einer nicht-experimentellen 'Real'-Fusion kann das, was man gerne prüfen möchte, nicht einem Test unterzogen werden, weil die zu fusionierenden spezifischen Variablen eben nur in einer Untersuchung, einer Teilstichprobe vorhanden sind. Auch hier sind statistische Maße (z. B. Vergleich von Signifikanzen bei den spezifischen Variablen vor und nach Fusion einschließlich der Überprüfung anhand weiterentwickelter statistischer Maßzahlen zur Schätzung der Güte einer Fusion, wie sie in anderen Beiträgen dieser wissenschaftlichen Tagung präsentiert wurden) nützlich, sie können aber nicht das Auge des Forschers ersetzen, 'face validity' ist gefordert - z. B. allein dadurch, dass 'Medienwährungen' durch eine Fusion nicht verändert werden sollten.

Die Festlegung auf ein statistisches Maß (z. B. Chi-Quadrat oder ähnliche Verfahren) kann im Übrigen für die Fusion kontraproduktiv sein. Wenn dieses Maß als Optimierungskriterium bei einer iterativen Zuordnung eingebunden wird, wird zwar die Zahl auffälliger Abweichungen minimiert, aber durchaus unter Umständen nicht das Fusionsergebnis verbessert, weil das Testverfahren nur eingeschränkt problemadäquat sein kann.

Voraussetzung für 'erfolgreiche' Fusionen sind eine möglichst große Zahl nicht nur demographischer gemeinsamer Merkmale, um den jeweils 'optimalen' Spender für einen Rezipienten bestimmen zu können. Datenüber-

tragungen allein anhand von Demographie entbehren nicht einer gewissen Ironie: einer der Hauptgründe für Fusionen ist, dass Demographie als Prädiktor für Konsumverhalten in modernen Gesellschaften unzureichend ist.

Aber auch hier muss nicht nur ein Programm zur technischen Abwicklung einer Fusion beherrscht werden. Es kann im Einzelfall nicht sinnvoll sein, alle gemeinsamen Merkmale in den Fusionsprozess einzubeziehen - z. B. wenn die Feldzeiten zwischen Spender- (Winter) und Empfänger-Datei (Sommer) stark differieren und eine gemeinsame Variable 'Freizeittätigkeiten' beinhaltet oder Panel-Effekte die Ausprägungen des einen oder anderen Merkmals beeinflussen könnten.

Fusionsergebnisse sind in aller Regel 'besser', wenn nur wenige Variablen übertragen werden. Für die Fusion stehen z. B. Konsumdaten zu Bereichen wie 'Garten/Zimmerpflanzen', 'Geld und Versicherungen', 'Gesundheit, Selbstmedikation', 'Haustiere', 'Körperpflege/Kosmetik' zur Verfügung. Wird nur einer dieser Bereiche übertragen, ist das Resultat für diese Variablen im Allgemeinen zufrieden stellender als eine Gesamtfusion aller Bereiche.

Ein Fehlschluss ist jedoch, dass das Fusionsergebnis insgesamt verbessert wird, wenn die Bereiche einer Spender-Datei jeweils separat übertragen werden. Der Preis ist eine Auflösung der Zusammenhänge zwischen den übertragenen spezifischen Merkmalen der Donoren, z. B. Katzenallergiker sind gleichzeitig Katzenbesitzer, Antialkoholiker mutieren zu heavy usern von Spirituosen o. ä.. Zusammenhänge können entstehen, die keine sind, Zusammenhänge können verschwinden oder invertiert werden. Wenn sich durch dieses spezifische Fusionsverfahren Zusammenhänge ändern, müsste dem 'Endverbraucher' jede simultane Verwendung von nicht gemeinsam übertragenen Merkmalen etwa in einer Kreuztabelle oder bei Zielgruppendefinitionen 'untersagt' werden - eine unrealistische Annahme.

Ebenso verhält es sich mit Ad Hoc-Fusionen (Fusion on the Fly) durch den 'Endverbraucher' - in der Regel kommen unterschiedliche Personen trotz gleicher Zielvorgabe zu unterschiedlichen Resultaten.

Die vorstehenden Anmerkungen verweisen auf Problemfelder. Fusionen sind nicht mechanistisch durchzuführen, sondern sind - um Friedrich Wendt zu zitieren - 'Kunsthandwerk'. Zu fordern sind Kenntnisse der Fusionsverfahren (Bewertung), aber insbesondere auch der Datenstrukturen: Stichproben, Feldzeit, Erhebungsmethode, Vergleich der Grundgesamtheiten, d. h. sind Daten bei formal gleicher Definition auch tatsächlich aus einer Grundgesamtheit (besonders auffällig im Vergleich von Random-Stichproben und Panel), Einfluss von Gewichtung u. v. m.

Die zunehmende Spezialisierung, Industrialisierung der Forschung mit der damit (häufig) verbundenen Beschränkung auf einige wenige Funktionen - sei es in der Studienleitung, Fragebogen-Redaktion, Datenerhebung oder EDV/

Statistik - erleichtert nicht den Umgang mit diesen hochkomplexen Anforderungen.

Bei Fusionen werden die Grenzen demnach gesetzt zum einen durch die Datenqualität der zu verbindenden Dateien und zum anderen durch den Kenntnisstand der Ausführenden: Und hier sehen wir Handlungsbedarf.

Der Marktforscher kann häufig nicht die Tragweite abschätzen, die durch den Einsatz bestimmter Verfahren generiert werden kann und der Statistiker, Mathematiker oder EDV-/IT-Spezialist neigt zur Beschränkung auf Formalia und nicht auf die wesentlichen Aspekte der Beurteilung verfahrensbedingter Qualitäten der zu verbindenden Daten.

Gleichwohl spricht pro Fusionen, dass sie auf Grund der Anforderungen des Marktes entwickelt wurden - ohne Nachfrage gäbe es keine Fusionen, für die vor allem zwei Gründe ausschlaggebend sind: detaillierte Bestimmung von Zielgruppen vergleichbar in mehreren Studien und zumindest für eine strategische Entscheidung Erstellung von Media-Mix-Plänen.

Kontra Fusionen oder 'Con-Fusion' sind häufig irrationale Argumente, die Fusionsverfahren tendenziell in die Ecke von Daten-Manipulationen stellen. Aber: Es werden keine neuen Daten generiert, sondern die Informationstiefe einzelner Studien wird mit akzeptierten mathematisch-statistischen Verfahren wesentlich erweitert.

Natürlich ist es unglücklich, dass das Ergebnis von Fusionen nicht direkt ohne eine Single-Source-Studie validiert werden kann, aber - wie eingangs erwähnt - wenn eine Single-Source-Studie vorliegt, benötigt man keine Fusion.

Häufig wird der Nutzwert von Fusionen bezweifelt, da man schließlich jahrelang ohne die Ergänzung von Datenbeständen ausgekommen sei. Gegen dieses Argument spricht z. B. bei Media-Planung nur über Analogieschlüsse, dass die Streuverluste wesentlich größer sein müssen, d. h. der ROI (Return on invest) von Werbeausgaben wird deutlich geringer sein.

Keine der großen Markt-Media-Studien in Deutschland verzichtet daher auf den Einsatz von Fusionsverfahren. Bei der AWA, der jährlich erscheinenden Allensbacher Werbeträger-Untersuchung mit ca. 21.000 Befragten, werden fehlende Informationen für Kernvariablen der Zielgruppenbestimmung für Medien-Selektion und Media-Planung, wie z. B. Angaben zum Einkommen, 'ergänzt'. Dazu werden geprüfte heuristische Verfahren eingesetzt.

Die zentralen Referenz-Dateien für die Planung der Schaltung von Funk-Spots resp. von Anzeigen sind die zweimal pro Jahr erscheinenden Studien der Arbeitsgemeinschaft Media-Analyse - ein Zusammenschluss von Werbeträgern, Werbetreibenden und Agenturen.

Für die Presse MA werden pro Welle ca. 39.000 Befragungen ausgewiesen. Seit 2004 werden die Daten über ein Split-Modell erhoben, da - ohne wesentlich andere Konsummerkmale - die Abfrage zur Nutzung von Print-Medien (Zeit-

schriften und Zeitungen) den zumutbaren Erhebungsumfang weit überschritten hätte:

- Split A: 13.000 Befragte mit Abfrage der Titelgruppen 1 und 2
- Split B: 13.000 Befragte mit Abfrage der Titelgruppen 1 und 3
- Split C: 13.000 Befragte mit Abfrage der Titelgruppen 2 und 3
  (1 = 53 Titel / 2 = 61 Titel / 3 = 55 Titel)

Damit vergleichend für alle Titel wie in einem Quasi-Single-Source-Datenbestand Zählungen und Analysen durchgeführt werden können, werden die jeweils fehlenden Titel-Informationen fusioniert. Wobei der besondere Charme dieses Verfahrens darin liegt, dass gemeinsame Variablen einbezogen werden können, die hoch mit dem medienkonsumptiven Verhalten korrelieren, z. B. bei Fusion der Titelgruppe 3 in Split A kann zurückgegriffen werden auf Titelgruppe 1 in Split B und Titelgruppe 2 in Split C, so dass die fallweise Zuordnung optimiert werden kann und die 'nationale' Media-Währung erhalten bleibt.

Aus einem kumulierten 3-Jahres-Bestand der obigen Presse MA wird der so genannte Tageszeitungs-Datensatz gebildet, damit hinreichend auswertbare Fälle auch für regionale Tageszeitungen zur Verfügung stehen. Es handelt sich zunächst um ein rein additives Zusammenführen. Die Aktualität des Gesamtbestandes wird sichergestellt durch Fusion der Medien-Nutzungs-Werte der letzten beiden Jahre in den Gesamtbestand, d. h. in die Erhebungsjahre 1 bis 3. Hier mag man durchaus kritisch anmerken, dass damit das Ziel der Ausweisung regionaler Daten z. B. auf Verbreitungsgebiets-Ebene für Zeitungen nach der Fusion zwar gegeben ist, aber die Zuverlässigkeit der Ergebnisse sich nur auf die beiden letzten Jahre beziehen kann, denn für die Validität ist das Sample maßgebend, in dem die Informationen originär erhoben wurden.

Zusätzlich mag es merkwürdig anmuten, dass auch die Originär-Informationen in dem Referenz-Datenbestand der letzten beiden Jahre per Fusion überschrieben werden, d. h. es werden - wenn auch in geringem Ausmaß - Leser im Referenz-Datenbestand zu Nicht-Lesern gemacht und vice versa.

Für strategische Entscheidungen wird auf Basis der Print-Tranche der MA eine Intermedia-Datei erstellt, in die per Fusion Fernseh-Nutzungs-Daten für Zeitschnitte (z. B. für ARD 1/2-Stunden-Schnitte in der Zeit von 17.30 bis 20.00 Uhr) auf Basis der sekundengenau gemessenen TV-Meter-Daten der Arbeitsgemeinschaft Fernsehforschung eingespielt werden.

Zusätzlich werden Funk-Reichweiten fusioniert. Hier mögen Kritiker einwenden, dass Fusionen aus Dateien, die auf Basis unterschiedlicher Stichproben und mit unterschiedlichen Verfahren erhoben wurden (Print: Face-to-Face / Funk: CATI = Computer Assisted Telephone Interviewing / TV: Panel), keine hinreichende Zuverlässigkeit aufweisen können.

Die Praxis zeigt das Gegenteil. Die Daten sind in sich stimmig und in hohem Maße nutzbar, wenn auch z. B. in Bezug auf TV keine Planungen auf Basis von

Werbeblöcken oder einzelnen Sendungen, Genres möglich sind. Dies wird durch den Aufbau der Intermedia-Datei auch nicht intendiert. Ziel ist u. a. die Ableitung von Parametern für strategische Entscheidungen der Verteilung von Media-Budgets nach Media-Gattungen.

Die Media-Analyse bleibt die Währung - eine Planung nach Konsummerkmalen wie Marken-Bekanntheit, Verwendung/Verbrauch von Konsumgütern und vieles mehr ist nur eingeschränkt möglich. Die VA - Verbraucher-Analyse - (1.650 Konsummerkmale mit insgesamt ca. 12.000 Ausprägungen und jährlich 31.000 Befragungen), die TdW - Typologie der Wünsche - (2.900 Konsummerkmale mit insgesamt ca. 14.000 Ausprägungen und jährlich 20.000 Befragungen) werden für die Print-Medien-Nutzung 'angepasst' an die jeweils aktuelle MA Presse: eine Anpassung nicht per Gewichtung, sondern per Fusion.

Zusätzlich werden aus der MA Intermedia die Funk- und TV-Daten 'übernommen', d. h. eine Fusion bereits fusionierter Daten. Auch hier zeigt die Praxis, dass diese Daten eine hohe Akzeptanz bei den Anwendern haben - mithin 'brauchbar' im Hinblick auf weit reichende monetäre Entscheidungen sind.

Für detailliertere Betrachtungen stehen darüber hinaus über 20.000 TV-Einzel-Wahrscheinlichkeiten - gebildet aus dem ACNielsen-TV-Panel - zur Verfügung (z. B. Stunden-Schnitte nach Wochentagen für Einzel-Sender).

Spezifisch für Radio wird die VuMA (Verbrauchs- und Media-Analyse) erhoben. Ausgewiesen werden nach einem rollierenden Verfahren ca. jeweils 31.000 Befragte (zweimal jährlich), wobei neue Erhebungspunkte in die jeweils zurückliegenden Wellen fusioniert und darüber hinaus die Fernseh-Reichweiten per Fusion an die der MA Funk angeglichen werden und die TV-Nutzung aus der Intermedia-Datei der MA übernommen wird.

Man mag annehmen, dass diese Festlegung auf eine gemeinsame Währung mit den entsprechend notwendigen Zusammenführungen von Dateien spezifisch deutsch sei.

Dies trifft nicht zu - in allen Gesellschaften sinkt die Antwortbereitschaft und die Ausschöpfung: Single-Source-Studien mit einer den Wünschen der Nutzer entsprechenden Gliederungstiefe sind nicht mehr zu vertretbaren Kosten und mit hinreichender Zuverlässigkeit der Ergebnisse durchführbar.

Das CMDC (Canadian Media Directors Council) hat daher das so genannte ‚Unity Project' zunächst für das Ballungsgebiet Toronto beauftragt, um letztendlich Entscheidungsgrundlagen dafür zu erhalten, für Gesamt-Kanada Daten aus unterschiedlichsten Quellen zusammenzuführen. Insgesamt wurden 10 Fusionen durchgeführt, um singulär erhobene Konsumdaten in allen Dateien zur Verfügung zu haben - wobei der Währungs-Charakter der jeweiligen Empfänger-Studien nicht tangiert werden durfte.

Ausgangsbasis sind folgende Dateien:

- PMB (Print) 4.769 Fälle
- BBM/RTS (Radio) 9.963 Fälle
- NAD Bank (Zeitungen) 4.504 Fälle
- BBM Meter (TV) 927 Fälle
- Nielsen (TV) 2.238 Fälle

Fusioniert wurden folgende Dateien:

Rezipient	<-----	Donor
PMB	<-----	NAD Bank
PMB	<-----	BBM/RTS
NAD Bank	<-----	PMB
BBM/RTS	<-----	PMB
BBM Meter	<-----	PMB
BBM Meter	<-----	NAD Bank
Nielsen	<-----	PMB
Nielsen	<-----	NAD Bank
Nielsen	<-----	BBM/RTS
BBM Meter	<-----	BBM/RTS

Hauptprobleme bei diesen Fusionen liegen in teilweise unzureichenden gemeinsamen Variablen, unterschiedlichen Fragestellungen mit gleichen Inhalten und Unterschieden in den Ergebnissen zwischen den einzelnen Studien, die auf unterschiedliche Erhebungsmodelle, Feldzeiten etc. zurückzuführen sind. Ein Beispiel mag dies verdeutlichen:

- PMB: 1 + movies attended in past 30 days 55,9 %
- BBM/RTS: movies attended at a theatre or drive-in once a month or more 6,5 %

Und: selbstverständlich erfordern gerade im Hinblick auf Donoren-Überhänge die Fusionen eine besondere Sorgfalt.

Die einzelnen Fusionen wurden in Bezug auf die Verteilung für alle übertragenen Variablen auf signifikante Unterschiede vor und nach Fusion mit Hilfe von Chi2 (95 % Niveau) getestet. Als Annahme galt, dass die jeweilige Fusion 'gelungen' ist, wenn nicht mehr als 5 % aller Tests signifikant waren - mithin im 'zufälligen' Bereich lagen. Dies wurde bei den Fusionen erreicht, bei denen die Print-Datei (PMB) nicht beteiligt war. Bei Fusionen im Zusammenhang mit der PMB zeitigten sich viele Auffälligkeiten, für die es jedoch eine einfache Erklärung gibt: die PMB ist stark disproportional nach sozio-ökonomischen Schichten angelegt, die per Gewichtung für eine bevölkerungsrepräsentative Darstellung proportionalisiert werden. Wird dieser Faktor bei der Beurteilung der Fu-

sionen berücksichtigt, zeitigt sich auch hier, dass die Fusionen insgesamt zu zufrieden stellenden Ergebnissen führen.

Gleichwohl gilt, die Originär-Information ist gegenüber der fusionierten zu bevorzugen.

Zu beachten ist, dass die Struktur der Empfänger-Datei auch die Struktur des Fusionsergebnisses bestimmt - d. h. Ergebnisniveaus können sich im Vergleich zwischen Donor und Rezipient verschieben. Dies ist dann unwahrscheinlich, wenn die Stichproben ein vergleichbares Niveau haben. Darüber hinaus sichern die modernen Verfahren in aller Regel, dass die Media-Währungen erhalten bleiben. Bei Ergebnisunterschieden sind wiederum die 'wissenden' Forscher für die Interpretation gefragt.

Jede Fusion ist einzigartig. Es gibt keine Verfahren nach dem Motto 'One size fits all'. Um es zu wiederholen: zu berücksichtigen sind Stichprobendesign, Gewichtung, genügend gemeinsame in Bezug auf die zu fusionierenden Themenbereiche aussagefähige Merkmale.

Uwe Czaia, geb. 1944, Diplom-Soziologe, Tätigkeit als Feldeinsatz- und Studienleiter bei Media-Markt-Analysen, Frankfurt, Gruner + Jahr AG, Hamburg, Geschäftsführer Getas GmbH, Bremen. Seit 1983 selbständig, Geschäftsführender Gesellschafter der CZAIA Marktforschung GmbH. Spezialisiert auf Media- und Kommunikationsforschung, Werbeträger- und Werbemittelforschung, Neue Medien/Internet, Marketingforschung. 1989 Gründung der Tochterfirma IMMEDIATE Gesellschaft zur Entwicklung und zum Vertrieb von Software für Marketing und Mediaplanung mbH. Spezialisiert auf Datenaufbereitungen, multivariate Analysen, Fusionen sowie PC-gestützte Auswertungs- und Mediaplanungs-Software.

Die Media-Analysen
Synopse des Datenbestands und Nutzungschancen für Sekundäranalysen des sozialen Wandels in Deutschland seit 1954[1]

Heiner Meulemann, Jörg Hagenah, Haluk Akinci

Seit 1954 erhebt die Arbeitsgemeinschaft Media Analyse in den Leser- und Media-Analysen die Nutzung von Druck- und elektronischen Medien, um die Werbewirksamkeit einzelner Publikationsorgane zu ermitteln. Nach einer technischen wie inhaltlichen Aufbereitung bieten diese für einen kommerziellen Zweck erhobenen Daten viele Chancen für wissenschaftliche Sekundäranalysen des Wandels der Mediennutzung und der Sozialstruktur in Deutschland. Diese Aufbereitungsarbeiten und Analysechancen sollen im folgenden Beitrag beschrieben werden. Im ersten Teil stehen die technischen Aspekte der Zugänglichkeit im Vordergrund. Im zweiten Teil wird die inhaltliche Erschließung der Daten – die Bildung zusammenfassender Zielvariablen, die Konstruktion von Zeitreihen und die Integration externer Daten – dargestellt.

1 Formale Erschließung der Daten

Die Arbeitsgemeinschaft Media-Analyse (AG.MA), ein Verbund von Medienbetreibern, Werbeagenturen und Werbungstreibenden, beauftragt seit 1954 sozialwissenschaftliche Erhebungsinstitute, in repräsentativen Befragungen der Bevölkerung ab 14 Jahren die Nutzung einzelner Druck- und elektronischer Medien im Tagesablauf zu erheben: in der sog. Leser-Analyse (LA) und der Media-Analyse (MA). LA und MA ermitteln die Werbewirksamkeit jedes einzelnen Publikationsorgans und erstellen eine gemeinsame „Werbewährung", die für die Planung von Programmen und die Platzierung von Werbung erforderlich ist. Das Zentralarchiv für Empirische Sozialforschung der Universität zu Köln (ZA) hat diese Daten von Anfang an archiviert. Das Medienwissenschaftliche Lehr- und Forschungszentrum der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln (MLFZ) hat seit 2003 in Zusammenarbeit mit

1 Teile dieses Artikels erscheinen auf Englisch in „Schmollers Jahrbuch. Journal of Applied Social Science Studies/Zeitschrift für Wirtschafts- und Sozialwissenschaften" in der Rubrik Data Watch, in der regelmäßig Datensätze vorgestellt werden.

dem ZA die Daten so aufbereitet, dass sie für wissenschaftliche Sekundäranalysen mit Standard-Analyseverfahren zur Verfügung stehen.

1.1 Dokumentation

Erhebungsform, Erhebungsdichte und erhobene Medienarten der LA und der MA sind in Tabelle 1 mit ihren Wandlungen dargestellt. Von 1954 bis 1958 wurde die Nutzung der Pressemedien in den LA alle zwei Jahre und dann bis 1971 jährlich erhoben. Seit 1972 wird in den MA auch die Nutzung der elektronischen Medien erhoben. Seit 1987 werden in der MA nicht mehr alle Medienarten – Radio, Fernsehen, Zeitungen, Zeitschriften und Kino – gemeinsam abgefragt, sondern getrennt für Pressemedien (MA-PM) und elektronische Medien (MA-EM), also Radio und Fernsehen. Seit 1997 werden Daten zur Fernsehnutzung nicht mehr senderspezifisch erhoben und der Schwerpunkt der MA-EM liegt auf der Radionutzung, so dass 2000 die MA-EM in MA-Radio umbenannt wurden. Seit 1997 wird die MA-PM, seit 2000 auch die MA-Radio alle halbe Jahre erhoben.

Tabelle 1 Chronologie der Leser-Analysen (LA) und der Media-Analysen (MA)

Jahre	Studie	Erhebungsform	Erhebungsdichte	Erhobene Medienarten
1954 - 1958	LA	persönlich	zweijährig	Presse
1960 - 1971			jährlich	
1972 - 1986	MA			Presse + Radio/ TV
1987 - 1996	MA-PM			Presse
	MA-EM			Radio/ TV
1997 - 1999	MA-PM		halbjährlich*	Presse Radio
	MA-Radio		jährlich*	
Seit 2000	MA-PM	persönlich**	halbjährlich	
	MA-Radio	telefonisch	halbjährlich***	

PM = Pressemedien (Zeitungen, Zeitschriften); EM = Elektronische Medien (Radio, TV)

* 1998 wurde die MA Radio halbjährlich und die MA-PM nur einmalig durchgeführt

** Seit der 2. Erhebungswelle PM 2004 wird eine 10%-Substichprobe mit der CASI-Methode (Computer-Assisted Self-administered Interview) befragt (www.agma-mmc.de, 2005).

*** Im Jahr 2000 wurde nur eine MA-Radio erhoben.

Die Erhebungsform war von 1972 bis 1999 das persönlich-mündliche Interview. Seit 2000 wird die MA-Radio telefonisch erhoben. Heute werden zu jedem Zeitpunkt ca. 60.000 Personen telefonisch und 30.000 Personen persönlich von mehreren Marktforschungsinstituten befragt. Die jeweils aktuellen Daten

aus dem zurückliegenden Erhebungsjahr stehen zunächst nur den Mitgliedern der AG.MA für planungsrelevante Entscheidungen zur Verfügung und werden nach einer ca. einjährigen Wartezeit an ZA und MLFZ für wissenschaftliche Sekundäranalysen weitergegeben.

Die Originaldatensätze unterscheiden sich hinsichtlich Speicherformat und Größe von anderen sozialwissenschaftlichen Datensätzen. Sie wurden als binäre Dateien gespeichert, die nicht mit Standard-Statistikprogrammen analysiert werden können, und daher in eine entsprechende Dateiform konvertiert werden müssen. Tabelle 2 gibt einen Überblick über die konvertierten Datensätze. Weil die Erhebungen aus bis zu neun erstellten Einzeldateien bestehen, liegen aus den Jahren 1954 bis 2002 im MLFZ-Archiv insgesamt 281 Dateien mit ca. 1,8 Mio. Befragten vor.

Tabelle 2: Ausschnitt aus dem MA-/ LA-Datenbestand des MLFZ

Erhebung		**Erhebungsinformationen**		**ZA-Bestand**	**MLFZ-Bestand**
		Medium	*Fallzahl*	*Anzahl Binär-Dateien*	*Anzahl SPSS-Dateien*
LA	1954	PM	13258	1	1
…	…	…			
LA	1971	PM	17517	1	1
MA	1972/I	EM + PM	14641	1	1
…	…	…			
MA	2002	EM I	54888	9	9
		EM II	61113	9	9
		PM I	26188	9	9
		PM II	26096	9	9
Summe			**ca. 1,8 Mio.**	**281**	**281**

Anmerkungen: MA = Media-Analyse, LA = Leseranalyse; ZA = Zentralarchiv für empirische Sozialforschung Köln; MLFZ = Medienwissenschaftliches Lehr- und Forschungszentrum der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln

1.2 Inhalte

Zentraler Inhalt ist die Nutzung der Medien Radio, Fernsehen (bis 1996), Zeitungen, Zeitschriften, Kino und seit 1997 die Internetnutzung. Die Mediennutzung wird in zwei Komplexen, *unspezifisch* und *spezifisch*, abgefragt. Die unspezifische Abfrage erhebt die generelle Mediennutzung, die spezifische Abfrage die Nutzung von bestimmten Sendern oder Titeln.

Die *spezifische Mediennutzung* wird in vier Schritten abgefragt. *Erstens* wird im *Generalfilter* für jeden Sender oder Pressetitel gefragt, ob er schon mal gehört, gesehen oder gelesen wurde. Bei allen schon mal genutzten Sendern oder Titeln wird *zweitens* im *Zeitfilter* gefragt, wann dies zuletzt geschehen ist. *Drittens* wird mit der *Frequenzabfrage* die Nutzungshäufigkeit jedes Senders oder Titels gemessen. *Viertens* wird für den letzten Tag vor dem Interview der gesamte *Tageslablauf*, insbesondere die Nutzung von Radio und Fernsehen, abgefragt.

Neben der Mediennutzung werden alle üblichen sozial-demographischen Merkmale der Befragten, die technische Ausstattung der Haushalte und einige Freizeit- und Konsumpräferenzen erhoben.

Tabelle 3: Ausschnitt aus der Variablenübersicht „madatsyn 1.0 short“

Identifikationsmerkmale	**Erhebungen**	**ZEITRAUM VARIABLEN**	
Erhebungszuordnung	*Anzahl*	*Erste Erhebung*	*Letzte Erhebung*
Soziodemographische Merkmale			
z.B. Geschlecht	68	LA 54	MA 02 PM II
Interviewsituation			
z.B. Bereitwilligkeit zum Interview	46	MA 79	MA 02 PM II
Einstellungen des Befragten			
z.B. Parteipräferenz des Befragten	22	MA 78	MA 02 PM II
Einstellungen zum Konsum			
z.B „Ich achte bei Körperpflegemitteln auf den pH-Wert“	3	MA 97 EM	MA 99 EM
Einkaufsgewohnheiten des Befragten bei Lebensmitteln und Getränken			
z.B. Kauf in Discountgeschäft	25	MA 78	MA 96 EM
Öffentliche Verkehrsmittel			
z.B. Städtische Straßenbahnhaltestelle	23	MA 76	MA 90 PM
Freizeitverhalten des Befragten			
z.B. Stricken, Häkeln, Selberschneidern	37	MA 87 EM	MA 02 PM II
Reisen		MA 87 EM	MA 02 PM II
z.B. Ferienreisen (Ziel)	46	MA 72/I	MA 02 PM II
Unspezifische Mediennutzung: Hörhäufigkeit pro Zeitabschnitt in einer normalen Woche			
z.B. 07:00 - 08:00 Uhr	37	MA 87 EM	MA 02 PM II
Spezifische Mediennutzung: Wöchentliche Zeitschriften (Lesehäufigkeit)			
z.B. Spiegel, Der	31	MA 76	MA 02 PM II

Das MLFZ hat alle Variablen der LA und MA in einer Excel-Liste „madatsyn 1.0“ inventarisiert, die alle der ca. 32.000 mindestens einmal erhobenen Einzelvariablen mit ihren Erhebungszeitpunkten enthält, so dass sich für jede Variable für die Zeit zwischen 1972 und 2002 ermitteln lässt, wie oft und wann sie abgefragt wurde und welche Variablen-Nummer sie zu jedem Erhebungszeitpunkt hat. Zu „madatsyn 1.0“ wurde eine Kurzfassung „madatsyn 1.0 short“ erstellt, die nur den ersten und letzten Erhebungszeitpunkt und die Anzahl der Erhebungen enthält. Tabelle 3 gibt einen Ausschnitt aus „madatsyn 1.0 short“, der für jeden Variablenbereich ein Beispiel enthält.

Drei Variablengruppen erlauben Analysen des sozialen Wandels jenseits der Mediennutzung und werden im Folgenden genauer dargestellt. *Erstens* werden *soziodemographische* Variablen, über die Tabelle 4 einen Überblick gibt, seit 1954 erfragt – also noch vor dem erstmals 1958 erhobenen Mikrozensus. Ihre jährliche Erhebung in Stichproben von bis zu 60.000 Befragten erlaubt eine tief gegliederte und zeitlich genaue Analyse des sozialstrukturellen Wandels der alten Bundesrepublik über fast fünfzig Jahre. *Zweitens* ermöglichen es die Angaben zur *Interviewsituation*, über die Tabelle 5 einen Überblick gibt, den Wandel der Bereitwilligkeit zum Interview und des Interesses am Interviewthema seit 1979 in Abhängigkeit von soziodemographischen Variablen zu untersuchen.

Tabelle 4: Soziodemographische Variablen. Ausschnitt aus „madatsyn 1.0 short“

Identifikationsmerkmale	**Erhebungen**	**ZEITRAUM VARIABLEN**	
Erhebungszuordnung	*Anzahl*	*Erste Erhebung*	*Letzte Erhebung*
Merkmale des Befragten			
Geschlecht	68	LA 54	MA 02 PM II
Alter des Befragten	68	LA 54	MA 02 PM II
Schulbildung des Befragten (5-stufig)	68	LA 54	MA 02 PM II
Berufstätigkeit des Befragten (12-stufig)	68	LA 54	MA 02 PM II
Nettoeinkommen des Befragten (13-stufig)	65	LA 54	MA 02 PM II
Konfession des Befragten (3-stufig)	61	LA 62	MA 02 PM I
Familienstand des Befragten (5-stufig)	66	LA 54	MA 02 PM I
Merkmale des Haushalts			
Personen im Haushalt insgesamt (12-stufig)	66	LA 60	MA 02 PM II
Personen mit eigenem Einkommen (7-stufig)	66	LA 54	MA 02 PM I
Haushaltsnettoeinkommen (12-stufig)	67	LA 58	MA 02 PM II

Tabelle 5: Interviewinformationen. Ausschnitt aus „madatsyn 1.0 short“

Identifikationsmerkmale	**Erhebungen**	**ZEITRAUM VARIABLEN**	
Erhebungszuordnung	*Anzahl*	*Erste Erhebung*	*Letzte Erhebung*
Technische Angaben zum Interview			
Tag des Interviews	57	LA 69	MA 02 PM II
Institut (9-stufig)	46	MA 75	MA 02 PM II
Interesse am Befragungsthema (4-stufig)	46	MA 79	MA 02 PM II
Bereitwilligkeit zum Interview (4-stufig)	46	MA 79	MA 02 PM II
Zustandekommen der Einkommensangaben des Befragten (4-stufig)	47	MA 76	MA 02 PM I
Zustandekommen der Einkommensangaben des Haushalts (4-stufig)	45	MA 76	MA 02 PM I
Anzahl Besuche, um das Interview zu erreichen (10-stufig)	46	MA 77	MA 02 PM II
Interviewdauer in Minuten	43	MA 82	MA 02 PM II

Drittens erlaubt die Abfrage zum *Tagesablauf*, die in Abbildung 1 dargestellt ist, den Wandel des Zeitbudgets nicht nur der Mediennutzung, sondern auch der Arbeit und Freizeit zu untersuchen. Sie wird seit 1987 für jede Viertelstunde zwischen 5 und 24 Uhr des letzten Tages vor dem Interview erhoben. Obwohl der Schwerpunkt die Nutzung einzelner Radio- und Fernsehsender ist (siehe Spalte 54 - 60 für das Radio), werden auch andere Aktivitäten „zu Hause“ und „außer Haus“ erhoben (Spalte 1 - 12). In weiteren, hier nicht aufgeführten Spalten wird die Nutzung von Fernsehen (seit 1997 nicht mehr senderspezifisch), Video, Schallplatten/Tonband/Kasseten/CD und PC (seit 1997) erfasst.

TAGESABLAUF von gestern:		**zu Hause**					**außer Haus**							**Radio hören**			
Wochentag von gestern eintragen!		Schlafen	Körperpflege / Anziehen	Essen / Mahlzeiten	Haus- / Berufsarbeit	andere Tätigkeiten / freie Zeit	im Auto unterwegs	Einkaufen / Besorgungen	Berufsarbeit	Schule / Studium	Besuch bei Freunden, Bekannten, Verwandten	Besuch von Kneipen, Gaststätten, Restaurants	andere Tätigkeiten / freie Zeit	SWF 1	SWF 3	S 2 Kultur	S4 Baden-Württemberg
Datum von gestern eintragen!																	
Kärtchen-Nr.		1	2	3	4	5	6	7	8	9	10	11	12	54	55	59	60
5.00 - 5.15	1	X															
5.15 - 5.30	2	X															
5.30 - 5.45	3	X															
5.45 - 6.00	4	X															
6.00 - 6.15	5	X															
6.15 - 6.30	6		X												X		
6.30 - 6.45	7			X											X		
6.45 - 7.00	8			X											X		
7.00 - 7.15	9						X								X		
7.15 - 7.30	10								X								
7.30 - 7.45	11								X								
7.45 - 8.00	12																

Abb. 1: Ausschnitt aus dem Tagesablauf-Fragebogen mit fiktiven Daten

1.3 Publizierte Sekundäranalysen

Für wissenschaftliche Publikationen wurde die MA bisher – vermutlich wegen der binären Datenstruktur der Originaldateien und der Unüberschaubarkeit der Variablenmengen – nur selten genutzt. Die wenigen bisher publizierten Sekundäranalysen lassen sich in vier Blöcke teilen.

Erstens werden seit 1991 (Franz, Klingler & Jäger 1991) jährlich die *aktuellen Eckwerte der Mediennutzung* von Autoren des Auftraggebers der Studien in

der Hauszeitschrift Media Perspektiven[2] berichtet und in der Regel mit Kennwerten der Vorjahre abgeglichen. Beispielsweise vergleichen Klingler und Müller (2004) die Daten der MA 2004 mit denjenigen der Jahre 2001 bis 2003. Franz, Klingler und Jäger (1991, 400 ff.) untersuchten die Entwicklung der Radionutzung von 1968 bis 1990.

Zweitens wurden die Daten genutzt um *medienwissenschaftliche Fragestellungen* zu analysieren. Kubitschke und Trebbe (1992) ermittelten eine medienübergreifende Nutzungstypologie. Weiß und Hasebrink (1995, 1997) untersuchten die Gewohnheiten von Radiohörern in Hamburg. Schönbach, Lauf, Stürzenbecher und Peiser (1997) gingen den Faktoren des Erfolgs von Tageszeitungen nach. Lauf (1999) verglich Zeitungsnutzungskennwerte aus den MA mit denjenigen der Allensbacher Werbeträgeranalyse und der Langzeitstudie Massenkommunikation.

Drittens bieten die Daten die Möglichkeit, den *sozialen Wandel in der Bundesrepublik* nachzuzeichnen. Wahl (1997, 2003) und Risel (2005) zeichneten Veränderungen von Lebensstilgruppen nach. Fachinger (2004) analysierte altersspezifisches Ausgabeverhalten.

Viertens wurden *methodologische Fragen* untersucht. Schnell (1997) untersuchte die Ausschöpfungsquoten der Stichproben. Außerdem wurde der Erhebungswechsel der Radio-Tranche vom persönlichen zum telefonischen Interview aus der Sicht der AG.MA (Klingler & Müller, 2000) beschrieben und ihre Auswirkungen auf Stichproben-Zusammensetzung und Antwortqualität wurden analysiert (Hagenah & Best, im Druck).

Vor 2002 haben also nur wenige Pioniere aus den MA publiziert, die sich trotz technischer Probleme durch den Datendschungel kämpften. Seit 2002 können Autoren auf SPSS-Datensätze zurückgreifen, müssen aber noch inhaltliche Aufbereitungsarbeiten durchführen (Variablenrecherchen, Labeln, Umkodierungen). Dennoch hat die Aufbereitung die Publikation erleichtert, wie im Herausgeberband von Hagenah und Meulemann (im Druck) ersichtlich ist.

2 Inhaltliche Erschließung der Daten

Nachdem die binären MA-Originaldaten in Standard-Analyseformat konvertiert wurden und eine Variablenübersicht erstellt worden ist, sind die MA-Daten zwar formal zugänglich, aber sie lassen sich noch nicht ohne weiteres inhaltlich analysieren. Dazu müssen *erstens* die Nutzungsdaten der einzelnen Sender und Presseorgane unter inhaltlich sinnvollen Gesichtspunkten zu – wie es hier genannt wird – *Zielvariablensummen* zusammengefasst werden. Dann sollen

2 Seit 1997 stehen Abstracts und seit 2000 auch alle Beiträge der monatlich erscheinenden Zeitschrift Media Perspektiven im Internet unter www.ard-werbung.de zum Download bereit.

zweitens bedeutsame *Trends* der Mediennutzung in den Gesamtstichproben und in Alterskohorten in elektronischer Form als Datenreport auf der MLFZ-Homepage veröffentlicht werden. *Drittens* soll damit begonnen werden, den Mediennutzungsdaten *externe Informationen* über Inhalte der Medien (z.B. Programmraster) und Ereignisse der Umwelt zuzuspielen.

2.1 Bildung von Zielvariablensummen

Für eine sozialwissenschaftliche Analyse, die nicht mehr die „Werbewährung" einzelner Sender, sondern die Publikumsneigung zu Senderkategorien ermitteln will, müssen aus den sender- und titelspezifischen Variablen Summen nach theoretischen Kategorien gebildet werden, die wir *Zielvariablensummen* nennen. Fünf solcher Kategorien für die Bildung der Zielvariablen sind in den Zeilen von Tabelle 6 für die jeweils einschlägigen Medien dargestellt. *Erstens* müssen Radio- und Fernsehsender nach ihrem privat- oder öffentlich-rechtlichen Status gebündelt werden. *Zweitens* können Hörfunksender und Zeitungen nach dem Verbreitungsgebiet und ihrer lokalen oder überregionalen Ausrichtung zusammengefasst werden. *Drittens* können Zeitungen in Abonnement- und Straßenverkaufszeitungen untergliedert werden und haben *viertens* unterschiedliche politische Ausrichtungen. Alle vier Medien sollen *fünftens* nach ihrer thematischen Spezialisierung zusammengefasst werden. Z.B. gibt es beim Hörfunk Talkradios mit einem überwiegenden Wortanteil, „Full Service-Sender" mit einer Mischung aus Musik und einem höheren Wortanteil und reine Musiksender, die jeweils nach ihrer „Klangfarbe" – Oldies, Mainstream, Contemporary Hit Radio etc. – unterschieden werden können.

Tabelle 6: Sender- und Titelkategorien mit Beispielen für Zielvariablensummen pro Medium

Titel- und Senderkategorien	Einheitlich für alle Media?	Zielvariablensummen pro Medium (Anzahl)			
		Radio	Fernsehen	Presse: Zeitschriften	Presse: Zeitungen
Rechtsstatus	nein	öffentlich-rechtlich vs. privatrechtlich (2)	öffentlich-rechtlich vs. privatrechtlich (2)	-	-
Region/ Verbreitungsgebiet	nein	Gebühreneinzugsgebiete: z.B. RB, NDR, WDR … (10)	-	-	regional/lokal vs. überregional (2)
Vertriebsart	nein	-	-	-	Abo- und Straßenverkaufszeitungen (2)
Politische Ausrichtung	nein	-	-	-	rechtskonservativ, links-liberal, neutral (3)
Thematische Spezialisierung	ja	z.B.Talkradio, Full Service Sender, Musiksender (8)[1]	Voll-, Spartenprogramme (4)[2]	z.B. Frauenzeitschriften, Wirtschaftsmagazine (9)[3]	Qualitäts-, Boulevard- und Agenturzeitungen (3)

[1] die Musiksender unterteilen sich in sechs Klangfarben: Mainstream Adult Contemporary, Mainstream Contemporary Hit Radio, Mainstream Rock/Album oriented Rock, Easy Listening/ Beautiful Music, Melodie-Schlager, Oldies

[2] drei Spartenprogramme: Sport, Nachrichten und Zuschaueranrufe/Spiele (keine Musiksender)

[3] Weitere Zeitschriftenarten: Nachrichtenmagazine, Populäre Illustrierte, Programmzeitschriften, Männermagazine, Jugendmagazine, Special-Interest-Zeitschriften und Very-Special-Interest-Zeitschriften

Wie die Zielvariablensummen erstellt werden, soll für die Kategorie Rechtsstatus anhand eines fiktiven Beispiels gezeigt werden. Die Tabelle 7 enthält in den Spalten 1 - s die ersten vier und den letzten der 125 Sender sowie in den Spalten 7 und 8 die beiden neu zu berechnenden Zielvariablensummen. Außerdem ist für jeden Sender die Information öffentlich-rechtlich (ör) oder privat (pr) in der Zeile „Rechtsstatus“ zu finden. Die Zeilen sind geordnet nach den Variablenbereichen (a) Generalfilter und (b) Zeitfilter. Die Ergebnisse zum Abfrageblock

(a) Generalfilter mit den Antwortmöglichkeiten 1 = „schon mal gehört“ und 0 = noch nicht gehört befinden sich für die Person 1 an dieser Stelle. Die öffentlich-rechtlichen Sender 3 und 4 wurden demnach schon mal gehört, so dass die Summe schon mal gehörter Sender für Person 1 = 2 ist. Diese Summe „gför = 2“ befindet sich in der Spalte 7 „Zielvariablensumme 1“.

Tabelle 7: Ermittlung der Zielvariablensummen (Radio) pro Senderkategorie Rechtsstatus (privat vs. öffentlich-rechtlich)

Spalte/ Zeile	1	2	3	4	…	s	7	8
	Sender 1	Sender 2	Sender 3	Sender 4	…	Sender 125	Zielvariablensumme 1	Zielvariablensumme 2
Rechtsstatus	privat	privat	öffentlich-rechtlich	öffentlich-rechtlich	…	privat	öffentlich-rechtlich	privat
a.)	*Generalfilter (GF): 1 = schon mal gehört; 0= nicht gehört*						GF Summe bekannter ör-Sender	GF Summe bekannter pr-Sender
1	1	0	1	1	…	1	gför = 2	gfpr = 2
…	…	…	…	…	…	…	…	…
n	1	1	1	0	…	1	gför = 1	gfpr = 3
b.)	*Zeitfilter (ZF): 1 = länger her; 2 = 2-4 Wochen her; 3 = letzte 14 Tage*						ZF Hörintensität ör	ZF Hörintensität pr
1	1	0	3	2	…	2	zför = 5	zfpr = 3
…	…	…	…	…	…	…	…	…
n	3	1	2	0	…	3	zför = 2	zfpr = 7

Anmerkungen: s = Sender; ör = öffentlich-rechtlich; pr = privat; gför = Generalfilter Summe bekannter öffentlich-rechtlicher Sender; gfpr = Generalfilter Summe bekannter privater Sender; zför = Zeitfilter Hörintensität öffentlich-rechtlicher Sender; zfpr = Zeitfilter Hörintensität privater Sender

2.2 Erstellung von Zeitreihen

Die gebildeten Zielvariablensummen sollen für die gesamte Zeitspanne in drei Abbildungstypen elektronisch veröffentlicht werden; Beispiele dazu werden im Folgenden für wenige Zeitpunkte vorgestellt.

Der *erste Abbildungstyp* stellt eine *Zielvariablensumme über Zeit* dar. Das Beispiel in Abbildung 2 enthält die Zielvariablensumme thematische Spezialisierung für den Zeitraum von 1972 bis 1990, die aus den Angaben zum Generalfilter gebildet wurde. Gezeigt wird die Entwicklung des Bekanntheitsgrades von Qualitäts- und Boulevardzeitungen. Auffällig sind der sinkende bzw. stagnierende Bekanntheitsgrad beider Zeitungstypen im Jahre 1980 und der darauf folgende deutliche Anstieg. In der geplanten jährlichen Darstellung ließe sich fest-

stellen, ob bei beiden Zeitungsarten zwischen 1976 und 1979 ein kontinuierliches Abfallen und nach 1980 ein entsprechender Anstieg zu beobachten ist. Dies könnte den „Knick" möglicherweise erklären. Ansonsten müsste recherchiert werden, ob er unter Umständen methodische Gründe hat. Fraglich bleibt auch, warum Qualitätszeitungen im Vergleich zwischen 1985 und 1990 konstante Werte aufweisen, während Boulevardzeitungen eher bekannter werden. Dies könnte möglicherweise eine erste Folge einer Boulevardisierung durch das private Fernsehen sein, das seit Mitte der 1980er ausgestrahlt wird.

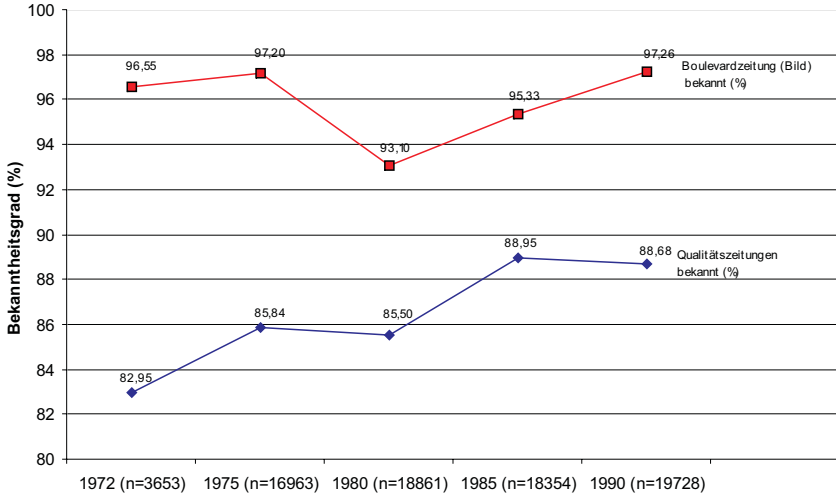


Abb. 2: Bekanntheitsgrad von Qualitätszeitungen und Boulevardzeitungen in %, 1972-1990 (Erster Abbildungstyp: Zielvariablensumme über Zeit)

Der *zweite Abbildungstyp* stellt *Zielvariablensummen über Zeit nach soziodemographischen Merkmalen* dar. In Abbildung 3 wird der Bekanntheitsgrad von Qualitätszeitungen 1972-1990 nach Bildung gezeigt. Auffällig ist, dass der Bekanntheitsgrad von Qualitätszeitungen bei niedrig gebildeten Personen von 1975 bis 1980 deutlich sinkt und bei den anderen beiden Gruppen steigt. Offenbar ergibt sich der „Knick" der Bekanntheit von Qualitätszeitungen in Abbildung 2 vor allem aus den Angaben der Hauptschüler. Ab 1975 ergibt sich eine Kluft zwischen den Bildungsgruppen: Die besser Gebildeten kennen Qualitätszeitungen deutlich häufiger als die Hauptschulabsolventen. Dies muss in einer genauen Betrachtung nach Jahren überprüft werden.

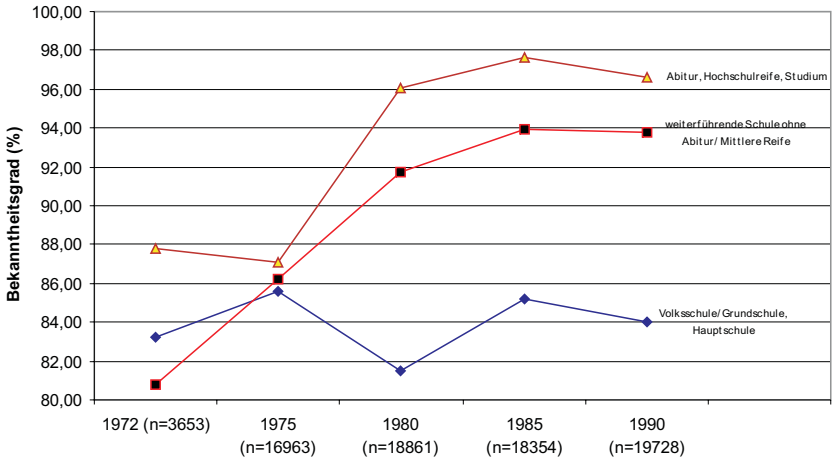


Abb. 3: Bekanntheitsgrad von Qualitätszeitungen in Bildungsgruppen in %, 1972-1990 (Zweiter Abbildungstyp: Zielvariablensumme über Zeit nach soziodemographischen Merkmalen)

Der dritte Abbildungstyp erfasst *Zielvariablensummen über Zeit nach Geburtskohorten*. Abbildung 4 beschreibt die Entwicklung des Bekanntheitsgrades von Qualitätszeitungen in acht Kohorten. Über alle Kohorten hinweg ist der beschriebene „Knick" im Jahr 1980 zu erkennen. Dieser müsste mit Hilfe der benachbarten Jahre genauer untersucht werden. Interessant erscheint vor allen Dingen die Annäherung bei jüngeren Geburtsjahren. K4 bis K8 unterscheiden sich von Anfang an über die Jahre kaum, während bei K1 bis K3 die 1972 festzustellende Bekanntheitsschere bis 1990 zunehmend kleiner wird. Dies ist möglicherweise das Ergebnis eines Selektionseffektes, da Personen mit höherer Bildung erfahrungsgemäß länger leben; es könnte überprüft werden, indem die Zusammensetzung der Kohorte über die Zeit untersucht wird.

Die bisherigen Zeitreihen beziehen sich auf Mediennutzungsentwicklungen. Ein *vierter Abbildungstyp Soziodemographie über Zeit nach Geburtskohorten* erfasst die Entwicklung soziodemographischer Merkmale. Abbildung 5 zeigt, dass bei den jüngeren Kohorten K6 bis K8 der Anteil von Personen mit Abitur deutlich höher ist als bei den vor 1940 geborenen Bundesbürgern; sie demonstriert die oft untersuchte Tatsache der Bildungsexpansion, die sich in den geplanten jährlichen Zeitreihen noch exakter bestimmen ließe. Diese Ergebnisse können auch mit entsprechenden Mikrozensus-Daten verglichen werden.

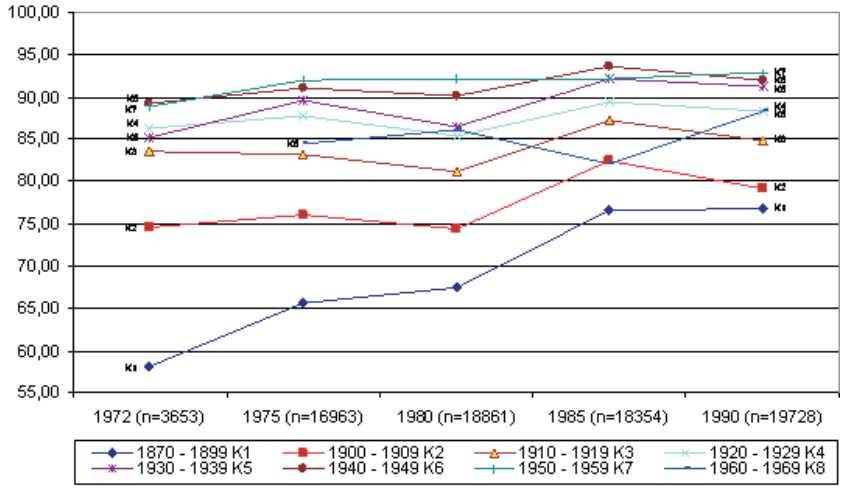


Abb. 4: Kohortenanalysen für den Bekanntheitsgrad von Qualitätszeitungen in %, 1972-1990 (Dritter Abbildungstyp: Zielvariablensumme über Zeit nach Geburtskohorte)

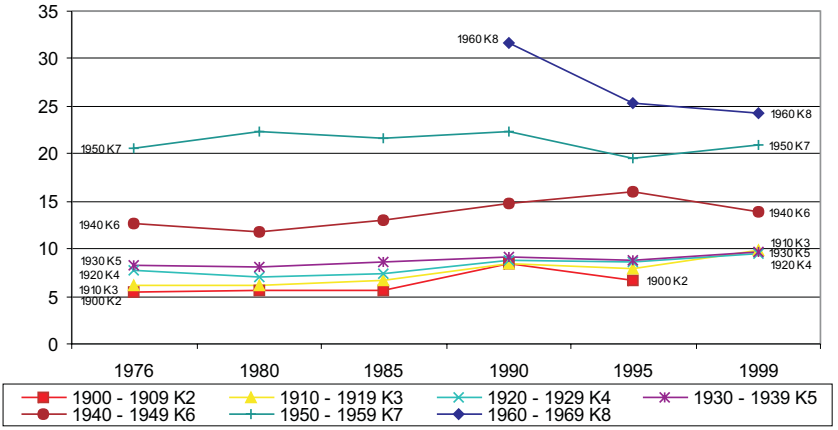


Abb. 5: Kohortenanalyse der Personen mit Abitur in %, 1976-1999 (Vierter Abbildungstyp: Soziodemographisches Merkmal nach Zeit in Geburtskohorte)

2.3 Integration von externen Daten

Die Mediennutzung unterliegt *externen* Einflüssen des Medien-Angebots und der Umwelt, die nicht in den MA-Daten enthalten sind: also der Programme und des Geschehens, die ins Programm eingehen. Die Quellen dazu sind in Tabelle 8 nach Informationsnähe und Informationsstufe dargestellt.

Tabelle 8: Externe Einflüsse auf die Mediennutzung, geordnet nach Informationsnähe und Informationsstufe

Informationsnähe	**Individualdaten**	**Aggregatdaten**	
Informationsstufen	*Programme und Inhalte*	*Medienorganisation*	*Ereignisse*
1. niedrig	*Senderart*: Voll- vs. Spartenprogramm/ überregionale vs. regionale Zeitungen	*Grundordnung*: Einführung Duales Rundfunksystem	*global*: Katastrophen, Kriege, olympische Spiele
2. mittel	*Programmraster* (Information, Unterhaltung)	*Angebotsstruktur*: Entstehung von Spartensendern/Gründung von Zeitungen, Zeitschriften	*national*: Bundestagswahl, Bundesliga-Spiele
3. hoch	*Inhalte* von spezifischen Sendungen /Titeln	*Anbieterstrategie*: Organisatorische Änderungen bei Einzelsendern/Zeitungen/Zeitschriften	*regional*: Landtagswahlen

Die *Informationsnähe* erfasst, ob Informationen Nutzern individuell zugeschrieben werden können oder nicht. Wenn Informationen zu *Programmen und Inhalten* mit den gegebenen Daten der Nutzung verbunden werden, wird der einzelne Nutzer charakterisiert und die Basis der *Individualdaten* erweitert; wenn die Entwicklung der *Medienorganisation* oder *Ereignisse* der politischen und sozialen Umwelt verfolgt werden, wird jeder Nutzer eines Zeitpunkts in gleichem Maße charakterisiert und eine zweite Datenbasis, nämlich von *Aggregatdaten*, geschaffen. Wir konzentrieren uns vorerst auf eine Datenintegration auf Individualdatenebene, wollen aber die zusätzlichen Möglichkeiten, die durch eine Integration von Aggregatdaten entstehen, im Blick behalten.

Die *Informationsstufe* ergibt sich aus der Genauigkeit, mit der Informationen über das Angebot erhoben werden. Mit steigender Informationsstufe werden Programme und Inhalte über Senderart, Programmraster oder Sendungsinhalte charakterisiert. Wie bereits bei der Bildung der Zielvariablensummen angedeutet, wollen wir in einem ersten Schritt Inhalte auf der niedrigsten Informationsstufe integrieren. Als nächstes wollen wir die Integration von Programmrastern

austesten. Die Integration von stärker differenzierenden Inhaltsanalysen wollen wir lediglich vorbereiten.

3 Zusammenfassung und Ausblick

Um eine inhaltliche Analyse der MA-Daten zu erlauben, wollen wir theoretisch begründete Zielvariablensummen bilden und in Trend- und Kohortenanalysen darstellen. Weiterhin ist geplant, die Mediennutzungsdaten um externe Daten des Programmrasters zu erweitern, die wir im Augenblick zusammenstellen. Außerdem recherchieren wir weitere externe Datenquellen. Beispielsweise ist es denkbar, bereits vorliegende Inhaltsanalysen gesendeter Programme sekundäranalytisch mit MA-Daten zu verknüpfen. Auch der Wandel der Medienorganisation und Ereignisse der politischen Umwelt sollten erfasst werden, wofür bestehende Ereignisdatenbänke wie der Media Tenor (2005) genutzt werden können.

Literatur

Fachinger, U., 2004: Einkommensverwendung im Alter. Expertise für die Sachverständigenkommission. 5. Altenbericht der Bundesregierung. Berlin: Deutsches Zentrum für Altersfragen. Verfügbar über World Wide Web: http://www.bmfsfj.de/RedaktionBMFSFJ/Abteilung3/Pdf-Anlagen/fachinger-einkommensverwendung-im-alter.pdf, gelesen am 09.02.2005.

Franz, G., Klingler, W. & Jäger, N., 1991: Die Entwicklung der Radionutzung 1968 bis 1990. In: *Media Perspektiven*, 6/1991, 400-410.

Hagenah, J. & Best, H., im Druck. Die Rolle von Auswahl- und Befragungsverfahren am Beispiel der Media-Analyse. Grundgesamtheit und Inhalte im Vergleich zwischen telefonisch und persönlich-mündlich erhobenen Daten. In: *V. Gehrau; B. Fretwurst; B. Krause & G. Daschmann (Hrsg.). Auswahlverfahren in der Kommunikationswissenschaft*. Köln: Herbert von Halem Verlag.

Klingler, W. & Müller, D. K., 2004. MA 2004 Radio II: Hörfunk behauptet Stärke. In: *Media Perspektiven*, 9/2004, 410-420.

Klingler, W. & Müller, D. K., 2000. MA 2000 Radio: Erstmals mit Telefoninterviews erhoben. In: *Media Perspektiven*, 9/2000, 414-426.

Kubitschke, L. & Trebbe, J., 1992. Zur Ermittlung einer medienübergreifenden Nutzungstypologie. Eine explorative Sekundäranalyse der Media-Analyse 1988. In: *Media Perspektiven*, 3/1992, 199 – 212.

Lauf, E., 1999. Primär sekundäranalysiert: Tageszeitungsnutzung in der Media-Analyse, der Allensbacher Werbeträger Analyse und in der Langzeitstu-

die Massenkommunikation. http://www.dgpuk.de/fg_meth/abs14.htm, gelesen am 12.02.2005.

Media Tenor, 2005. Media Tenor. Institut für Medienanalyse. http://www.mediatenor.de/, gelesen am 05.07.2005.

Risel, M., 2005: Westdeutsche Lebensstile Ende des 20. Jahrhunderts. Eine empirische Untersuchung zum Zusammenhang von Sozialstruktur und Lebensstil, Universität Tübingen: unveröffentlichte Magisterarbeit.

Schnell, R., 1997. Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen. Opladen: Leske + Budrich.

Schönbach, K., Lauf, E., Stürzenbecher, D. & Peiser, W., 1997. Faktoren des Zeitungserfolgs. In: *Schönbach, K. (Hrsg.). Zeitungen in den Neunzigern: Faktoren ihres Erfolgs. 350 Tageszeitungen auf dem Prüfstand*. Bonn: ZV Zeitungsverlag Service GmbH.

Wahl, A., 1997: Strukturierte Pluralität. Lebensstile zwischen vertikalen Strukturbedingungen und intervenierenden Faktoren, Frankfurt am Main/Berlin u.a.: Peter Lang.

Wahl, A., 2003. Veränderung von Lebensstilen. Frankfurt a.M.: Campus Verlag GmbH.

Weiß, R. & Hasebrink, U., 1995. Hörertypen und ihr Medienalltag. Eine Sekundärauswertung der Media-Analyse 94 zur Radiokultur in Hamburg. Berlin: Vistas Verlag GmbH.

Weiß, R. & Hasebrink, U., 1997. Hörertypen und ihr Medienalltag. Plädoyer für eine hörerzentrierte Nutzungsanalyse. In: *Publizistik*, 42/2, 164 – 180.

Ergebnisse des Zensustests
Einfluss von Dubletten auf die Qualität der Melderegister

Hans Gerd Siedt

Vorbemerkung

Gegenstand des Vortrages sind ausgewählte Ergebnisse des Zensustests. Es wird über den Teil des Projektes Zensustest berichtet, der den Zusammenführungen von Daten vorausgehen sollte und in dem man sich vergewissert, dass der Datenbestand „sauber“ ist in dem Sinne, dass jedes Objekt oder jede Person nur einmal im Datenbestand enthalten sind.

Doppelerfassungen sind bei Volkszählungen - unabhängig von der Methode - immer ein Problem. Bei Volkszählungen, die sich der Zählermethode bedienen, werden Doppelerfassungen erst im Nachgang der Erhebungen mit Hilfe von Stichprobenerhebungen zur Qualitätskontrolle erkannt. In der Regel konnten die Sätze mehrfach gezählter Personen nicht mehr aus den Daten entfernt werden, weil das Zensusergebnis schon amtlich verkündet war und/oder Verfahren zur Bereinigung des Zensus aufgrund von Stichprobenergebnissen erst in den letzten Jahren entwickelt und angewandt wurden. In den USA wurden beim 2000er Zensus 1,5 % der Bevölkerung mehrmals gezählt.

Bei einem registergestützten Zensus ist die Möglichkeit gegeben, die Register, die ausgezählt werden sollen, um Dubletten zu bereinigen. Welche Erfahrungen beim Zensustest in diesem Zusammenhang gemacht wurden und wie beim Zensus verfahren werden soll, ist Gegenstand meines Vortrags.

1 Ziele und Anlage des Zensustests

Die Gründe für den Methodenwechsel beim Zensus liegen in den Erfahrungen und vor allem Wahrnehmungen der 87er Volkszählung. Ein Zensusverfahren, das alle Bürgerinnen und Bürger unter Einsatz von Zählern befragt, ist mit Schlagworten wie „hohe Kosten, geringe Akzeptanz und antiquierte Methode“ besetzt und hat seit 1987 ein negatives Image.

Ein Meilenstein für den Methodenwechsel beim Zensus war die 1996 getroffene Entscheidung der Bundesregierung gegen eine herkömmliche Volkszählung im Rahmen der EU-Zählungsrunde 2001. Diese Entscheidung ist 1998

vom Parlament über alle Fraktionen hinweg sowie von der Innenministerkonferenz bestätigt worden. In Vorbereitung dieses Methodenwechsels wurden in den Jahren 2001 bis 2003 auf der Grundlage des Zensusvorbereitungsgesetzes vom 27. Juli 2001 (BGBl. I S. 1882) Zensus-Testhebungen durchgeführt und ausgewertet.

Das Untersuchungsdesign des Zensustests ist der folgenden Abbildung zu entnehmen.

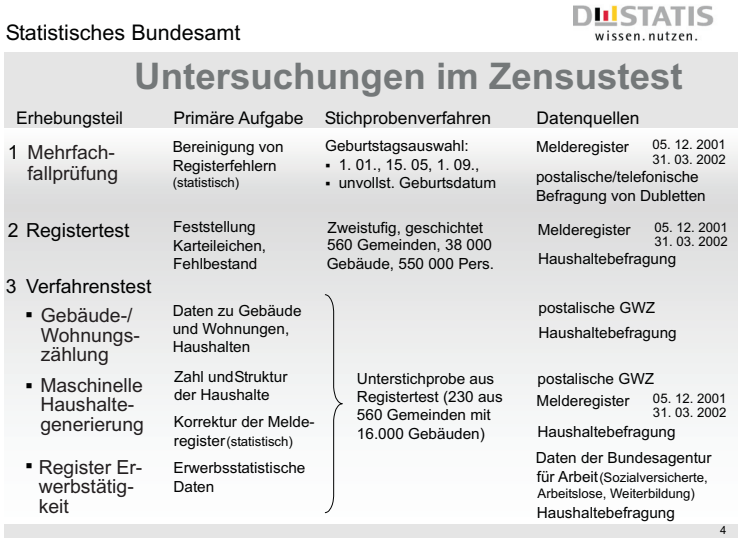


Abb. 1: Untersuchungen im Zensustest

Im Wesentlichen sollten über folgende Sachverhalte zuverlässige Erkenntnisse erlangt werden:

- die Qualität der Melderegister im Hinblick auf Über- und Untererfassungen;
- den Wirkungsgrad von Verfahren zur statistischen Bereinigung der Melderegister um Mehrfachfälle, Übererfassungen und Fehlbestände sowie über
- die Unterschiede in den Ergebnissen zwischen einer postalischen Erhebung der Wohnungsdaten bei den Gebäude-/Wohnungseigentümern (GWZ) und deren Erhebung durch eine direkte Befragung der Haushalte (Wohnungsnutzer) über Erhebungsbeauftragte;
- die Möglichkeiten der Weiterentwicklung des Verfahrens der maschinellen Generierung von Haushaltszusammenhängen durch kombinierte Nutzung

der Melderegisterdaten und der in der Gebäude- und Wohnungszählung erhobenen Daten sowie über die Zuverlässigkeit der Generierungsergebnisse;

- die Nutzungsmöglichkeiten und Qualität der Personenregister der Bundesagentur für Arbeit.

Auf der Grundlage der Testerhebungen wurden verschiedene Modelle eines registergestützten Zensus entwickelt. Die Ergebnisse des Zensustests und die Modellbeschreibungen sind in dem Bericht der Statistischen Ämter des Bundes und der Länder zusammengefasst und Ende 2003 den Dienstaufsichtsbehörden der Statistischen Ämter zugeleitet worden.[1] Die Statistik hat daraufhin den Auftrag erhalten, die methodischen Vorarbeiten für den registergestützten Zensus mit Priorität fortzuführen.

2 Dublettenprüfung im Zensustest

Die Melderegister sind bei einem registergestützten Zensus die wichtigste Datenquelle, aus der die demographischen Grunddaten je Person (Geschlecht, Alter, Familienstand, Staatsangehörigkeit, Wohnort) gewonnen werden sollen.

Zur Prüfung auf Mehrfachmeldungen in den Melderegistern wurden von allen Meldebehörden zu den Stichtagen Datensätze der Einwohner, die am 1. Januar, 15. Mai oder 1. September geboren sind sowie der Einwohner mit unvollständigem Geburtsdatum angefordert. Die Anforderung der Daten von Personen mit unvollständigem Geburtsdatum ist darin begründet, dass ein Teil der im Ausland geborenen Bürger das Geburtsdatum erfahrungsgemäß nur mit Monat/Jahr bzw. nur mit dem Geburtsjahr benennen kann. Eine andere Erfahrung ist, dass Personen mit entsprechenden Wissenslücken auf den 1. Januar ausweichen. Dies erklärt die Auswahl des Geburtsdatums 1. Januar.

Erwartungsgemäß weist die Gruppe der am 1. Januar Geborenen mit 386.000 eine deutlich stärkere Besetzung aus. Die beiden anderen Geburtstage (15. Mai, 1. September) sind mit Personenzahlen um die 250.000 vertreten. Die Zahl der Personen mit unvollständigem Geburtsdatum beläuft sich mit 90.500 auf weniger als 10 % der 971.000 Datensätze.

Die Melderegister werden in Deutschland gemeindeweise geführt. Es gibt gut über 12.000 Gemeinden und um die 5.400 Meldebehörden.[2]

Bei dezentral geführten Melderegistern ist nicht auszuschließen, dass Personen nicht, oder nicht nur in einer Gemeinde, sondern in mehreren Gemeinden

1 Statistische Ämter des Bundes und der Länder, Ergebnisse des Zensustests, Wirtschaft und Statistik, Heft 8/2004, S. 813 - 833

2 Die geringere Zahl der Meldebehörden kommt dadurch zustande, dass Gemeinden sich zu Verwaltungsgemeinschaften, Verbandsgemeinden u.ä. zusammengeschlossen haben, die u.a. die Melderegister gemeinsam führen. Andererseits gibt es in größeren Städten oftmals mehr als eine Meldebehörde.

gleichzeitig mit alleiniger Wohnung oder mit Hauptwohnung gemeldet, oder ausschließlich mit Nebenwohnung registriert sind. Solche Fehler können durch eine nicht zeitgleich stattfindende An- und Abmeldung sowie ihre (verzögerte) verwaltungsmäßige Bearbeitung oder durch unterlassene Abmeldungen usw. entstehen.

Hinzu kommt, dass das Melderecht den mehrfachen Eintrag des Wohnsitzes zulässt (Hauptwohnung/Nebenwohnung).

Bei einer Nutzung der Meldedaten zu Zensuszwecken ohne weitere Prüfung der Angaben durch die statistischen Ämter besteht daher die Gefahr, dass Personen nicht oder mehrfach, am falschen Ort oder mit falschem Wohnstatus gezählt und dadurch unzutreffende Einwohnerzahlen festgestellt werden.

Um Doppelzählungen bei einem registergestützten Zensus zu vermeiden, muss der Registerbestand auf Duplikate untersucht und gegebenenfalls bereinigt werden. Die Dublettensuche kann man datentechnisch als eine Zusammenführung eines Datenbestandes mit sich selbst betrachten, mit der Zielsetzung, die Datensätze zu kennzeichnen, die zur gleichen Person gehören, sich aber in der Wohnanschrift unterscheiden. Die folgende Abbildung veranschaulicht die Vorgehensweise.

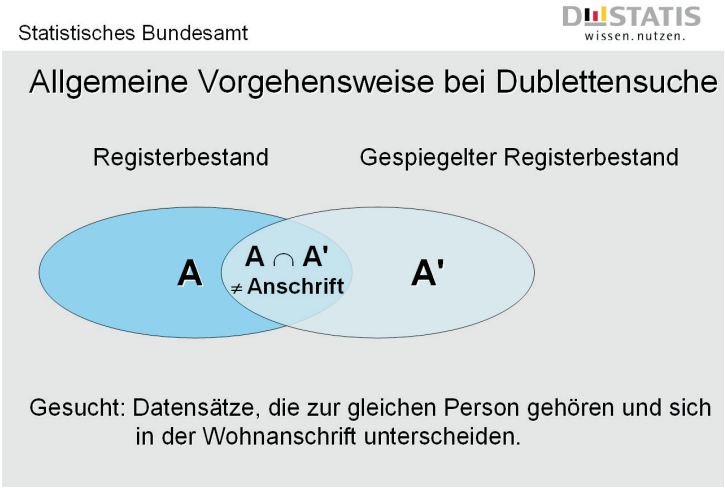


Abb. 2: Allgemeine Vorgehensweise bei Dublettensuche

Grundannahme bei der Suche von Dubletten in Datenbeständen von Personen ist, dass es Merkmale in den Datensätzen gibt, anhand derer Daten eindeutig der gleichen Person zugewiesen werden können.

Die Melderegister enthalten kein bundeseinheitlich durchgängiges und zeitlich konstantes Ordnungsmerkmal. Aus dem Datenkranz der Melderegister[3] waren also solche Merkmale zu bestimmen, die sich für Personen im Zeitablauf in der Regel nicht ändern und die miteinander kombiniert Schlüsselcharakter erhalten. Es wurde im Vorfeld der Mehrfachfallprüfung davon ausgegangen, dass Datensätze sich dann auf die gleiche Person beziehen, wenn die Datensätze in den Schlüsselmerkmalen Geschlecht, Geburtsjahr, Geburtsmonat, Geburtstag, Geburtsort, Geburtsname und Vorname übereinstimmten. Wenn diese Übereinstimmung bspw. in zwei Datensätzen vollständig gegeben ist, ist die Wahrscheinlichkeit, dass diese Datensätze zu der gleichen Person gehören und diese Person doppelt gemeldet ist, sehr hoch.

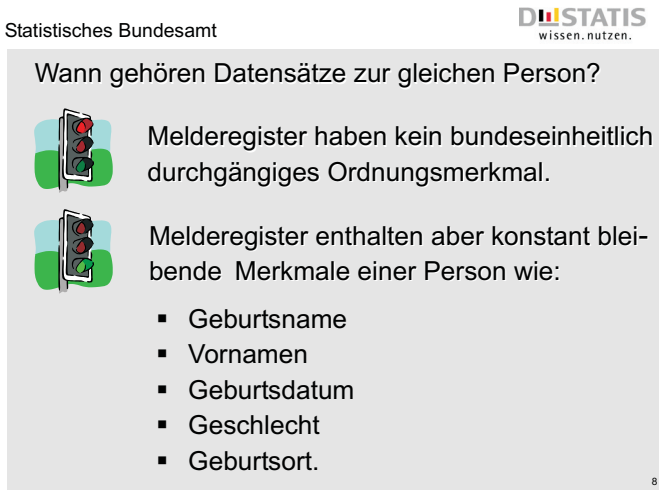


Abb. 3: Wann gehören Datensätze zur gleichen Person?

3 Aus den Melderegistern wurden für die o.g. Einwohner gemäß § 2 Zensustestgesetz folgende Merkmale erhoben: 1. als Erhebungsmerkmale: Geburtsmonat und –jahr; Geschlecht; Staatsangehörigkeiten; bei im Ausland Geborenen: Geburtsstaat; Familienstand; Wohnort; Status der Wohnung (alleinige Wohnung, Haupt- oder Nebenwohnung); 2. als Hilfsmerkmale: Namen; Vornamen; gegenwärtige Anschriften; Tag der Geburt; Geburtsort; Standesamt und Nummer des Geburtseintrags; Anschrift und Status der künftigen Wohnung oder der Wohnung, in die der Einwohner laut Rückmeldung verzogen ist; Anschrift und Status der Wohnung in der Gemeinde, aus der der Einwohner zugezogen ist; Zuzug aus dem Ausland; Anschrift der zuletzt bewohnten Wohnung in der Gemeinde; Datum des Beziehens der Wohnung; Datum des Auszugs aus der Wohnung; Datum des Fortzugs ins Ausland; Datum der Anmeldung bei der Meldebehörde; Datum der Abmeldung bei der Meldebehörde; Datum des Wohnungsstatuswechsels.

Die Eingabefelder und die Feldformate der in den Melderegistern gespeicherten Daten sind bundeseinheitlich normiert über den Datensatz für das Meldewesen (DSMeld).[4]

Obwohl die oben genannten Merkmale im Zeitablauf konstant bleiben, sind sie nicht in allen Melderegistern gleichlautend gespeichert. Die dezentrale Führung der Melderegister mit voneinander getrennten An- und Abmeldevorgängen sowie die auf viele Mitarbeiter verteilte Bearbeitung von Meldevorgängen beeinträchtigen die Konstanz der Merkmale etwa durch Fehler bei der Datenaufnahme, unterschiedliche Schreibweisen – auch bei der Übertragung von fremdsprachlichen Namen ins Deutsche - oder Verwendung von Abkürzungen bspw. bei Geburtsorten.

In Abbildung 4 werden Gründe für die Beeinträchtigung der Konstanz der Kernmerkmale für die Dublettensuche aufgelistet.

Abb. 4: Gründe für die Beeinträchtigung der Konstanz von Merkmalen

Marginale Unterschiede in der Schreibweise von Namen (z.B. „oe“ anstatt „ö“) können darauf zurückzuführen sein, dass veraltete Software verwendet wird, die keine Sonderzeichen wiedergeben konnte. Andererseits kann die Schreibweise „oe“ aber korrekt sein.

Die Anwendung unterschiedlicher Softwareversionen in den Meldebehörden mit unterschiedlichen Zeichensätzen in einem Datenbestand, die Markierung

4 S. Bundesvereinigung der kommunalen Spitzenverbände, Datensatz für das Meldewesen. Einheitlicher Bundes-/Länderteil (DSMeld), Köln 2004

von Merkmalen aus programmtechnischen Gründen[5], aber auch das „papierlose“ Meldeamt[6] sind ursächlich dafür, dass auch Eingabefelder, die in amtliche Papiere eingedruckt werden, nicht immer der DSMeld-Norm entsprechend gefüllt sind. Hinzu kommen Unterschiede in der Schreibweise, die bei Systemwechsel der Software durch unzulängliche Adaptionen des Datenbestandes in die neue Software verursacht wurden.

Bei dem Merkmal „Geschlecht“ ist das Transsexuellengesetz zu berücksichtigen, dessen personenstandsrechtliche Schutzbestimmungen natürlich einzuhalten sind.

Als Geburtsort sollte in den Melderegistern der Ortsname vermerkt sein, der sich in der Geburtsurkunde bzw. in anderen amtlichen Dokumenten findet. Häufig sind damit in den Melderegistern durch Verwaltungsreformen überholte Ortsnamen aufgeführt, aber auch solche in ihrer neuen Form. Im Datenbestand der Testerhebung ist beispielsweise der Geburtsort „Frankfurt am Main“ in 58 Variationen enthalten.

Obwohl die Qualität der Dateninhalte bei den aufgeführten Merkmalen sehr hoch ist - auch weil der Bürger sie gegenprüft, wenn er bspw. seinen Personalausweis erhält -, folgt aus der Beeinträchtigung der Konstanz der Merkmale, dass eine byteweise Prüfung auf Übereinstimmung der Sachlage nicht angemessen ist und Datensätze auch dann noch der gleichen Person zugeschrieben werden können, wenn die Überstimmung in den Schlüsselmerkmalen nicht 100 Prozent beträgt, die Ausprägungen der Merkmale sich also in hohem Maße ähnlich sind.

Die Dublettenprüfung erfolgt entsprechend mit einem maschinellen Verfahren, bei dem Datensätze miteinander verglichen werden und bei Übereinstimmung als „gleich“ bzw. als „mit hoher Wahrscheinlichkeit gleich“ gesetzt werden.

Das Ergebnis des Verfahrens muss sich in der „Realen Welt“ beweisen, etwa indem die Personen, die mit mehreren Datensätzen im Registerbestand enthalten sind, nach ihren Meldeverhältnissen gefragt werden.

Abbildung 5 zeigt mögliche Fallkonstellationen als Ergebnis der Überprüfung der maschinellen Dublettenprüfung in der Empirie.

Wenn die Dublettenprüfung fehlerfrei verlaufen würde, würde sich der Datenbestand aufteilen in Dubletten und Unikate. Bei den Dubletten würde anhand des Anmeldedatums entschieden, an welchem Ort die Person gezählt und wo sie gelöscht wird. Ziel des Verfahrens ist, jeder Person - bezogen auf den Wohnsta-

5 Anstatt das Feld „Rufnamen“ zu füllen, wurde im Feld „Vornamen“ ein Namensteil z.B. mit „/“ markiert.

6 Manche Sachbearbeiter machen ihre Notizen in Eingabefelder, die normalerweise selten benutzt werden.

tus „Alleinige Wohnung/Hauptwohnung“ - einen und nur einen Datensatz zuzuordnen.

Es kann aber nicht davon ausgegangen werden, dass das Ergebnis der Dublettenprüfung fehlerfrei ist, weil bereits auszuschließen ist, dass die dem Verfahren zugrunde liegenden Datenfelder immer fehlerfrei sind.

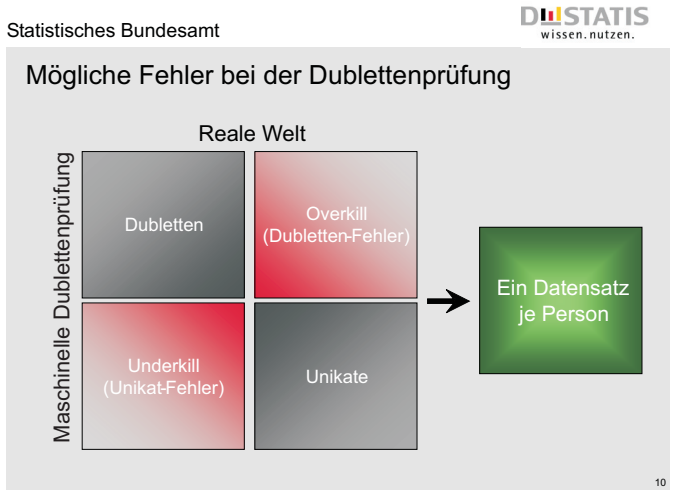


Abb. 5: Mögliche Fehler bei der Dublettenprüfung

Es sind zwei Arten von Fehlern möglich. Zum einen kann Personen die Dublettenеigenschaft fälschlicherweise zugewiesen werden („Overkill“). Oder es wird nicht erkannt, dass Personen mehr als einmal im Datenbestand vorhanden sind („Underkill“).

Die allgemein definierte Aufgabe der Dublettensuche ist, „Overkill“ und „Underkill“ gleichzeitig zu minimieren, wobei beim „Overkill“ zumindest noch die Chance gegeben ist, den Fehler zu entdecken, wenn - wie beim Zensustest durchgeführt - eine Befragung der „Dubletten“ erfolgt.

Es liegt auf der Hand, dass eine Dublettensuche nur möglich ist, wenn die dezentral gehaltenen Registerbestände in eine zentrale Datei gebracht werden. Erste Voraussetzung hierfür waren bundeseinheitliche Plausibilitätsprüfungen und die Überführung in ein einheitliches Satz- und Zeichenformat.

Vor dem Zensustest hatte die amtliche Statistik nahezu keine Erfahrungen mit Verfahren der Dublettensuche in großen Datenbeständen. Es wurde daher „externes Wissen eingekauft“ und entsprechend ausgewiesene Softwarehäuser beauftragt, den Datenbestand auf Dubletten zu prüfen. Es wurden sechs Softwarevarianten, die nach unterschiedlichen Methoden arbeiten, getestet. Um auszu-

schließen, dass das Ergebnis der Dublettensuche durch inadäquate Parametervorgaben beeinträchtigt wird, erfolgte die Dublettensuche jeweils in den Softwarehäusern mit dem dort vorhandenen geschulten Personal.

Parallel dazu wurden im Statistischen Bundesamt Konzepte erarbeitet, deren Umsetzung Überprüfungen der Ergebnisse der externen Dublettensuche ermöglichten. Neben maschinellen Prüfungen erfolgte eine Befragung der Personen, deren Datensätze als doppelt oder mehrfach vorhanden gekennzeichnet worden waren, zu ihrem Wohnsitz am Stichtag.

Ein besonderes Anliegen war, den Bürger nicht mit Programmfehlern zu konfrontieren und ihm fälschlicherweise mitzuteilen, er sei mit mehr als einer Hauptwohnung gemeldet. Dies hatte zur Konsequenz, dass die Ergebnisse der externen Softwareprogramme vor Durchführung der Befragung überprüft werden mussten.

Das Ergebnis der Dublettensuche mit Hilfe der sechs Softwarevarianten und das Ergebnis der Evaluation der Dublettensuche zeigt Abbildung 6. Die Evaluation der Dublettensuche über maschinelle Verfahren und postalische/telefonische Befragungen hatte etwas mehr als 39.000 „echte“ Dubletten (vor Hochrechnung) zum Ergebnis.

Statistisches Bundesamt

DESTATIS wissen.nutzen.

Ergebnis der Dublettensuche der Softwarevarianten

- vor Hochrechnung -

Gefundene Dubletten	Softwarevariante A 1	A 2	A 3	A 4	A 5	A 6
Übereinstimmend	25.961	25.961	25.961	25.961	25.961	25.961
Richtig	31.784	32.509	32.194	31.720	34.665	33.629
Falsch	2.368	2.837	2.389	10.540	6.869	2.518
Nicht	7.551	6.826	7.141	7.615	4.670	5.706
Summe Fehler	9.919	9.663	9.530	18.155	11.539	8.224

12

Abb. 6: Ergebnis der Dublettensuche der Softwarevarianten

Im Abbildung 6 sind die von den Softwarevarianten übereinstimmend gefundenen Dubletten wiedergegeben sowie die richtig bzw. falsch klassifizierten Du-

bletten. Die Zahl der nicht gefundenen Dubletten und der Gesamtfehler ist ebenfalls ausgewiesen.

Übereinstimmend gefunden wurden 25.961 Dubletten. Um die relative Leistungsfähigkeit der Programme einschätzen zu können, ist die dritte und vierte Zeile aufschlussreich. Besonders schwer wiegen die nicht gefundenen Dubletten (Underkill-Fehler), weil diese im Bestand verbleiben würden.

Die Variante A6 weist den niedrigsten Gesamtfehler aus. Die Variante A4 hat den höchsten Gesamtfehler, insbesondere weil in dieser Variante bei ausländischen Namen sehr intensiv zusammengeführt wurde. Die Variante A5 kommt der Zahl der echten Dubletten mit 34.600 am nächsten, weil sie die Programmparameter relativ weich eingestellt hat, also einen grobkörnigen Filter nutzt.

Als Zwischenergebnis ist festzuhalten:

- Keine Software bietet ein hinreichend genaues Ergebnis.
- Beim besten Ergebnis liegt der Fehleranteil noch bei 21,5 %.
- Bei Anwendung derart eingestellter Software beim Zensus würden entweder sehr viele Bürger fälschlicherweise als Dublette angeschrieben oder die angezeigten Dubletten müssten aufwändig in einem maschinell gesteuerten Dialogverfahren geprüft werden.

Als Konsequenz dieses Ergebnisses des Zensustests wurde der Schluss gezogen, dass die amtliche Statistik für den Zensus eigenständig leistungsfähige Suchroutinen entwickelt, die bei einem flächendeckenden Zensus praktikabel sind. Begründet ist diese Vorgehensweise auch in dem Sachverhalt, dass die sehr intensive Beschäftigung mit dem Datenmaterial erkennen ließ, warum manche Sätze als „Doppel“ ausgegeben wurden. Diese Anhaltspunkte gilt es dann zu systematisieren und programmtechnisch umzusetzen.

3 Methode der Dublettenprüfung beim Zensus

An die programmtechnische Umsetzung der Dublettenprüfung beim Zensus sind hohe Anforderungen gestellt. Die Dublettensuche erfolgt in einem Datenbestand von etwa 88 Millionen Datensätzen. Sie muss zeitnah zum Stichtag abgeschlossen sein. Die Datenfelder müssen zur Vorbereitung der Dublettenprüfung in besonderer Weise standardisiert und erweitert werden. Es sind die Datensätze zu markieren, die mehrfach vorhanden sind (Personen mit Haupt- und Nebenwohnungen, Personen mit mehreren Hauptwohnsitzen) bzw. die Datensätze von Personen, die ausschließlich mit Nebenwohnung gemeldet sind, was nach deutschem Melderecht nicht zulässig ist.

Zunächst gilt es, die primären Merkmale zu bestimmen, anhand derer der Datenbestand nach Dubletten durchsucht wird. Die primären Merkmale, die be-

reits im Zensustest herangezogen wurden, haben sich bewährt, so dass die Merkmale Geburtsname, Vornamen, Geschlecht, Geburtsdatum und Geburtsort diese Funktion übernehmen.

Aus Gründen der Laufzeitoptimierung und weil es angezeigt ist, die Dublettensuche als Ähnlichkeitssuche zu organisieren, muss die Menge der Daten, in der nach mehrfach vorhandenen Sätzen gesucht wird, eingeschränkt werden. Zu diesem Zweck wird der „Daten-Raum", in dem nach Dubletten gesucht wird (Such-Raum), nicht auf vorgefundene Schreibweisen von Namen beschränkt. Statt dessen wird der Such-Raum erweitert, indem bspw. die Merkmale „Geburtsname" und „Vorname" phonetisch umgewandelt[7] und standardisiert werden. Datensätze bspw. der Personen „Meyer", „Maier", „Meier" bilden so einen gemeinsamen Such-Raum. Es entsteht der „standardisierte Datenbestand".

Unter Beachtung des Prinzips „Teilen, Suchen, Finden" wird der „standardisierte Datenbestand" so geteilt, das eine Optimierung der Suchabläufe erreichbar ist. Es entsteht der in Abbildung 7 gezeigte Suchbaum.

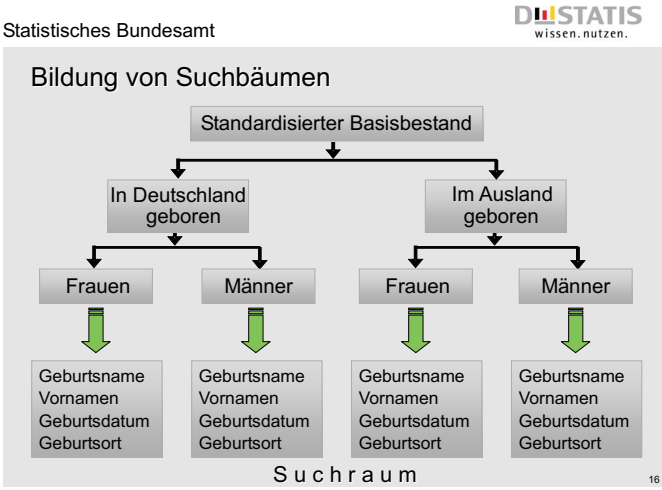


Abb. 7: Bildung von Suchbäumen

Das erste Splitting des „standardisierten Basisbestandes" ist in zweifacher Hinsicht von Vorteil: die Datenmenge wird reduziert und die komplexere Suche bei ausländischen Namen wird auf eine Teilmenge reduziert, die etwa 10 % des Ge-

7 Die gängigen Verfahren wie „Soundex" oder „Kölner Verfahren" werden in Richtung einer „phonetic-light-Variante" modifiziert, um die ansonsten gegebene große Zahl von unsinnigen Vergleichen einzuschränken.

samtbestandes ausmacht.[8] Mit dem zweiten Splitting nach Männern und Frauen halbieren wir den jeweils zu durchsuchenden Bestand.

Wie mit den verbleibenden vier Merkmalen „Geburtsname“, „Vornamen“, „Geburtsdatum“, „Geburtsort“ weiter verfahren wird, ist in Abbildung 8 dargestellt.

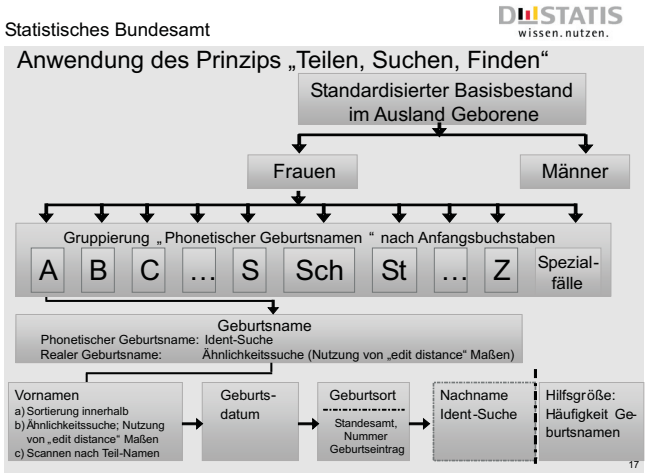


Abb. 8: Anwendung des Prinzips „Teilen, Suchen, Finden“

Bei der zum Einsatz kommenden „phonetischen“ Variante wird der Anfangsbuchstabe der Namen beibehalten, so dass die Datenmenge anhand des Anfangsbuchstabens des „Phonetischen Geburtsnamens“ gruppiert werden kann. Innerhalb der Gruppe der Personen mit gleichem Anfangsbuchstaben des „Phonetischen Geburtsnamens“ findet eine Ident-Suche (byteweiser Vergleich) statt. Bei Sätzen mit phonetisch-gleichem Geburtsnamen erfolgt eine Ähnlichkeitssuche innerhalb der „realen Geburtsnamen“.

In der Menge der „ähnlichen realen Geburtsnamen“ werden dann die Merkmale Vornamen, Geburtsdatum, Geburtsort aufeinander folgend in den Suchprozess einbezogen. Bei Personen, die mit der Eheschließung den Namen gewechselt haben, kann dieser zusätzlich in die Prüfung gehen. Die Häufigkeit des Vorkommens von Geburtsnamen kann ebenso als Entscheidungshilfe mit herangezogen werden.

Abbildung 8 weist zusätzlich zu der Gruppierung der phonetischen Geburtsnamen nach Anfangsbuchstaben auf Spezialfälle hin, die gesondert bearbeitet

8 Diese Erleichterung des Suchprozesses hatten wir auch den externen Software-Programmen zugestanden.

werden müssen. Zum einen sind es Namen aus dem asiatischen Raum, die in Meldeämtern nicht immer in der richtigen Weise auf Vor- und Geburtsnamen aufgeteilt werden. Zum anderen sind es Namen, die aus dem Russischen ins Deutsche (zurück)übertragen wurden. Da sich bei diesen Übertragungen (z.B. wird aus Ganiman Hahnemann oder aus Vajmer Weimer) auch die Anfangsbuchstaben ändern können, wird das Verfahren der Dublettensuche an dieser Stelle zu variieren sein. Dank des oben erwähnten „papierlosen" Meldeamtes konnten Erfahrungen gesammelt werden, wie Namen aus anderen Sprachräumen ins Deutsche transkribiert werden. Diese Informationen werden genutzt, um gezielte Suchen durchführen zu können.

Das Programm zur Ähnlichkeitssuche versucht, aus einem gegebenen Namen einen Zielnamen zu erstellen. Das Programm zur Ähnlichkeitssuche verwendet eine generalisierte Form der Berechnung der „Levenshtein Distance". Das vereinfachte Prinzip der Vorgehensweise bei der Bestimmung der „Levenshtein Distance" erklärt sich aus dem schrittweisen Aufbau der Diagonalwerte der Matrix in Abbildung 9.

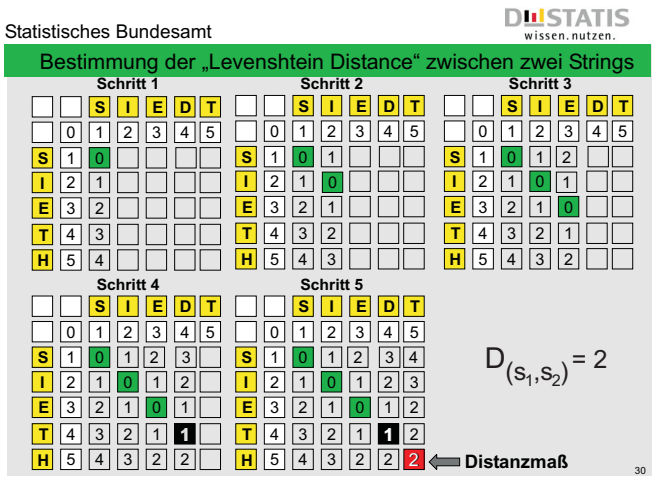


Abb. 9: Bestimmung der „Levenshtein Distance" zwischen zwei Strings

Gemessen wird die Anzahl und Art der Manipulationen, um einen Namen in einen anderen zu überführen. Bei Buchstaben, die ohne „Manipulation" überführt werden können, erhalten die Diagonalelemente den Wert „0". Anderenfalls den Wert „1", der sich zum Gesamtpunktwert addiert.

In dem komplexen angewandten Programm werden Manipulationen wie Einfügen, Löschen, Ersetzen, Drehen, Verdoppeln von Buchstaben unterschiedlich

gewichtet und ihre Gewichte addiert (Kosten der Manipulationen). Die Kosten der Manipulationen werden zur Stringlänge ins Verhältnis gesetzt und auf das Intervall zwischen 0 und 1 normiert. Für das Distanzmaß gilt:

$$D_{(s_1,s_2)}1 = \frac{\sum Kosten\ der\ Manipulation}{Stringlänge}$$

Bei identischen Geburtsnamen erhält das Distanzmaß den Höchstwert „1“. Ansonsten werden in Abhängigkeit von der Aktion Punkte abgezogen.

Als Ergebnis der Ähnlichkeitssuche entsteht eine Datei mit den Sätzen, die hohe Ähnlichkeitswerte in den primären Merkmalen haben (s. Abbildung 10).

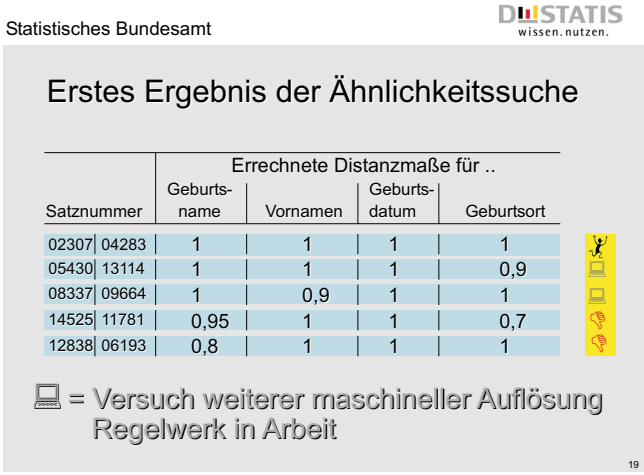


Abb. 10: Erstes Ergebnis der Ähnlichkeitssuche

Das Regelwerk, wie bei bestimmten Distanzmaß-Konstellationen zu verfahren ist, muss noch entwickelt werden. Ziel ist die Zahl der Fälle, die maschinell nicht lösbar sind, möglichst klein zu halten.

Mit einem Verfahren, das weniger weit entwickelt ist, als das soeben vorgestellte, konnten im Zensustest 95 % der Fälle, die in die Befragung gegangen sind, maschinell aufgelöst werden. In Bundesländern mit einem geringen Ausländeranteil wurden 98 % und mehr maschinell aufgelöst.

Maschinell nicht auflösbare Fälle werden zunächst einer visuellen Kontrolle unterzogen, in deren Verlauf entschieden wird, ob diese Fälle im Rahmen einer Befragung geklärt werden müssen.

Im Zensustest ergab die Dublettenprüfung folgendes Bild:

In den Melderegistern waren zum 5.12.2001 82.031.000 Personen mit alleiniger Wohnung bzw. mit Hauptwohnung gemeldet (Bevölkerung am Ort der Hauptwohnung). In diesem Registerbestand waren hochgerechnet 856.000 Dubletten enthalten, wobei rund 370.000 Dubletten aufgrund von Wohnungswechseln um den Stichtag, d.h. umzugsbedingt zu erklären sind.[9] 378.000 Dubletten wurden in der ersten Datenlieferung entdeckt,[10] 108.000 Dubletten wurden in der stichtagsgenauen Zusammenführung der beiden Datenlieferungen gefunden.

Damit beträgt die von den Dubletten statistisch bereinigte Bestandszahl der Melderegister zum 5.12.2001 81.175.000 Personen (Bevölkerung am Ort der Hauptwohnung). Die entsprechende Bevölkerungszahl gemäß Bevölkerungsfortschreibung beträgt zum 31.12.2001 82.440.000. Die zu erwartende Korrektur der Bevölkerungszahl durch den registergestützten Zensus liegt damit im Saldo bei mindestens - 1,3 Mill. Personen. Bei der Volkszählung 1987 wurde die Bevölkerungszahl im Saldo um - 77 000 Personen korrigiert. Die Anwendung der neuen Einwohnerzahlen bei Umsatzsteuerverteilung und Finanzausgleich führte seinerzeit zu Umverteilungen zwischen den Ländern in Höhe von 455 Mill. DM.[11]

4 Qualitätsbericht zu den Melderegistern

Der zur Vervollständigung des Bildes gegebene Qualitätsbericht zu den Melderegistern umfasst die Definition der Merkmale und Dateiformate, den Erfassungsgrad der Melderegister, ihre Update-Mechanismen sowie die Möglichkeit, Bezugszeiträume mit den Daten der Melderegister exakt abzubilden.

Zur Bewertung der Ergebnisse des Zensustests im Hinblick auf Überlegungen zum registergestützten Zensus 2010/2011 ist es förderlich, sich in Erinnerung zu rufen, welche Maßnahmen im Meldewesen in Umsetzung begriffen sind, die Einfluss auf die Qualität der Registerdaten haben könnten. Neben der Optimierung der Verfahrensabläufe im Meldewesen[12] und der konsequenten

9 Um auch die Einwohner am richtigen Wohnort zählen zu können, die sich nach dem Stichtagsdatum 5. Dezember 2001 bei den Meldebehörden rückwirkend an- oder abmelden, war es notwendig, für den ausgewählten Personenkreis einen weiteren Melderegisterauszug anzufordern, der zum Stichtag 31. März 2002 erstellt wurde.

10 Einfacher relativer Standardfehler 2,09; das Ergebnis liegt mit einer Wahrscheinlichkeit von 68 % zwischen 370.000 und 386.000.

11 Quelle: Deutscher Bundestag, Bericht des Haushaltsausschusses, BT-Drs. 11/6621 vom 08.03.1990.

12 Von besonderer Bedeutung für den Zensus ist hierbei die Änderung des Melderechtsrahmengesetzes vom 4. April 2002. Die Novellierung verfolgte u.a. das Ziel, die Nutzung neuer Medien zuzulassen, um Geschäftsprozesse des Meldewesens effizienter, effektiver und attraktiver gestalten zu können.

Anwendung des Instrumentariums des § 4a MRRG sind hier vor allem die Realisierung der „Vernetzung“ der Melderegister sowie die Realisierung der ID-Nummer für Besteuerungsverfahren zu nennen.

In Zusammenhang mit der angestrebten „Vernetzung“ der Melderegister werden zur Zeit große Anstrengungen unternommen, den Datenaustausch zwischen den Meldebehörden und anderen Verwaltungsstellen auf ein zeitgemäßes Niveau zu heben. So wird das bisherige Datenaustauschformat DSMeld ersetzt durch das XML-Format XMeld.[13]

Abb. 11: Qualitätsbericht zu den Melderegistern - Definition der Merkmale, Dateiformate

Der Zensustest hat gezeigt, dass trotz der bundeseinheitlichen Vorgabe des DSMeld die Definitionen der Merkmale zum Teil in Abhängigkeit von der verwendeten Software variieren und Fehler enthalten, die weit in der Vergangenheit entstanden sind und nicht auffällig wurden, weil die Personen seit Jahrzehnten keine Notwendigkeit sahen, das Meldeamt aufzusuchen. Da bei der Umstellung auf das XML-Format bislang keine Konsolidierung des Datenbestandes erkennbar ist, steht zu befürchten, dass die Fehler zumindest bei „Altfällen“ lediglich in ein anderes Datenformat gebracht werden.

Die Aufgabe der Meldebehörden gem. § 1 MRRG ist die Registrierung der im Zuständigkeitsbereich wohnhaften Personen, um deren Identität und Woh-

13 S. hierzu OSCI-XMeld Projektteam, OSCI-XMeld Version 1.1 Spezifikation des bundeseinheitlichen Datenaustauschformates für die Übermittlung von Daten des Meldewesens, Bremen, Juli 2003.

nungen feststellen zu können. Trotz dieser hohen Selbstverpflichtung hat der Zensustest - zusätzlich zu den Ergebnissen der Dublettenprüfung - festgestellt, dass die Melderegister „Karteileichen“ in Höhe von 2,3 % und Fehlbestände in Höhe von 1,7 % enthalten. Die „Karteileichen“- und Fehlbestandsraten streuen in den Ländern und vor allem in den Gemeinden stark. Die amtliche Statistik hat daraus den Schluss gezogen, dass beim Zensus 2010/2011 die Registerauszählungen über primärstatistische Stichprobenerhebungen gegengeprüft werden müssen. Es ist allerdings zu erwarten, dass sich die Qualität der Melderegister infolge der bundesweiten Überprüfung bei Einführung der Steuer-Identifikationsnummer infolge der vorgesehenen postalischen Mitteilung der Nummer an alle Bürger verbessern wird.

Abb. 12: Qualitätsbericht zu den Melderegistern - Erfassungsgrad der Melderegister

Die bundesweite Überprüfung bei Einführung der Steuer-Identifikationsnummer gibt Hinweise auf die Vorteile, die die Melderegister als „lebende Register“ haben. Sie sind in vielfältige Verwaltungsprozesse eingebunden, die für Updates genutzt werden können (s. Abbildung 13).

Ein Nachteil der Melderegister, der aus der dezentralen Vorhaltung der Daten herrührt, nämlich der zum Teil große zeitliche Abstand zwischen Wirksamwerden der Anmeldung unter der neuen Adresse und Wirksamwerden der Abmeldung in der Fortzugsgemeinde wird über die ab 2007 vorgesehene elektronische Rückmeldung, die den Austausch von An- und Abmeldedaten synchronisiert, größtenteils behoben.

Trotz der größeren zeitlichen Übereinstimmung der Zu- und Abbuchungen in den Melderegistern der Gemeinden wird es auch künftig besondere Anstrengungen erfordern, die Meldedaten stichtagsgenau abzubilden, da der Meldevorgang i.d.R. vom Bürger ausgelöst werden muss.

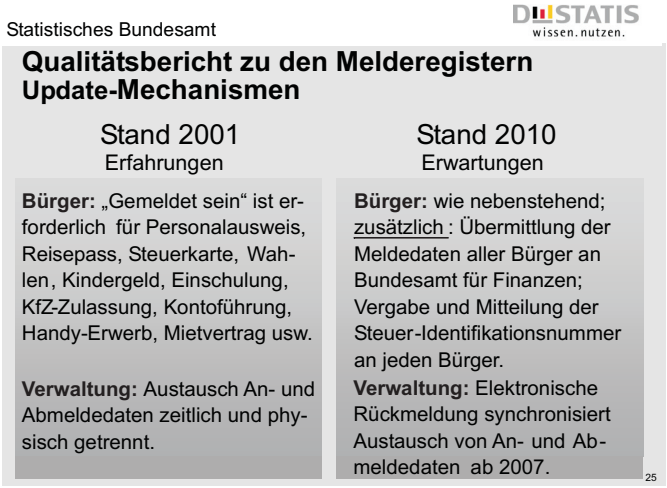


Abb. 13: Qualitätsbericht zu den Melderegistern - Updatemechanismen

Abb. 14: Qualitätsbericht zu den Melderegistern - Exaktheit der Bezugszeiträume

5 Schlussfolgerungen

Gegen den registergestützten Zensus, verstanden als reine Auszählung der Melderegister, wird oftmals eingewandt, er benachteilige die Gemeinden, die alle Anstrengungen unternehmen, um ihren Registerbestand von „Karteileichen“ zu bereinigen, etwa indem sie Abmeldungen zügig bearbeiten, den unzustellbaren Lohnsteuerkarten, Wahlbenachrichtigungen oder anderen Hinweisen aus der Verwaltung nachgehen. Andererseits würden die Gemeinden, die keine Maßnahmen zur Bereinigung ergreifen, beim Finanzausgleich belohnt, weil sie überhöhte Einwohnerzahlen haben. Mit der Dublettenprüfung werden zu derartigen „Karteileichen“ die aktuellen Gegenstücke in anderen Gemeinden gefunden und bei den Gemeinden, die sich einen Vorteil erhofft haben, nicht gezählt.

Ein weiterer Bonus der Dublettenprüfung ergibt sich aus den folgenden Überlegungen:

Beim Zensus 2010/2011 ist vorgesehen, die dann aktuelle „Karteileichen“- und Fehlbestandsrate der Melderegister über primärstatistische Stichproben zu ermitteln und die Registerzahlen entsprechend zu korrigieren.

Mit dem Instrumentarium der Mehrfachfallprüfung ist es möglich, zu überprüfen, ob die als „Karteileichen“ festgestellten Personen nicht unter anderen Anschriften im Melderegister neu angemeldet sind (nach Stichtag).

Der Sinn dieser Maßnahme wird einsichtig bei der Vorstellung, dass die vermeintliche „Karteileiche“ innerhalb einer Gemeinde umgezogen ist.

Wird diese Korrekturmöglichkeit nicht genutzt, wird der Gemeinde fälschlicherweise ein Einwohner abgezogen. Die Dimension dieses von den Gemeinden nachweisbaren Fehlers wird deutlich, wenn die Zahl der innerstädtischen Wanderungen betrachtet wird. In Hamburg ziehen bspw. pro Jahr etwa 110 000 Personen innerhalb der Stadt um, in Berlin sind es im Jahr 2001 etwa 450 000 Personen gewesen. In den Gemeinden mit 20 000 u.m. Einwohnern sind es über 2,1 Mill. Personen. Entsprechendes gilt für die Wanderungen innerhalb der Bundesländer und - bezogen auf die Feststellung der amtlichen Einwohnerzahl für Deutschland sowie internationale Vergleiche - für Wanderungen innerhalb des Bundesgebietes.

Es ist also festzustellen:

> Die Dublettenprüfung ist aktives Kontrollinstrument des registergestützten Zensus 2010/2011. Sie korrigiert die beim Registerzensus gegebene systematische Benachteiligung der Gemeinden, die auf die Qualität ihrer Register besonders bedacht sind und eröffnet darüber hinaus Möglichkeiten, primärstatistische Korrekturverfahren raumbezogen auszurichten.

Die Vorbereitungen des Zensus 2010/2011 sind in vollem Gange. Es zeichnet sich bereits jetzt ab, dass der Zensus methodisch und datentechnisch sehr kom-

plex werden wird, um ein Vielfaches komplexer als die 87er Zählung. Dies ist der Preis dafür, dass der registergestützte Zensus etwa ein Drittel von dem kosten wird, was eine Zählung nach dem 87er Konzept kosten würde (etwa 340 Mill. zu 1,1 Mrd.).

Data Matching: Integration von Umfrageergebnissen und Unternehmensdaten

Stefan Tuschl

Für einen langfristigen Erfolg am Markt ist es für ein Unternehmen essentiell, die wichtigen Kunden (hoher Kundenwert) von den weniger wichtigen Kunden zu differenzieren. Langfristige Beziehungen zu den wichtigen Kunden aufzubauen und zu unterhalten, ist das Schlüsselelement für eine erfolgreiche Unternehmensstrategie. Wenn nun Unternehmen - um dieses Ziel zu erreichen - (zusätzliche) Informationen über Ihre Kunden gewinnen wollen, können Sie zu Analysezwecken grundsätzlich auf zwei Arten von Datenquellen zurückgreifen: auf die eigenen Kundendaten (üblicherweise abgelegt in einem Data Warehouse oder in Datenbanken) oder auf Marktforschungsergebnisse, also Daten, die aus anonymen Kundenbefragungen stammen. Beide Datenquellen werden in den Unternehmen allerdings primär „per se", also getrennt voneinander, verwendet.

Mit CRM fing alles an …

Mit dem Entstehen von CRM, das auf den systematischen Aufbau und die Pflege dauerhafter und profitabler Kundenbeziehungen zielt, rückten die in den Unternehmensdatenbanken verfügbaren Kundendaten in den Mittelpunkt des Interesses. Durch die schnell fortschreitenden Weiterentwicklungen im IT-Bereich ist es einem Unternehmen mittlerweile möglich, sämtliche Daten von Kunden und alle Transaktionen mit diesen Kunden in Datenbanken zu sammeln, und zwar preiswert, schnell und automatisch. Data Warehouses, Data Marts und Marketing-Datenbanken wurden eingerichtet, mit dem Ziel, die intern verfügbaren Daten zu konsolidieren und durch deren Verknüpfung neue Informationen und Einblicke zu gewinnen. Dabei werden diese Daten idealerweise so integriert und aufbereitet, dass sie an jeder Stelle im Unternehmen in der jeweils passenden Zusammenstellung zur Verfügung stehen. Zusätzlich werden Techniken des Data Mining zur Analyse der Daten eingesetzt, um ein tieferes und weiterführendes Kundenverständnis zu bekommen. Bei diesem analytischen CRM kommt es darauf an, möglichst viel Relevantes aus den in den Kundendaten erhaltenen Informationen zu gewinnen und so Eigenschaften, Verhaltensweisen und Wertschöpfungspotenziale von Kunden besser erkennen und einschätzen zu können.

Data Mining-Techniken sind das Herz jedes Database Marketing-Systems, ermöglichen sie es doch, den Kunden in jedem Stadium seines Lebenszyklus besser zu verstehen, ihn dadurch gezielter mit Angeboten anzusprechen und ihn damit schließlich auch für das Unternehmen profitabler zu machen. Die zentralen Fragen, mit denen sich das Database Marketing befassen muss, sind:

Schlüsselfragen im Database Marketing

Der Schlüssel zum Data Matching: Interne Daten + … !

Data Mining wird meistens ausschließlich in Bezug auf die Auswertung von intern im Unternehmen vorliegenden Datenquellen verstanden und durchgeführt. Diese Daten umfassen in aller Regel Bestandsdaten von Kunden sowie Transaktionsdaten über die Produktnutzung dieser Kunden bei dem betreffenden Unternehmen. D.h., es sind sämtliche datentechnisch erfassten Geschäftsvorfälle der Kunden mit dem eigenen Unternehmen verfügbar, also in erster Linie Bestell-, Lieferungs- und Abrechnungsdaten. Diese Datenbestände dokumentieren, welche Kunden wann welche Produkte des eigenen Unternehmens genutzt haben, allerdings bleiben dabei zwei wichtige Informationsdimensionen unberücksichtigt:

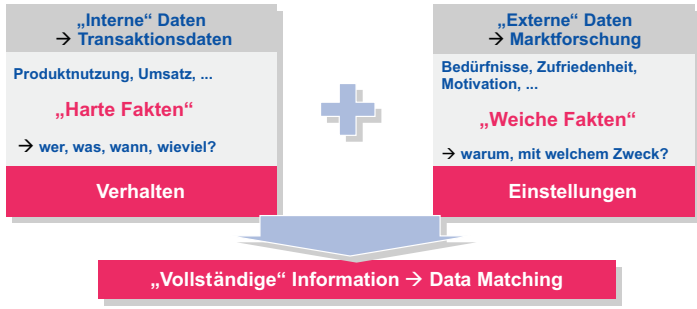
a) Die „Weichen Faktoren", die im Zusammenhang mit der jeweiligen Transaktion stehen, wie z.B.: Kaufmotivation für ein bestimmtes Produkt, Verwendungszweck, Zufriedenheit mit einem Produkt bzw. mit einem Unternehmen.

b) Informationen über die Aktivität der eigenen Kunden bei der Konkurrenz: Der „Blick auf die Kunden" ist auf die Vorgänge im eigenen Haus begrenzt. Auf dieser Informationsbasis sind jeweils nur Analysen des Kundenverhaltens aus interner Sicht und begrenzt auf interne Vorgänge möglich.

Für ein effizientes CRM sind also mehr Informationen nötig, als eine interne Datenbank liefern kann, die sich normalerweise auf die vier W-Fragen Wer? Was? Wann? Wie viel? beschränkt. Diese Perspektive können nun Daten, die aus Marktforschungserhebungen stammen, besonders gut ergänzen und die bestehende Lücke „über den Tellerrand hinaus" hin zu Motiven, Einstellungen, Bedürfnissen und auch zum Gesamtmarktüberblick schließen:

- Marktforschungsuntersuchungen sind häufig auf die Analyse der Ursachen und Zusammenhänge fokussiert, welche zu den getätigten Geschäften/Vorgängen führten. Das heißt, sie liefern die fehlenden Erklärungen und Zusammenhänge hinter den „harten Fakten" der internen Kundendaten.
- Marktforschungsstudien können auch so konzipiert werden, dass die gesamte Marktstruktur inklusive Konkurrenzsituation abgebildet und nicht lediglich die Produktnutzung bei einem Unternehmen analysiert wird.

CRM und Marktforschung



Mit der Erkenntnis, dass interne Firmendaten und Marktforschungsdaten unterschiedliche Informations-Dimensionen abdecken, entstehen im Database Marketing neue Zielsetzungen. Im Zuge eines so genannten Database Enrichments können z.B. einstellungsbezogene Daten (z.B. Zufriedenheiten, Interesse) oder auch Marktstrukturdaten (z.B. Geschäftsbeziehungen zur Konkurrenz) die bestehende Datenbasis des Kunden „bereichern". Für ein derartiges Database Enrichment werden so genannte Data Matching-Modelle verwendet, die zu den „neueren" multivariaten Verfahren zu zählen sind. Ein erfolgreiches „Matching" von internen Datenquellen und Marktforschungsdaten ermöglicht zusätzliche und neue Erkenntnisse über die Kunden eines Unternehmens; die durch das Database Enrichment zusammengeführten verschiedenen Informationsdimensionen fördern somit eine „ganzheitliche" (holistische) Sicht von Marketing und Vertrieb auf die Kunden.

Wie funktioniert Data Matching?

Bei der kombinierten Nutzung von Marktforschungsdaten und internen Datenquellen muss berücksichtigt werden, dass interne Kundendaten personenbezogen vorliegen, Marktforschungsdaten aufgrund datenschutzrechtlicher Bestimmungen aber anonym erhoben werden. Daraus folgt, dass man keine direkten Verbindungen auf Personenebene zwischen den beiden „Dateninseln" herstellen kann, dass also individuelle Umfragedaten nicht einfach an die jeweiligen persönlichen Einträge der Befragten in der Kundendatenbank angefügt werden können:

- Primär wäre dies eine klare Verletzung von Datenschutz- und Standesrichtlinien und ist damit faktisch ausgeschlossen.
- Selbst wenn diese Vorgehensweise zulässig wäre, würde dies dem Unternehmen dennoch nur eingeschränkt nützlich sein, da i.d.R. nur Stichproben von Kunden befragt werden, und die Zusatzinformationen dann nur für einen Bruchteil aller Kunden in der Datenbank bekannt wäre.

Es müssen also andere Verfahren und Möglichkeiten gefunden werden, die Datenquellen „zusammenzubringen", die nicht auf Personenebene stattfinden. Das Grundprinzip des Data Matching beruht nun darauf, mit Hilfe von speziellen Verfahren des Data Mining und der multivariaten Statistik Regeln und Muster in den Befragungsdaten zu identifizieren und diese auf die Kundendatenbank anzuwenden. Man versucht dabei zunächst, die anonymen Marktforschungsergebnisse für möglichst homogene Untergruppen zu modellieren. Diese Untergruppen werden im nächsten Schritt in der Kundendatenbank identifiziert bzw. durch Anwenden eines gefundenen Regelwerkes nachgebildet. Im Analogieschluss werden dann den Personen der entsprechenden Gruppen in der Datenbank die durch Marktforschung erhobenen Eigenschaften, ergänzt um den Wahrscheinlichkeitswert dieser Ausprägung, zugewiesen.

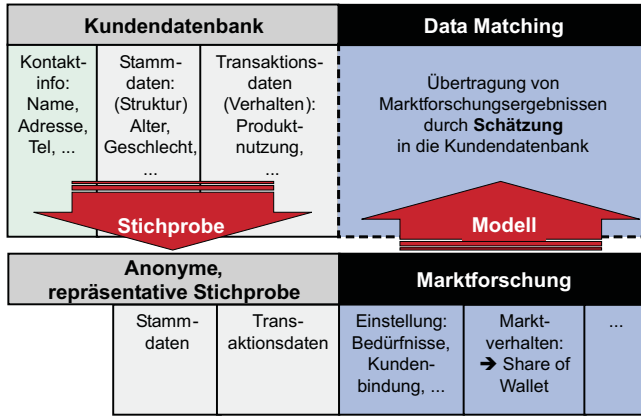
Die einzelnen Schritte eines Data Matching-Prozesses

Das Data Matching von Kundendaten und Umfrageergebnissen (Database Enrichment) lässt sich grob in die folgenden Schritte unterteilen:

- Die Marktforschung erhebt Informationen über Kundenbedürfnisse/-einstellungen (z.B. über Finanzverhalten, Produktaffinität) auf Basis einer Stichprobe aus der Kundendatenbank.
- Über Link-Variablen, die aus der Kundendatenbank stammen, wird in den Stichprobendaten dann z.B. die Produktaffinität mit Hilfe mathematischer Modelle erklärt.
- Dieses Erklärungsmodell wird auf die Kundendatenbank angewendet.

- Für jeden Kunden in der Kundendatenbank kann somit eine Wahrscheinlichkeit angegeben werden, mit der er/sie z.B. eine bestimmte Produktaffinität aufweist (z.B. „stark an Produkt X Interessierte“).
- Diese Informationen können dann für gezielte Marketingmaßnahmen (Produkt-, Preis- und Kommunikationspolitik) genutzt werden.

Database Enrichment:
Der Data Matching Prozess im Überblick



Beispiel 1: In Befragungsdaten wird festgestellt, dass bei leitenden Angestellten in der Altersklasse 30 bis 50 mit mindestens 2 Kindern ein deutlich überproportionaler Anteil zum lukrativen Typ („… wertvolle Häufigreisende“) gehört. 63% dieser Kundengruppe befinden sich im Zielsegment gegenüber lediglich 23% bei der Gesamtheit aller Kunden. Vorausgesetzt den Fall, dass es sich hier um ein robustes und signifikantes Ergebnis handelt (was durch die vorliegenden Fallzahlen und über ein „Hold-Out Sample“ nachgewiesen wurde), kann diese Erkenntnis auf die Kundendatenbank übertragen werden und im Analogieschluss im Sinne einer Wahrscheinlichkeit interpretiert werden: Wenn in der definierten Untergruppe der Stichprobe 63% der Befragten zum lukrativen Typ gehören (und wenn die Stichprobe repräsentativ für die Kundendatenbank ist), dann weisen die Personen in der Kundendatenbank, auf die diese Definition („leitende Angestellte mit …“) zutrifft, eine Wahrscheinlichkeit von 63% auf, zu diesem Typus zu gehören.

Beispiel 2: Die Analyse von Marktforschungsdaten ergibt, dass männliche Kunden zwischen 30 und 45 Jahren, welche bei der X-Bank ein Girokonto mit Onli-

ne-Zugang sowie ein Sparkonto unterhalten, zu 80% festverzinsliche Wertpapiere bei einem Konkurrenzinstitut besitzen. Im Analogieschluss gilt daher, dass diejenigen Kunden der X-Bank, auf welche die genannte Merkmalskombination zutrifft, mit 80% Wahrscheinlichkeit zu der anzusprechenden Zielgruppe für eine geplante „Abwerbeaktion" gehören, da sie ein interessantes Verlagerungspotenzial zur X-Bank aufweisen.

Voraussetzung für die Übertragung der Ergebnisse ist das Vorhandensein der maßgeblichen Strukturvariablen sowohl in den Marktforschungsdaten als auch in den internen Daten. Diese stellen die „Brücke" des Data Matching zur Übertragung der Ergebnisse dar, die so genannten „Link-Variablen". In der Konzeptionsphase eines Data Matching-Projektes muss das Vorhandensein von inhaltlich aussagekräftigen Link-Variablen überprüft werden. Dabei ist einige praktische Erfahrung notwendig, um abschätzen zu können, welche Art von Link-Variablen (z.B. Produktnutzung oder Kaufhistorie) für das entsprechende Marktforschungs-Thema die besten Ergebnisse beim Data Matching erwarten lässt und daher gezielt zur Datenbasis ergänzt und gepflegt werden sollte. Reine Soziodemographie ist in den meisten Fällen dafür sicher nicht ausreichend.

Marktforschungsdaten müssen anonym sein

Die Anonymität der Marktforschungsdaten erfordert die Erstellung komplexer Data Matching-Modelle zur Übertragung der Ergebnisse auf personalisierte, interne Unternehmensdaten.

Diese aufwändigere Vorgehensweise hat aber auch klare Vorteile, was die Verlässlichkeit der Informationen sowie die mit den Erhebungen verbundenen Kosten betrifft:

- Die Anonymität der Befragungssituation ermöglicht es, „echte" und verlässlichere Aussagen von den Befragten zu bekommen. Bei personalisierten Befragungen muss man damit rechnen, dass „politische" Antworten gegeben werden.
- Mit Data Matching-Modellen werden die Ergebnisse von einer Stichprobe auf die gesamte Kundendatenbank übertragen, wodurch es nicht mehr nötig ist, alle Kunden der internen Datenbank zu befragen. Diese Vorgehensweise

ermöglicht – gerade bei sehr hohen Kundenzahlen – nicht zu vernachlässigende Kosteneinsparungen.

Anwendung des Data Matching Modells
≠ Übergabe der Studienergebnisse!

Kunden-ID in Kunden-datenbank	Geschätzte Kaufwahr-scheinlichkeit
100356092	0,47
100356106	0,33
100356237	...

Schätzung: Anwendung des Data Matching Modell

Interview-ID in Studie	Produkt-affinität
0023	Sehr hoch
0049	Hoch
0105	...

Reale Studiendaten

➔ Es gibt keine Übergabe der personalisierten Studiendaten an den Kunden!

Voraussetzungen für ein gutes Gelingen

Damit das Data Matching auch zielgerichtet und erfolgreich verläuft, sind schließlich noch einige Punkte zu beachten:

- Die Kundendatenbank sollte mit ihren Inhalten vorab „gesichtet“ werden, um die generelle Machbarkeit des geplanten Data Matchings fundiert und umfasssend zu prüfen.
- Zur Bestimmung der besten Link-Variablen sind mehrere Testläufe notwendig (Feasibility-Test).
- Alle ausgewählten Link-Variablen müssen auch in der Kundendatenbank vorhanden sein.
- Die Trennschärfe des Data Matching-Verfahrens ist durch die Verfügbarkeit von geeigneten Link-Variablen limitiert.
- Die Feinheit der Umfrageinformationen, welche mit Hilfe von Data Matching übertragen werden können (z.B. die Anzahl an Segmenttypen aus einer Kundensegmentierung), ist durch die Menge der verfügbaren Fälle in der Befragungsstichprobe limitiert.

Fazit

Data Matching-Modelle sind eine bewährte Technik, um die Lücke zwischen internen Kundendaten und externen Marktforschungsergebnissen (Einstellungsdaten, Marktstrukturdaten) zu schließen. Sie ermöglichen Unternehmen, über den eigenen Tellerrand hinaus zu schauen, sie schärfen den Blick auf deren Kunden, lassen die Unternehmen ihre Kunden holistischer begreifen und damit besser verstehen. Diese Zusatzerkenntnisse kann ein Unternehmen effizient und gewinnbringend einsetzen, um

- in sich homogene Zielgruppen zu identifizieren und anzusprechen, da dadurch Kosteneinsparungen und Effizienzgewinne in Bezug auf Response-Raten, z.B. bei Direkt-Mailing-Aktionen, zu erwarten sind,
- Re-selling-/Cross-selling-/Up-selling-Möglichkeiten zu identifizieren und zu nutzen,
- Abwanderungs-/Kündigungsrisiken zu erkennen und Gegenaktionen zu starten, um profitable Kunden zu halten, bevor diese kündigen sowie
- den Kundenwert (Customer Lifetime Value) seiner Kunden zu steigern, sowohl durch gezieltes Investment in profitable Kunden als auch durch Minimierung der Ansprache nicht-profitabler, für das Unternehmen uninteressanter Kunden.

Weiterführende Literatur

Schweiger, A.; Wilde, K.J. (1993): Database Marketing – Aufbau und Management. In: Hilke, W. (Hrsg.): Direkt-Marketing, Wiesbaden, S. 89–125.

Witten, H., Frank, E. (2000): Data Mining – practical machine learning tools and techniques with JAVA implementations. Morgan Kaufmann Publishers, Inc. San Francisco, CA.

Fayad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996): Advances in Knowledge Discovery and Data Mining. AAAI Press, MIT Press, Cambridge, Massachusetts, USA.

Integration von Umfragedaten und mikrogeografischen Informationen

Raimund Wildner[1]

1 Vorbemerkung

Die GPS-Technologie (dabei steht das Kürzel für „Global Positioning System"), die seit dem 8.12.1993[2] eine satellitengestützte Standortbestimmung auf ca. 10 m Genauigkeit erlaubt, hat die Möglichkeiten der Mikrogeografie vervielfältigt: Zu ihrer Anwendung wurde ganz Deutschland (wie andere Länder auch) mit seinen Adressen und dem Wegenetz verortet. Dadurch wird es möglich, von jeder Adresse zu jeder anderen Adresse in Deutschland die Wegezeit per Auto oder zu Fuß zu bestimmen.

Für die Marktforschung bietet diese Technologie die Möglichkeit, neue und für den Marktforschungskunden interessante Auswertungen zu generieren, wenn Interviews mit den Wohnortkoordinaten des Interviewten versehen werden. Einige dieser Möglichkeiten sollen im Folgenden dargestellt werden.

2 Die Ausgangsdaten und ihre Verknüpfung

Üblicherweise werden bei Umfragen keine Adressen erhoben und zusammen mit den Umfragedaten gespeichert. Dem stehen schon Erwägungen des Datenschutzes entgegen, denn zumindest bei Einfamilienhäusern wäre dann eine einfache und eindeutige Identifizierung der antwortenden Person möglich.

Eine Ausnahme bilden die Panelteilnehmer; hier muss schon aus Gründen der Kommunikation mit den Panelteilnehmern und ihrer Vergütung die Adresse vorhanden sein. Doch auch hier scheidet eine Auswertung der Paneldaten (z.B. beim Haushaltspanel die Einkaufsdaten) in Verbindung mit der Adresse aus den genannten Datenschutzgründen aus.

Es ist jedoch möglich, die Anschriften so genannten „Mikrozellen" zuzuordnen. Mikrozellen in dem GfK-Instrument „Point Plus" sind kleine Einheiten mit durchschnittlich je sieben Haushalten. Für diese Zellen liegen folgende Informationen vor:

---

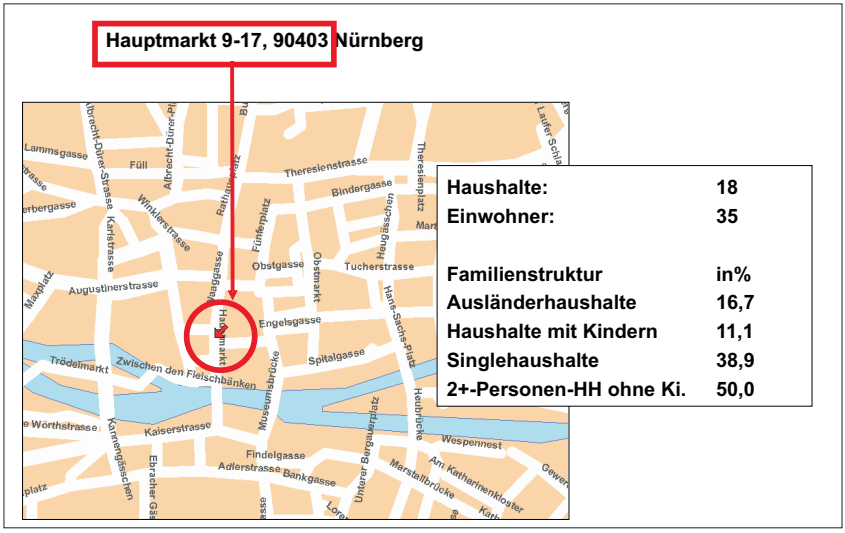
1 Dr. Raimund Wildner ist Leiter der Methoden- und Produktentwicklung der GfK AG und Geschäftsführer des GfK-Nürnberg e.V. (raimund.wildner@gfk.de)

2 vgl. http://www.kowoma.de/gps/Geschichte.htm

- Altersstruktur
- Sozialer Status
- Kinderanteil / Singleanteil
- Ausländeranteil
- Bebauungsstruktur / Haushaltsgröße
- Kaufkraft
- Gewerbestruktur
- PKW-Bestand
- Regional- und Straßentyp

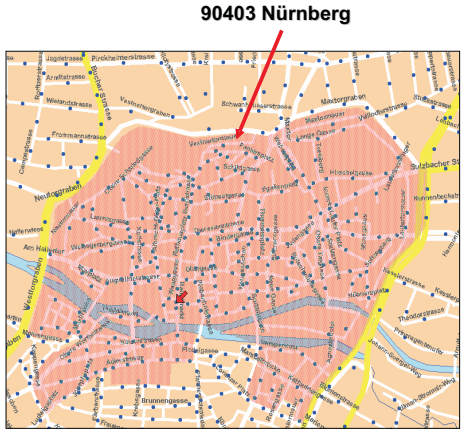
Dabei sind die in der Datenbank enthaltenen Daten Ergebnisse umfangreicher Modellierungen. So wird die Altersstruktur aufgrund einer Datei ermittelt, die von 15 Millionen Personen Vornamen und Altersangaben enthält. Dies ermöglicht eine Vornamensanalyse, die teilweise auf regionaler Basis durchgeführt werden muss. Diese Zusammenhänge werden dann auf die Mikrozellen übertragen. Mehrere Mikrozellen lassen sich dann zu Postleitbezirken oder auch politischen Einheiten wie z.B. Gemeinden zusammenfassen

Abbildung 1 zeigt beispielhaft eine solche Mikrozelle, Abbildung 2 eine Aggregation zu einem Postleitbezirk.



Quelle: GfK

Abb. 1: Beispiel für Mikrozelle



Quelle: GfK

Abb. 2: Beispiel für Aggregation zu Postleitzahlenbezirk

Neben den Point Plus-Regionaldaten bilden die Haushaltspaneldaten der GfK die zweite wichtige, in die Regionaldaten zu integrierende Datenbasis. Im Haushaltspanel der GfK berichten 17.000 Haushalte folgende Informationen zu ihren Einkäufen in den Bereichen Nahrungsmittel und Getränke, Wasch-, Putz- und Reinigungsmittel, Körperpflege und Kosmetik:

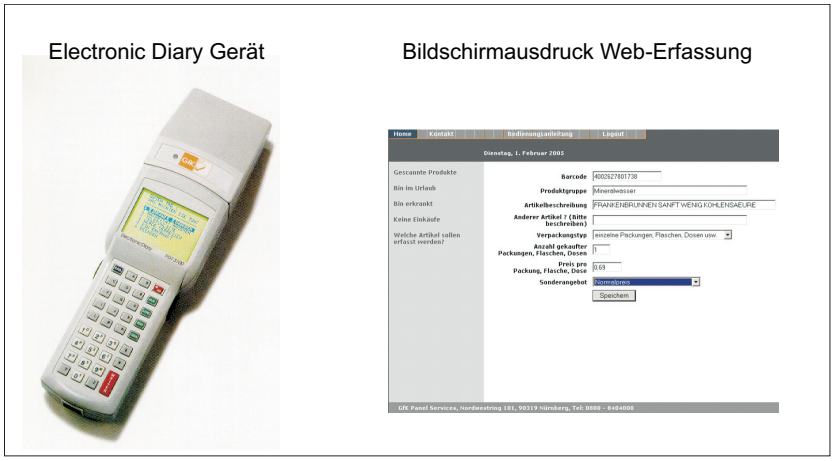
- Wann gekauft? → Tag des Einkaufs
- Was und wie viel gekauft? → Artikel und gekaufte Menge
- Wo gekauft? → Einkaufsstätte
- Wie viel bezahlt? → Preis
- Besondere Bedingungen? → Aktionskauf oder Normalkauf.

Diese Daten werden auf zwei verschiedenen Wegen erfasst:

- 13.000 Haushalte verwenden das „Electronic Diary-Gerät", einen Mikrocomputer, der mit einem CCD-Scanner und mit alphanumerischer Tastatur ausgestattet ist und der eine einfache, menügesteuerte Eingabe der Daten erlaubt.
- 4.000 Haushalte verwenden einen kleinen Handscanner, mit dem die Artikel gescannt werden. Diese gelesenen EAN werden über die USB-Schnittstelle zunächst auf den Computer und dann über das Internet an die GfK übertragen. Von dort werden die Artikeltexte wieder zurück gespielt und der Panel-

haushalt gibt zu jedem Artikel die weiteren Informationen wie z.B. die gekauften Stück und den Preis ein.

Abbildung 3 zeigt das Electronic Diary-Gerät sowie einen Bildschirmausdruck der Erfassung über das Internet.



Quelle: GfK

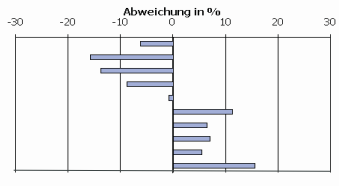
Abb. 3: Datenerfassung im Haushaltspanel

Wenn die Daten nun integriert werden sollen, dann ist zu berücksichtigen, dass trotz der Größe des Haushaltspanels von 17.000 Haushalten im Vergleich zu den 5 Millionen Mikrozellen ein krasses Missverhältnis besteht. Die Integration geht daher auf folgende Art und Weise:

1. In einem ersten Schritt werden zu den Haushalten die jeweiligen Mikrozellen ihrer Wohnadresse zugeordnet.
2. Sodann werden die Merkmale des Wohnumfelds aus der Point Plus-Datei dazugespielt.
3. In einem dritten Schritt wird der Zusammenhang zwischen den Einkaufsintensitäten und den Merkmalen des Wohnumfelds analysiert. Hierbei ergeben sich dann Analysen, wie sie die Abbildung 4 am Beispiel eines Merkmals für den Kauf von Gemüse zeigt.
4. Fasst man diese Analysen zusammen, so ergibt sich ein Steckbrief für den Kauf der jeweiligen Warengruppe. Am Beispiel von Frischgemüse sieht der wie folgt aus:
   - Haushalte mit Kinder: Ein hoher Anteil von Haushalten mit Kindern führt zu hohen Gemüseausgaben.

- Große Wohnhäuser: Erhöht sich der Anteil der 20- und mehr Familienhäuser, so erhöhen sich auch die Ausgaben für Gemüse.
- Hohes Einkommen: Ausgaben für Gemüse sind höher in Ortsteilen mit hoher Kaufkraft als in solchen mit geringer Kaufkraft.
- Junge Haushalte: Ausgaben für Gemüse sinken bei zunehmendem Anteil von Senioren im Ortsteil.

Kaufkraft im Ortsteil	Haushalte	Durchschnittsausgaben für Gemüse	Index der Gemüseausgaben	Abweichung vom Durchschnitt
0: sehr gering	464	137,5	93,9	-6,1
1	499	123,5	84,3	-15,7
2	481	126,4	86,3	-13,7
3	530	133,6	91,2	-8,8
4	473	145,3	99,2	-0,8
5	455	163	111,3	11,3
6	473	156	106,5	6,5
7	522	156,9	107,1	7,1
8	491	154,5	105,5	5,5
9: sehr hoch	496	169,3	115,6	15,6
insgesamt	4.884	146,5	100,0	0,0



Lesebeispiel: Haushalte, die in einem Ortsteil mit sehr hoher Kaufkraft wohnen kaufen 15,6% mehr Frischgemüse als der Durchschnitt der Haushalte

Quelle: GfK

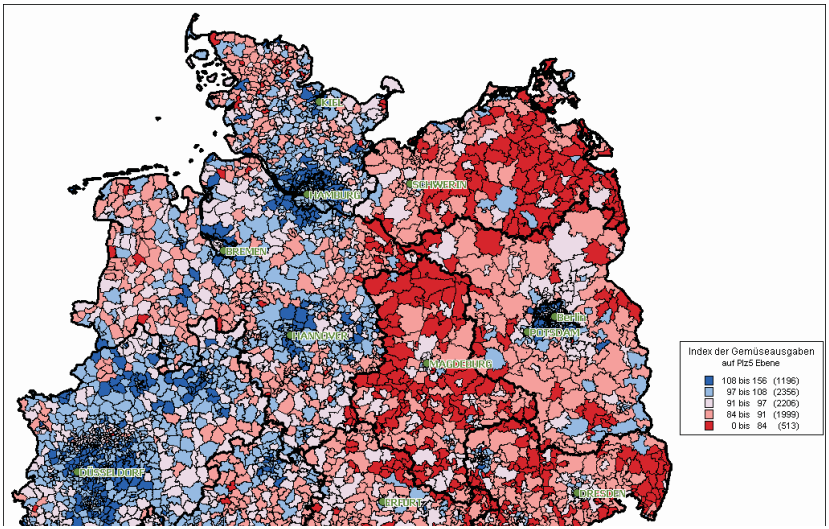
Abb. 4: Statistische Analyse für die Kaufmenge von Gemüse

3 Die Nutzung der Datenverknüpfung

3.1 Standortbewertung und Optimierung von Direktmarketingmaßnahmen

Mit dieser Information lassen sich nun die verschiedenen Intensitäten der Ausgaben für eine Kategorie darstellen. Die kartografische Darstellung (vgl. Abbildung 5) erlaubt dem Handel eine Bewertung von potenziellen Standorten sowie die Optimierung der Streuung von Postwurfsendungen und Probenverteilungen.

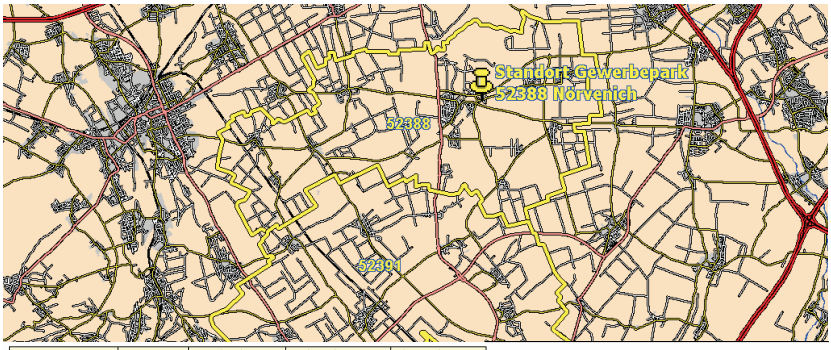
So sind die dunklen Bereiche in Ostdeutschland rot eingefärbt und signalisieren stark unterdurchschnittliche Pro-Kopf-Ausgaben für Gemüse, während die dunklen Bereiche um Hamburg sowie im Ruhrgebiet stark überdurchschnittliche Ausgaben anzeigen.



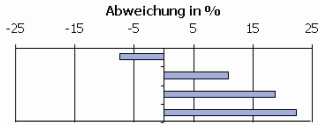
Quelle: GfK

Abb. 5: Karte für die Ausgabenintensität für Gemüse nach PLZ

Betrachtet man sich nun einzelne (potenzielle) Standorte mit ihren jeweiligen Einzugsgebieten, so zeigt sich, an welchen Standorten welche Warengruppen unter- und welche überdurchschnittliche Chancen haben. So ist der eine Standort in Nörvenich als ländlicher Raum bei Gemüse eher benachteiligt, hat aber Chancen bei Kaffee, Waschmittel und Wasser (vgl. Abbildung 6), bei einem anderen Markt in Düsseldorf ist es dagegen umgekehrt (vgl. Abbildung 7).

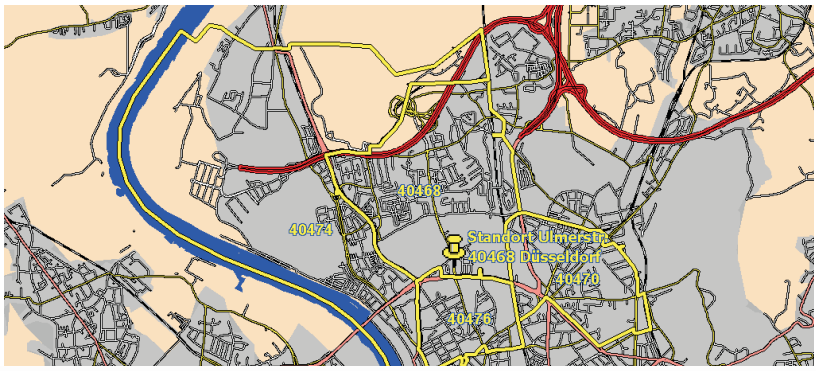


Warengruppe	Haushalte	Durchschnittsausgaben für Warengruppe	Index der Warengruppenausgaben	Abweichung vom Durchschnitt
Gemüse	7.451	132,7	92,5	-7,5
Kaffee	7.451	82,4	110,8	10,8
Waschmittel	7.451	46,2	118,7	18,7
Wasser	7.451	68,8	122,4	22,4

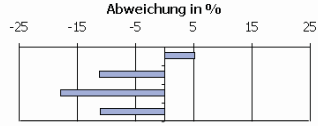


Quelle: GfK

Abb. 6: Bewertung eines Standorts: Beispiel 1



Warengruppe	Haushalte	Durchschnittsausgaben für Warengruppe	Index der Warengruppenausgaben	Abweichung vom Durchschnitt
Gemüse	23.540	150,8	105,1	5,1
Kaffee	23.540	66,1	88,9	-11,1
Waschmittel	23.540	32,0	82,1	-17,9
Wasser	23.540	50,0	88,9	-11,1



Quelle: GfK

Abb. 7: Bewertung eines Standorts: Beispiel 2

Zur Optimierung von Direktmarketingmaßnahmen (adressierte Werbebriefe oder Probenverteilungen) ermittelt man erst die Zielgruppe im Verbraucherpa-

nel, in einem Testmarkt oder auch in einem Produkttest. Dann sucht man nach den Merkmalen, welche in den mikrogeografischen Zellen vorhanden sind und welche die Zielgruppe von den anderen Haushalten möglichst gut trennen, z.B. mit Hilfe einer CHAID- oder einer Diskriminanzanalyse. Die Mikrozellen mit einem besonders hohen Anteil an Zielgruppenhaushalten bzw. –personen liefern dann die Zieladressen der Maßnahme.

3.2 Bewertung von Handelsunternehmen

Die Möglichkeiten der Mikrogeografie und des Verbraucherpanels eignen sich besonders auch, ganze Handelsunternehmen mit allen Filialen zusammen zu beurteilen.

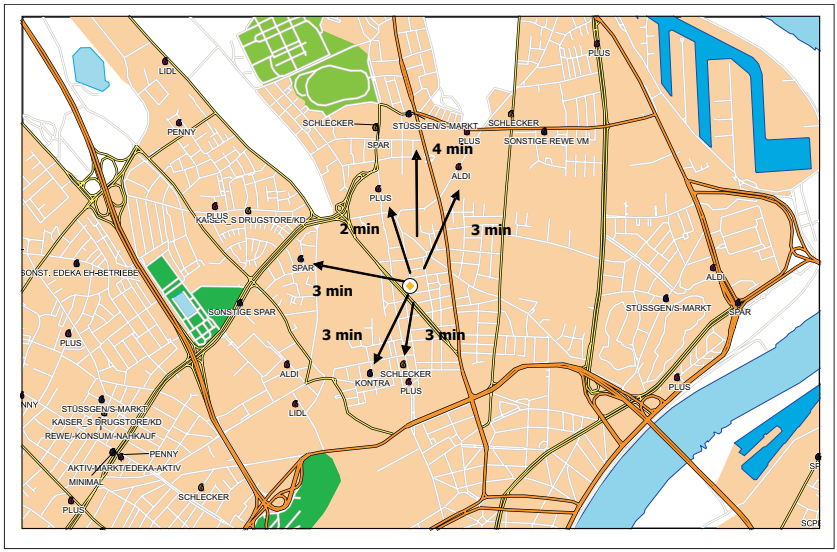
Ein Haushalt kauft pro Jahr im Durchschnitt 218 mal im Jahr Güter des täglichen Bedarfs ein. Dabei besucht er im Durchschnitt 11,7 verschiedene Einkaufsstätten. Welche Einkaufsstätte er aufsucht, hängt ab von

- der Art des Einkaufs (z.B. Kleineinkauf oder Versorgungskauf),
- der Attraktivität der Einkaufsstätte (Sortiment, Preise etc.),
- den persönlichen Gegebenheiten (z.B. ob sie / er über einen PKW verfügt) sowie von
- der Entfernung zur Einkaufsstätte.

Jeder Haushalt hat dabei ein ganz bestimmtes, individuelles Spektrum an Möglichkeiten. Abbildung 8 zeigt beispielhaft das Optionsspektrum eines beliebigen Panelhaushalts XYZ in Köln.

Zentral ist also die Messung der Fahrzeit sowie das Vorhandensein aller Handelsstandorte. Diese Informationen werden in der Kooperation mit der Firma Geni den Haushaltspaneldaten zugespielt. Man erhält dann eine Übersicht über das Optionsspektrum des Haushalts, die Nutzung der Einkaufsstätten sowie den zeitlichen Aufwand, der zum Aufsuchen der Einkaufsstätte erforderlich ist.[3] Die folgende Tabelle zeigt beispielhaft diese Aufstellung für den Haushalt XYZ.

3 Dabei werden derzeit nur die Möglichkeiten vom Wohnort aus analysiert. Es bleibt künftigen Erweiterungen vorgesehen, zusätzlich die Möglichkeiten bei der Arbeitsstätte bzw. auf dem Weg von bzw. zur Arbeitsstätte mit zu berücksichtigen.



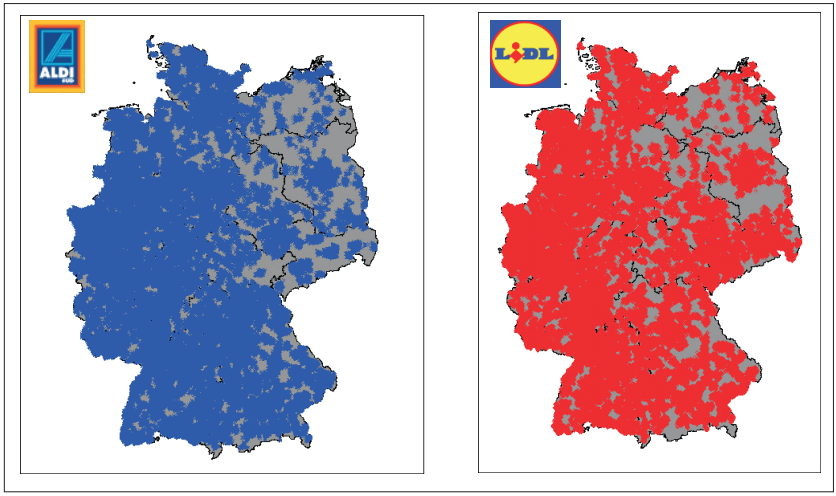
Quelle: GfK / Geni

Abb. 8: Optionsspektrum des Panelhaushalts XYZ in Köln

Tabelle der Optionen des Haushalts XYZ und ihrer Nutzung

KeyAccount	Typ	Fahrzeit (min)	Umsatz (€)
KAUFLAND	SBW	7	
REAL	SBW	9	624,10
TOOM	SBW	11	
WAL-MART	SBW	15	216,13
GLOBUS	SBW	17	
SPAR	VM/SM	3	
KONTRA	VM/SM	3	
STÜSSGEN/S-MARKT	VM/SM	4	49,59
HL	VM/SM	6	520,46
MINIMAL	VM/SM	7	
HIT	VM/SM	9	
EDEKA NEUKAUF	VM/SM	12	
EXTRA	VM/SM	20	
PLUS	DISC.	2	270,88
ALDI	DISC.	3	232,73
LIDL	DISC.	5	150,22
PENNY	DISC.	10	
SCHLECKER	DM	3	13,40
KAISER´S DRUGSTORE/KD	DM	4	
DM-WERNER	DM	5	212,52
Gesamt			2.290,03

Vergleicht man die beiden Discounter Aldi und Lidl, so stellt man fest, dass beide eine vergleichbare Haushaltsabdeckung erreichen: Ein Aldi-Geschäft kann innerhalb von 15 Minuten von 86,5% aller Haushalte erreicht werden. Bei Lidl beträgt dieser Wert 87,2% (vgl. Abbildung 9).



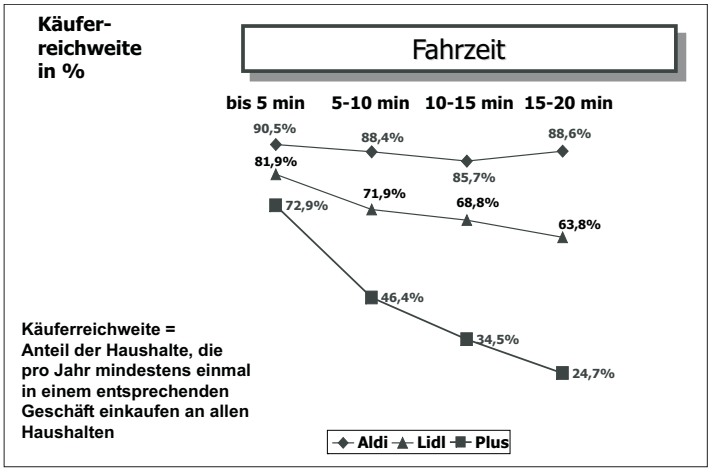
Quelle: GfK / Geni

Abb. 9: Einzugsgebiete bis 15 Minuten Fahrzeit für Aldi und Lidl

Betrachtet man jedoch die Nutzung der Geschäfte durch die Haushalte (vgl. Abbildung 10), so stellt man fest,

1. dass der Aldi schon bei sehr kurzer Entfernung (bis 5 Minuten) von einem höheren Anteil der Haushalte genutzt wird als der Lidl (90,5% vs. 81,9%),
2. dass weiter der Anteil der Haushalte, die den Aldi nutzen, bei wachsender Entfernung sich nur wenig ändert, während beim Lidl ein deutlicher Rückgang zu verzeichnen ist. So nutzen bei einer Entfernung von 15 bis 20 Minuten noch 88,6% der Haushalte den Aldi aber nur noch 63,8% den Lidl.
3. Zum Vergleich dazu erweist sich der Plus als reiner Nahversorger. Nutzen diesen noch 72,9% der Haushalte, die höchstens 5 Minuten von ihm entfernt wohnen, so sinkt der Anteil mit einer Entfernung von 5 bis 10 Minuten bereits auf unter 50% und bei einer Entfernung von über 15 Minuten auf unter ein Viertel.

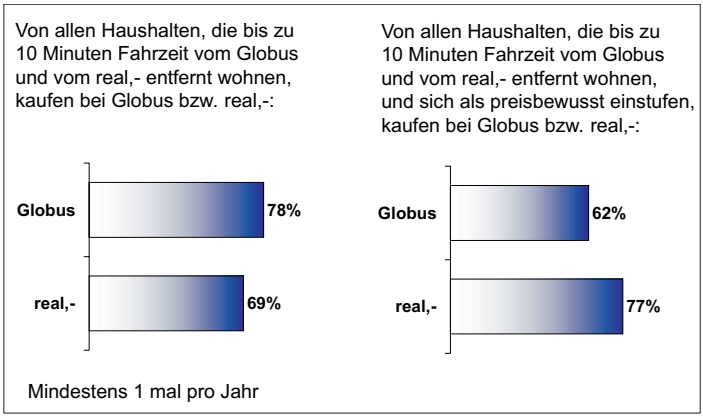
Da kann es nicht überraschen, dass eine Aldi-Filiale pro Jahr durchschnittlich 6,6 Millionen Euro Umsatz verbucht, eine Lidl-Filiale 3,9 Millionen und eine Plus-Filiale etwa 2,5 Millionen.



Quelle: GfK / Geni

Abb. 10: Käuferreichweite in Abhängigkeit von der Entfernung für verschiedene Handelsunternehmen

Möglich ist auch der direkte Vergleich von Handelsunternehmen miteinander, sowohl insgesamt als auch in Bezug auf bestimmte Zielgruppen. Abbildung 11 zeigt, dass bei gleicher Entfernung der Globus zunächst mehr Haushalte anziehen kann als der real,-, dass dieses Verhältnis sich jedoch umkehrt, wenn man die besonders preisbewussten Haushalte herausgreift. Dies wird verständlich, wenn man bedenkt, dass der real,- sich aufgrund attraktiver Aktionen ein besonders gutes Preisimage verschaffen konnte.



Quelle: GfK / Geni

Abb. 11: Vergleich von Handelsunternehmen

4 Ausblick

Die Möglichkeiten der Verknüpfung von Umfragedaten mit mikrogeografischen Informationen wurden anhand einiger Beispiele aufgezeigt. Dabei ist zu berücksichtigen, dass die GPS-Technologie noch immer sehr neu ist, dass daher weitere und sehr interessante Möglichkeiten noch zu entwickeln sind.

Dabei ist jedoch stets auch der Datenschutz im Auge zu behalten. Dies gilt zunächst einmal, weil selbstverständlich die gesetzlichen Bestimmungen und die vom Arbeitskreis der Deutschen Markt- und Sozialforschungsinstitute (ADM) gesetzten Standesregeln zu beachten sind. Dies gilt aber auch im wohlverstandenen Eigeninteresse: Die Auskunftswilligkeit der Befragten ist ein hohes Gut, mit dem pfleglich umgegangen werden muss. Nicht alles, was möglich ist, ist erlaubt und wenn es nicht erlaubt ist, dann ist es auch nicht klug, dies zu tun.

Ersatz von Umfragedaten durch Regionalisierung
Wohnquartiersbeschreibung zur Beschreibung von Interviewausfällen

Jürgen H.P. Hoffmeyer-Zlotnik

1 Einleitung

Die Fragestellung, die sich in der Umfrageforschung immer häufiger stellt, ist die nach dem Umgang mit dem Unit-Nonresponse. Eine Bevölkerungsumfrage zu einem allgemein interessierenden Thema muss akzeptieren, dass nur 50 bis 60 Prozent der mit einem Zufallsauswahlverfahren ausgewählten Zielpersonen bereit sind, sich an der Umfrage zu beteiligen. Wer verbirgt sich hinter den Ausfällen? Gibt es systematische Verzerrungen, einen demographischen und/oder schichtspezifischen Bias? Einige grobe demographische Merkmale bietet die Referenzstatistik des Mikrozensus. Reichen allerdings demographische Merkmale, wenn es darum geht, Einstellungen zu gewichten? Wie lässt sich ein durch Ausfälle entstandener Bias bei Einstellungen durch eine Gewichtung, orientiert an demographischen Merkmalen, korrigieren?

Da in der Regel bestenfalls von den Verweigerern ein paar demographische Eckdaten wie Geschlecht und Alter zu erfahren sind, aber weder von den Verweigerern und erst recht nicht von den Nicht-Erreichbaren Grundorientierungen an Einstellungen in Erfahrung zu bringen sind, bleibt nur der Versuch, über Gruppenspezifika rückzuschließen. Hierbei geht man davon aus, dass spezifische homogene Gruppen einen Konsens an Grundpositionen von Einstellungen aufweisen. Akzeptiert man diese Annahme, dann kann man auch davon ausgehen, dass spezifische Gruppen mit spezifischen Grundpositionen an Einstellungen auch eine gemeinsame Grundposition an Wohnpräferenz haben und danach trachten, unter ihresgleichen in einem homogenen räumlichen Gebiet potentieller sozialer Kontakte zu siedeln. Hinter dieser Annahme steht die Volksweisheit: „Sag mir, wo Du wohnst und ich sag Dir, wer Du bist!“ Die Technik, solche weitgehend homogen besiedelten kleinräumigen Einheiten zu identifizieren, ist die Regionalisierung. Die Technik, Personengruppen mit spezifischen Grundpositionen an Einstellungen zu identifizieren, ist die Netzwerkanalyse. Inwieweit stimmen allerdings Netzwerk und Wohnquartier überein?

Im Folgenden soll herausgearbeitet werden, welche theoretischen Bedingungen der sozial-räumlichen Differenzierung hinter dominant besiedelten Wohnquartieren stehen und wie man diese als Typen identifizieren kann. Des Weite-

ren wird aufgezeigt, wie man analysieren kann, welcher Personentyp hinter welchem Typ Wohnquartier anzutreffen ist. In einem dritten Teil wird anhand von zwei regionalen Fallstudien zur Netzwerkanalyse gezeigt, in welchen Fällen Netzwerk und Wohnquartier übereinstimmen.

2 Sozial-räumliche Differenzierung als theoretischer Rahmen

Die Untergliederung der Stadt und die Verteilung der städtischen Bevölkerung in dieser basiert auf einer Reihe von Annahmen (Hoffmeyer-Zlotnik 2000: 141 f.):

- Die Stadt weist eine räumliche Differenzierung auf, dergestalt, dass der Wert des Bodens im Allgemeinen vom Zentrum zur Peripherie kontinuierlich abfällt und dass Nutzungsart und Bebauungsdichte, da diese in einer Wechselbeziehung mit dem Bodenpreis stehen, in Abhängigkeit von der Entfernung vom Zentrum differieren (Carter 1980: 224 ff.). Das bedeutet, dass die Nutzung des Bodens, in der aktuellen deutschen Gesellschaft an der Rendite orientiert, von innen nach außen an Intensität abnimmt. Dieses impliziert aber auch, dass die Bebauungsstruktur (im Sinne einer Geschossflächenzahl) von innen nach außen an Dichte abnimmt (Schinz 1964: 203).
- Die urbane Gesellschaft weist eine soziale Differenzierung auf, die mit einem steigenden Grad der Arbeitsteilung und einer steigenden Selektivität des Zugangs zum Arbeitsmarkt das vorhandene System sozio-ökonomischer Statusgruppen weiter differenziert. Das Ergebnis ist ein steigender Grad der Segregation einer zunehmenden Anzahl von homogenen Statusgruppen.
- Eine unterschiedliche räumliche und bauliche Struktur befriedigt unterschiedliche Wohnansprüche und ermöglicht einen unterschiedlichen Lebensstil, entsprechend den sozio-ökonomischen (Dangschat 1994; 1997) und lebenszyklischen (Fischer 1982; Falk 1998) Erfordernissen (Knauss 1981; Friedrichs 1995).
- Der städtische Wohnungsmarkt ist kein einheitlicher. Er zerfällt in eine Reihe von Wohnungsteilmärkten mit unterschiedlichen Zugangsmöglichkeiten, entsprechend dem sozialen (Denton & Massey 1988; Kecskes 1997), ethnischen (Hoffmeyer-Zlotnik 1977; 1982; Friedrichs 1998) und/oder ökonomischen Status (Hamm 1982: 114 ff.; Eekhoff 1987) der Wohnungssuchenden (Ipsen 1980; Eekhoff 1987; Hoffmeyer-Zlotnik 1995). Die Zugänglichkeit zu den unterschiedlichen Wohnungsteilmärkten ist selektiert in Abhängigkeit von der Zugehörigkeit eines Haushalts zu einer Statusgruppe. Damit wirkt sich der selektive Wohnungsmarkt segregierend aus.

- Durch die Selektion des Zugangs und der daraus folgenden Segregation der Statusgruppen stellt das unmittelbare Wohnumfeld ein Kernstück des individuellen Aktionsraumes dar (Herlyn & Herlyn 1976: 103).
- Durch die homogenen Gruppenstrukturen ihrer Bewohner stellen Wohnviertel oder Nachbarschaften ein potentielles System sozialer Kontakte dar (Gans 1961: 136 f.; Zinn 1978; Hoffmeyer-Zlotnik 1995).

Hinter diesen Annahmen stehen Theorien der sozialen Differenzierung, der räumlichen Differenzierung und einer Verbindung beider miteinander.

2.1 Soziale Differenzierung

Soziale Schichtung (Hoffmeyer-Zlotnik 1986: 19-24) ist ein in jeder arbeitsteiligen Gesellschaft anzutreffendes Strukturmerkmal sozialer Differenzierung und vertikaler Gliederung der Bevölkerung. Der hierarchische Aufbau einer Gesellschaft differenziert sich mit steigender gesellschaftlicher Arbeitsteilung stärker aus. Über die soziale Schichtung (differenziert nach objektiven sozialen Merkmalen wie Beruf, Bildung, Einkommen und Vermögen) werden die unterschiedlichen Bevölkerungsgruppen bewertend in eine „gesellschaftliche Rangskala" eingestuft (Schäfers 1981: 54). Verbinden sich mit der Gruppenzugehörigkeit objektive Merkmale, so mündet das Gruppenzugehörigkeitsgefühl in einem „Wir-Bewusstsein" (Schelsky 1961: 414). Andererseits bestimmen subjektive Faktoren des Verhaltens die Gewichtung objektiver, „werthaltiger" Kriterien. Denn die Verbindung von einer bestimmten Art der Wertvorstellungen und die Begründung, warum gerade diese Werthaltungen bei den Mitgliedern einer Subgruppe eines sozialen Systems in gewisser Hinsicht übereinstimmen, macht das Gruppenbewusstsein und den Gruppenwillen aus. Objektiv werden die werthaltigen Kriterien über den sozialen bzw. sozio-ökonomischen Status einer Person oder Gruppe definiert und diese beinhalten die Gesamtheit der Positionen und Rollen, die ein Mensch in der Gesellschaft inne hat. Subjektiv bedarf es einer „Prestigezuordnung", da sich mit jedem sozial relevanten Kriterium ganz bestimmte Verhaltenserwartungen an ein Individuum verbinden.

Typisiert man die Subgruppen in abgrenzbare, hierarchisch geordnete Bevölkerungsgruppen mit eindeutiger Mitgliedschaft, so erhält man eine soziale Strukturierung der Gesellschaft. Allerdings findet innerhalb der Subgruppen ständig eine weitere Ausdifferenzierung auf regionaler wie positionaler Ebene statt. Und die vertikale Anordnung der Gruppen nimmt auch horizontale Aspekte der Differenzierung auf. Die Relation der Gruppen zueinander wird abhängig vom eigenen Standort. Der gültige Maßstab wird damit zu einem subjektiven. Soziale Distanz als „Grad von Ferne oder Nähe im sozialen Raum" regelt die Beziehung der Gruppen zueinander. Der Grad der sozialen Distanz zeigt das „Ausmaß der Annäherungsbereitschaft einer Person/Gruppe gegenüber einer

anderen Person/Gruppe“ (Friedrichs 1977: 85). Gruppenabgrenzung zur Demonstration sozialer Distanz geschieht oft über das Errichten oder Aufrechterhalten einer räumlichen Distanz. Dieses heißt: Soziale Gruppen tendieren dahin, segregiert unter ihresgleichen, in einem potenziellen System sozialer Kontakte zu leben (Gans 1961: 136 f.; Zinn 1978; Hoffmeyer-Zlotnik 1995).

2.2 Räumliche Differenzierung

Eine Stadt ist kein einheitliches Ganzes, sondern sie untergliedert sich in unterschiedliche Teilgebiete von unterschiedlicher Standortqualität. Diese Unterschiede sind gegeben durch die räumliche Lage eines Teilgebietes innerhalb einer Stadt, durch dessen Nutzung, durch dessen bauliche Struktur und durch dessen Ausstattung mit „Gelegenheiten“ einer öffentlichen und privaten Infrastruktur. Städtische Teilgebiete weisen also, bedingt durch ein Set von Faktoren, eine unterschiedlich zu bewertende Wohnqualität und, damit verbunden, Lebensqualität auf.

Nimmt man das idealtypische Modell der Stadtentwicklung von Burgess (1925; 1929) zum Ausgangspunkt der Betrachtung der Stadt (siehe Friedrichs 1995: 38 ff.) und transformiert dieses auf die Situation der modernen mitteleuropäischen Stadt von heute, so lassen sich für eine Wohnquartiersbeschreibung als relevant folgende Annahmen zur Struktur der Stadt zusammenfassen (siehe Hoffmeyer-Zlotnik 2000: 142 f.):

- Städte weisen eine innere Gliederung auf.
- Die innere Gliederung von Stadt ...
- weist bei der Bevölkerungsverteilung zonale (Lebenszyklus), sektorale (sozialer Status) und punktuelle oder geklumpte (ethnische) Strukturen auf (Lichtenberger 1998: 143 ff.; Friedrichs 1995: 81; Simon 1990);
- weist bei der Nutzungsverteilung zonale (Arbeitsstätten im sekundären und tertiären Sektor) und sektorale (Wohnen) Strukturen auf.
- Zentren unterschiedlicher Hierarchie stellen Klumpenstrukturen dar, die von einer zonalen Struktur umgeben sind, bestehend aus (potentiellem) Expansionsbereich und konkreten Einzugsbereich(en) (Hoffmeyer-Zlotnik 1977: 18).
- Nutzungen und Bevölkerungsgruppen sind ungleich, aber nicht zufällig, über das städtische Gebiet verteilt.
- Soziale Differenzierung führt zu räumlich differenziertem Siedeln von Bevölkerungsgruppen in gruppendominant besiedelten Wohnquartieren (Segregation).
- Sozialer Aufstieg von Personen/Haushalten veranlasst diese, in ein statusadäquates Wohnquartier umzuziehen. Statusadäquanz ist definiert über grup-

penspezifische Wohnpräferenzen. Sozialer Abstieg führt zu einem Verdrängungsprozess.

- Die Eigentümer und ihre Agenten (z.B. Makler und Wohnungsbaugesellschaften – aber auch Stadtplaner) können die einmal stattfindende Entwicklung auf dem Wohnungsmarkt nicht umkehren; sie können aber einen Trend verstärken (über die Qualität, über den Preis und über Zugangsbeschränkungen).
- Je größer und differenzierter eine Stadt wird, desto vielfältiger ist diese durch unterschiedliche Nutzungen (Wohnen, Arbeiten, Erholung, Verkehr, Versorgung, Bildung, Kommunikation; im Zentrum zusätzlich: Administration und Repräsentation) geprägt.

2.3 Sozial-räumliche Differenzierung

Dieses potenzielle System sozialer Kontakte in räumlicher Abgrenzung verlangt ein relativ homogenes soziales Umfeld, zumindest ein gruppendominiertes soziales Umfeld, das räumlich überschaubar ist, um hierin den einmal erreichten Gruppenstatus zu leben und zu demonstrieren und sich von den statusrangniedrigeren Gruppen mehr oder weniger stark (auch räumlich) abzugrenzen.

Dieses statusgruppenspezifische Wohnen wird unter anderem über einen segmentierten Wohnungsmarkt, also über Wohnungsteilmärkte gestützt und gelenkt (Ipsen 1984: 19-38). Nicht jeder kann in jeder Wohnlage siedeln, auch wenn er meinte, sich dieses von den Wohnkosten her leisten zu können. Es gibt wenigstens drei (bis vier) große gruppenspezifische Wohnungsteilmärkte, deren Zugänglichkeit weniger über den Mietpreis (vgl. Ipsen, Glasauer & Heinzel 1981: 33) als vielmehr über die Vergabepraxis und Zugangskontrollen der privaten Wohnungsanbieter reguliert wird (Kreibich 1985; Häußermann 1996: 18 f.). Die vorzufindenden Wohnungsteilmärkte, zwischen denen erhebliche Barrieren das Umzugsverhalten für die sozial schwächeren Gruppen erschweren, stehen jeweils einer bestimmten Teilpopulation der Gesamtbevölkerung offen. Allerdings haben die statushohen Bevölkerungsgruppen die größeren Wahlmöglichkeiten: Ihnen steht es weitgehend frei, auch in den Bereichen, die den statusniedrigeren Gruppen vorbehalten sind, Wohnungen anzumieten. Mit sinkendem sozialen Status sinkt allerdings auch die Wahlmöglichkeit auf dem Wohnungsmarkt: Die unteren sozialen Gruppen bleiben in der Regel auf den ihnen vorbehaltenen Teilmarkt beschränkt. Damit bietet die Teilmarktregulierung ein Instrument zur Aufrechterhaltung von sozial-räumlicher Differenzierung städtischer Bevölkerung.

Qualitätsänderungen auf dem Wohnungsmarkt geschehen:

- über Instandhaltung, womit der Qualitätsabfall eines Wohnquartiers in einem Prozess des „filtering down“ verlangsamt wird (Eekhoff 1987: 13),
- über Modernisierung, womit ein „filtering up“ dargestellt wird (Eekhoff 1987: 13)
- und über Neubautätigkeit.

Das Niveau der Instandhaltungsmaßnahmen wird von den Preisunterschieden zwischen den einzelnen Teilmärkten bestimmt (Eekhoff 1987: 13). Neubautätigkeit und Modernisierung, soweit nicht über staatliche Programme subventioniert, erschaffen ein Wohnraumangebot für gehobene Ansprüche zu gehobenen Preisen (Miete oder Eigentum). Die unteren sozio-ökonomischen Gruppen werden im Fall der Neubautätigkeit über den Filtering-Prozess bedient (Ratcliff 1949; Ipsen, Glasauer & Heinzel 1981; Glasauer 1986): Durch das zusätzliche Angebot für gehobene Statusgruppen oder Personen mit „gehobenem“ Einkommen entsteht ein Angebotszuwachs auf den Teilmärkten für die unteren Einkommen. Denn durch die Qualitätsveränderung von einzelnen Wohngebäuden oder ganzen Marktsegmenten und in Folge dessen durch Umzüge von einem Teilmarkt auf den, nächsten, werden Wohnungen der oberen Teilmärkte nach unten „durchgereicht“ (Eekhoff 1987: 8 ff.).

Neben der Regulierung des Wohnungsmarktes über das Angebot in Qualität und Menge sowie über den Preis gibt es drei weitere Aspekte der Regulierung:

- den schichtspezifischen Zugang (u.a. Kreibich 1985; Eekhoff 1987: 68; Leitner 1983: 96 ff.),
- den lebenszyklusbedingten Zuschnitt von Wohnung und Wohnquartier (Lichtenberger 1998: 144 f.; Falk 1998),
- den ethnischen Ausschluss (Hoffmeyer-Zlotnik 1977; 1982; Leitner 1983; Friedrichs 1998).

Sozialer Status führt zu einer sektoralen Verteilung über die Stadt, da die oberen Schichten in jenen Sektoren der Stadt anzutreffen sind, in Abhängigkeit von der Topographie, von der vorherrschenden Windrichtung und von der Lage der Industrie, in denen die „klimatischen“ (und Luft-) Verhältnisse die besten sind.

Die Stellung im Lebenszyklus führt zu einer konzentrischen Verteilung der Gruppen über die Stadt, da die jungen Einpersonenhaushalte in der Innenstadt inmitten des „städtischen Geschehens“ siedeln, die jungen Familien mit Kleinkindern an die städtische Peripherie „ins Grüne“ wollen und die Alten wieder in die Städte zur „städtischen Infrastruktur“ zurück wandern. Diese Wanderungsbewegungen geschehen unabhängig von der Schicht und damit innerhalb aller Sektoren.

Ethnische Zugehörigkeit führt (zunächst unabhängig vom sozio-ökonomischen Status und vom Lebenszyklus) zu einem Siedeln innerhalb ethnischer Kolonien. Diese sind punktuell über die Stadt verteilt, meist nach historischen Gegebenheiten: Welcher Siedlungsteilraum wurde den ethnischen Einwanderern überlassen, als diese auf den Wohnungsmarkt drängten, und wo liegen deren Erweiterungsgebiete? In den Städten der alten Bundesländer sind dieses in der Regel die ehemaligen (auch potenziellen) Stadterneuerungsgebiete und die peripheren Hochhaussiedlungen der 60er und 70er Jahre sowie Gebiete mit aktueller oder ehemaliger hoher/höherer Umweltbelastung.

3 Erfassen von Wohnquartieren: Entwicklung eines Instruments

Es gibt in der Stadtforschung einige Möglichkeiten, Quartiere oder Nachbarschaften anhand statistischer Daten (Zensusdaten) zu typisieren und abzugrenzen. Die bekannteste hiervon ist die von Shevky & Bell (1955) entwickelte Methode der Sozialraum-Analyse (in der Anwendung für eine deutsche Großstadt siehe auch Friedrichs 1977: 197 ff.). Die Probleme in der Bundesrepublik Deutschland liegen aktuell bei der Verfügbarkeit entsprechender Daten: Für eine Typisierung von Wohnquartieren benötigt man kleinräumig verfügbare Daten einer Volks- und einer Gebäude- und Wohnungszählung, die auf Blockebene (besser Blockseitenebene) aufbereitet und zugänglich sind. Als Alternative bleibt die Begehung oder die Beschreibung der Wohnquartiere in Verbindung mit dem Forschungsprozess oder die Nutzung der Quartier- oder Milieu-beschreibenden geographischen Systeme der Marktforschung wie z.B. jenes der GfK-Regionalforschung (z.B. in diesem Band) oder jene von infas-GEOdaten oder von microm (siehe: Regionale Standards 2005).

Die Wohnquartiersbeschreibung ist ein Instrument der Regionalisierung und stellt damit eine Klassifikation von dinglichen Merkmalen und/oder Individuen nach bestimmten Eigenschaftsdimensionen dergestalt dar, dass deren räumliche Ordnungsstruktur (Raumstruktur) weitgehend berücksichtigt und dass das Kontingenzprinzip gewahrt wird (Hard 1973: 87). Damit stellt Regionalisierung eine Typisierung nach Merkmalen in räumlicher Abgrenzung dar, wobei entweder benachbarte Raumelemente mit gleicher Merkmalsstruktur addiert oder über Verflechtungsmerkmale strukturiert werden. Im Endergebnis interessiert nicht der konkrete geographische Raum, sondern der Raumtyp.

3.1 Die zentralen Variablen:

Für eine Typisierung städtischer Wohnquartiere müssen jene Merkmale herausgefiltert werden, die spezifische Quartierstypen charakterisieren: Diese Merk-

male sind reduziert auf das Minimum des Notwendigen. Dieses sind die Variablen: *Lage*, *Dichte* und *Nutzung* (vgl. Burgess 1925; Boustedt 1966; Hoffmeyer-Zlotnik 1984). Erhoben werden diese Variablen für einen *Sichtbereich*. Ein Sichtbereich ist jener Bereich, den eine Person wahrnehmen kann, wenn diese sich an einen definierten Standort, z.B. vor eine Haustür, stellt und sich einmal um die eigene Achse dreht. Hierbei werden Lage, Dichte und Nutzung entsprechend den vorgegebenen Kategorien pro Sichtbereich notiert. In einem weiteren Schritt werden die Variablen zu Indizes zusammengefasst. Das Abgrenzen von Wohnquartieren setzt eine Quartiersbeschreibung über Begehung voraus. Wohnquartiere stellen damit das Aggregat von identischen Ausprägungen für die beschriebenen Sichtbereiche dar.

Die Wohnquartiersbeschreibung ist sowohl als Befragung der Zielpersonen einer Umfrage als auch als Begehung durch die Interviewer anzuwenden. Für eine Nonresponse-Analyse sollte diese als Begehung durchgeführt werden. Bei jenen, die an der Befragung teilnehmen, ist aber eine gleichzeitige Befragung sinnvoll, da hierüber Typenbildungen von Einstellungen in Quartierstypen möglich sind.

Lage

Alle Nutzungen, auch die Nutzung „Wohnen“, sind in ihrer Qualität abhängig von der Erreichbarkeit in Raum und Zeit. „Erreichbarkeit“ ist bedingt durch die Lage einer Nutzung bzw. eines Wohnquartiers innerhalb eines Siedlungsraumes. Lage ist gleichzusetzen mit der zurückzulegenden Distanz (zu messen in Wegstrecke) vom Wohnquartier bis zum nächsterreichbaren Zentralen Geschäftsbezirk eines großstädtischen Siedlungsraumes. Die Variable Lage verortet das zu beschreibende Wohnquartier in der inneren Gliederung der Stadtregion und ist ein Maß für die Distanz vom Zentrum (Hoffmeyer-Zlotnik 2000: 172 ff.).

Die zu stellende Frage: „Wie weit ist das (Großstadt-)Geschäftszentrum der Innenstadt von dem Haus, in dem Sie wohnen (bei Befragung)/vor dem Sie stehen (bei Begehung), entfernt?“

Dichte

Die Variable Dichte bezieht sich auf die bebaute Umwelt. Es wird nicht die Bevölkerungsdichte herangezogen, da diese nicht beobachtbar ist. Nur der Gebäudetyp kann bei einer Beobachtung als Indikator für Dichte dienen. Wichtig für die Charakterisierung eines Quartiers ist, dass nicht ein einzelnes Gebäude typisiert wird, sondern dass nach jenem Typ von Wohngebäude gesucht wird, der die unmittelbare Umgebung der Wohnung der Zielperson prägt. In diese Typisierung der in einem Quartier überwiegenden Gebäude gehen Annahmen über Größe und Höhe der Gebäude, Anzahl der Wohnungen und Geschlossenheit

und Kompaktheit der Bebauung ein. Die impliziten Annahmen schließen darüber hinaus auch auf die Art der vorhandenen Wohnungen und darauf, dass unterschiedliche Wohnformen ihrerseits unterschiedliche Lebensstile ermöglichen (Hoffmeyer-Zlotnik 2000: 174 ff.).

Die zu stellende Frage: „Von welcher Gebäudeart sind die Wohngebäude, die - rechts und links sowie gegenüber - oder vor oder hinter Ihrem Wohnhaus gelegen sind? Also, wie sind die Nachbarwohngebäude zu charakterisieren? Welcher Gebäudetyp überwiegt?“

Für diese Aufgabe wird der befragten/der beobachtenden Person eine optische Hilfe angeboten mittels einer Liste mit Fotos bzw. Skizzen von 10 unterschiedlichen Gebäudetypen: „Am besten passt Bild: ... “.

Nutzung

Die Abfrage der Nutzungsart dient der Charakterisierung des Wohnquartiers. Abgefragt werden die Funktionen „Wohnen“, „Arbeit“ und „Versorgung“. Mit Blick auf die Funktion „Arbeit“ werden Gebäudetypen aller drei Wirtschaftssektoren berücksichtigt. Es wird nicht nach der überwiegenden Nutzung, sondern nach dem Mix der vorhandenen Nutzungen gefragt - Mehrfachnennungen sind möglich. Die Nähe zu spezifischen und der Mix von verschiedenen Nutzungen macht den Charakter eines Wohnquartiers aus und ermöglicht Rückschlüsse auf die Lebensqualität in diesem Quartier und auf den Lebensstil von dessen Bewohnern (Hoffmeyer-Zlotnik 2000: 176 f.).

Nutzung wird gemessen über die Abfrage:

„Sind in unmittelbarer Nachbarschaft Ihres/dieses Hauses:

A: nur Wohngebäude,

B: auch eine Ansammlung von mindestens vier Läden mit Gütern für den täglichen Bedarf, die sich unter einem Dach befinden,

C: auch Wohngebäude mit Läden/Kneipen,

D: auch Fabrik(en),

E: auch mindestens ein Geschäfts-, Büro-(Hoch-)haus, Öffentliche Einrichtungen,

F: auch landwirtschaftlich genutzte Gebäude wie Stall, Scheune, Schuppen für Maschinen und ähnliches?“.

Nutzung wird sodann zu einem Index verarbeitet, der folgende 10 Ausprägungen hat (Tabelle 1):

Tabelle 1: Index *Nutzung*

1	landwirtschaftlich genutzte Gebäude
2	Gewerbegebiet neben Landwirtschaft
3	reines Wohnen
4	Wohngebäude mit Läden/Kneipen
5	Fabrik im/am Wohnquartier, Läden/Kneipen möglich
6	Fabrik und Büros/Verwaltung
7	Fabrik und Einkaufszentrum
8	Büros/Verwaltung im/am Wohnquartier
9	Einkaufszentrum im/am Wohnquartier
10	Einkaufszentrum und Büros/Verwaltung im/am Wohnquartier

Hoffmeyer-Zlotnik 2000: 177

3.2 Die Indizes

Lage, *Dichte* und *Nutzung* für sich allein betrachtet ermöglichen allerdings noch keinen Rückschluss auf das Wohnquartier. Hierzu müssen erst diese drei Merkmale kombiniert werden, um Quartiersspezifika hervorzuheben. Dieses geschieht mit Hilfe von drei Indizes: Index *Zentralität*, Index *Urbanität* und Index *Wohnquartier.*

Index Zentralität

Dichte in der Bedeutung von *Art der Bebauung* und Lage in der Bedeutung von Antreffbarkeit des Wohnquartiers in einer bestimmten *Entfernung zum Oberzentrum* sollen als Indikatoren für *Zentralität* gesehen werden. Der Index *Zentralität* wird gebildet über eine additive Verknüpfung der beiden Variablen *Lage* und *Dichte*. Diese Variable dient dazu, die Art der Bebauung in ein konzentrisches Modell von Stadt bzw. Stadtregion einzuordnen. Denn von einem – idealtypisch gedacht – konzentrischen Aufbau der Stadtregion um den Zentralen Geschäftsbezirk einer Großstadt und einem ebenfalls konzentrischen Aufbau unterschiedlicher Stadtteile um nachgeordnete Zentren, gliedern sich die städtischen Wohnquartiere nach *Lage* und *Dichte* grob in Zonen (Hoffmeyer-Zlotnik 2004: 84).

Index Urbanität

Da der Index *Zentralität* nur für Wohnquartiere zu interpretieren ist und, abgesehen vom Zentralen Geschäftsbezirk, weder nachgeordnete Zentren noch andere Bereiche mit Nicht-Wohnnutzung ausweist, müssen die Bereiche mit Nicht-Wohnnutzung einer gesonderten Betrachtung unterzogen werden. Hierzu dient der Index *Urbanität*. Dieser wird ebenfalls über eine additive Verknüpfung

zweier Variablen gebildet: die Addition der Variablen Lage mit dem Index *Nutzung*. Der Index *Urbanität* weist nachgeordnete Zentren auf, die einem monozentrisch ausgerichtetem Modell von Stadt widersprechen. Da alle Großstädte und auch viele Mittelstädte in Mitteleuropa heute eine Mehrzentrenstruktur aufweisen und die Zentren von Kleinstädten in der auf eine Großstadt ausgerichteten Stadtregion nachgeordnete Zentren darstellen, ist der Index *Urbanität* für eine Beschreibung von Wohnquartieren zwingend erforderlich, da sowohl ein Zentrum als auch die an ein Zentrum anschließende Zone von ihren Bewohnern her schwer zu erfassen sind. Ein Zentrum beherbergt Bewohnergruppen mit sehr unterschiedlichem sozialen Status (Hoffmeyer-Zlotnik 2004: 84 f.).

Index Wohnquartier

Der Index *Wohnquartier*, als der dritte und zentrale Index, zeigt auf, welchem Wohnungsteilmarktsegment ein Wohnquartier zuzurechnen ist. Dieser Index gilt immer dann, wenn sich ein selektierter Wohnungsmarkt bei Wohnungsüberschuss über das freie Spiel von Wohnungsangebot und Wohnungsnachfrage, ohne eine Wohnungszwangsbewirtschaftung, frei entwickeln kann. Damit erlaubt dieser Index für die überwiegende Mehrheit der Wohnquartiere in mitteleuropäischen Groß-, Mittel- und Kleinstädten Rückschlüsse auf die in einem bestimmten Wohnquartierstyp wahrscheinlich dominante Bevölkerungsgruppe.

Tabelle 2: Index Wohnquartier für die Großstadt

Code	Wert für *Lage* + *Dichte* + *Nutzung*	Einschränkungen D= *Dichte*, N= *Nutzung*	*Zentralität* (Z) *Urbanität* (U)
1	26 bis 30		Wenn Z= 9 oder 10
2	17 bis 25	wenn N =10	wenn U= 9 oder 10
3	23 bis 25	wenn N kleiner 10;	wenn U kleiner 9
	oder 20 bis 22	wenn N kleiner10 und D größer 7	
4	17 bis 19	wenn N kleiner 10 und D größer 5	
5	8 bis 16	wenn D= 9 oder 10	
6	20	wenn N kleiner 10 und D kleiner 8;	
	der 17 bis 19	wenn N kleiner 10 und D kleiner 6;	
	oder 14 bis 16	wenn D kleiner 9;	
	oder 8 bis 13	wenn D größer 6	
7	8 bis 13	wenn D kleiner 6	
	oder 6 oder 7	wenn D größer 1	
8	11 bis 13	wenn D= 1	
9	3 oder 5		

Hoffmeyer-Zlotnik 2004: 89

Der Index Wohnquartier wird gebildet über eine additive Verknüpfung der gewichteten Werte für *Lage* und *Dichte* mit dem Index *Nutzung*. Hierbei sind Einschränkungen bei den Werten von *Dichte* und *Nutzung* zu beachten sowie die Indizes *Zentralität* und *Urbanität* bei den innerstädtischen Wohnquartierstypen zu berücksichtigen. Der Index *Wohnquartier* besteht für die ethnische Majorität aus drei unterschiedlichen Varianten, je einer für die Großstadt, die Mittelstadt und die Kleinstadt. Tabelle 2 zeigt exemplarisch die Indexbildung für den Bereich der Großstadt. Die Bildung aller drei Varianten ist nachzulesen unter Hoffmeyer-Zlotnik 2004: 85 ff.

Der Index *Wohnquartier* stellt eine Typisierung von Wohnquartieren dar, mit deren Hilfe über einen Quartierstyp auf die in diesem dominante Bewohnergruppe rückgeschlossen werden kann. Die Ausprägungen des Index *Wohnquartier* sind wie folgt (Tabelle 3):

Tabelle 3: Ausprägungen des Index *Wohnquartier*

Wertebereich			**Wohnquartierstyp**
Groß-stadt	Mittel-stadt	Klein-stadt	
1	-	-	Zentraler Geschäftsbezirk, Zentrum erster Ordnung
2	2	2	nachgeordnete Zentren, B- und C-Zentren;
	2	2	auch A-Zentrum von Mittel- und Kleinstadt
3	3	3	„Zone im Übergang“, Innenstadtbereiche
3			altes „Westend“, auch: innerstädtische Altindustrien, Hafen
4	4	-	Mietskasernenquartiere, kompakt bebauter Innenstadtrand
5	5	5	periphere Hochhausgebiete (der Suburbanisierung)
6	6	6	Wohnquartiere der Reihen und Zeilen, in den neuen Bundesländern: auch Plattenbauten
7	7	7	Wohnquartiere der peripheren Einzelhausbebauung
8	8	-	Villenviertel
9	9	9	ländlicher Bereich, Peripherie der Stadtregion

Hoffmeyer-Zlotnik 2000: 179

Die einzelnen Typen des Index Wohnquartier stehen für die nachfolgenden neun Quartierstypen (Hoffmeyer-Zlotnik 2000: 179-181):

(1) Ein „*Zentrum erster Ordnung*“ ist allgemein charakterisiert durch eine zentrale Lage, eine kompakte Bebauung und einen hohen Anteil an Läden und Büros. Über die Fragestellung (für Lage) ist es das Geschäftszentrum der Großstadt. Es ist der Ort der höchsten Vielfalt an Angeboten von Konsum-

gütern und zentralen Dienstleistungen. Die Wohnnutzung spielt in diesem Quartierstyp eine untergeordnete Rolle.

(2) „*Nachgeordnete Zentren*" sind in der Regel in mittlerer Entfernung vom zentralen Geschäftsbereich gelegen und stellen in den Großstädten Stadtteilzentren dar.

(3) Mit „*Zone im Übergang*" wird jener Bereich der Stadt bezeichnet, der sich unmittelbar an ein Zentrum, unabhängig ob erster oder zweiter Ordnung, anschließt. In großen Städten stellen diesen Bereich die das Oberzentrum umgebenden Stadtteile der Innenstadt dar. Ausgewiesen ist dieser Quartierstyp durch eine kompakte Bebauung. Hier befinden sich, zusätzlich zu einer Wohnnutzung mit breitem qualitativen Spektrum, auch Büros und/oder Läden. Die „Zone im Übergang" weist eine Ansammlung heterogener Bevölkerungsgruppen auf, die von den „Gentrifiern" (Blasius 1993; Friedrichs & Kecskes 1996) bis zu den ethnischen Kolonien der alten „Gastarbeiter"-Gruppen (Hoffmeyer-Zlotnik 1982; Friedrichs 1998) reichen.
Sofern vorhanden, ist das „Westend" der Gründerzeit sowohl Expansionsgebiet für einen gehobenen Dienstleistungssektor als auch ein Wohngebiet im Spannungsfeld zwischen „gold coast" und „slum".

(4) In *Hafenstädten* oder altindustrialisierten Städten zählen Teile des Hafengebietes und/oder innenstadtnahe alte Industriewerke (ob „saniert" oder nicht) ebenfalls zur dritten Zone.

(5) Die „*Mietskasernenquartiere*" zeichnen sich in ihrer Mehrzahl durch eine mittlere bis geringe Entfernung zum Zentrum aus. Sie weisen in der Regel eine kompakte Bebauung auf. Es handelt sich hierbei um Gebiete mit gemischter Nutzung, d.h. hier liegen auch Gewerbehöfe und Büros. Die Bewohnerstrukturen sind, insgesamt betrachtet, relativ homogen, unterscheiden sich aber über eine unterschiedliche historische Entwicklung der Regionen oder Städte.

(6) Die „*peripheren Hochhausgebiete*" sind, je nach Lage und nach vorhandenem Zentrum, unterschieden. Es handelt sich hierbei in den westlichen Bundesländern um die Suburbs der 60er und 70er Jahre, die schon seit Anfang der 90er Jahre einen „filtering down"-Prozess aufweisen. Die Mehrheit dieser Quartiere ist heute von statusniedrigen Gruppen bewohnt.

(7) Die „*Wohnquartiere der Reihen und Zeilen*" weisen überwiegend eine mittlere Entfernung zur Innenstadt auf. Die Bebauung besteht aus Mehrfamilien- und Reihenhäusern, es kann aber auch eine Zeilenbauweise vorhanden sein. In den neuen Bundesländern einschließlich Ost-Berlins befinden sich in diesem Quartierstyp die industriell gefertigten „sozialistischen Reihenhäuser" in Plattenbauweise. Die „Reihen und Zeilen" sind oft benachbart

zur Industrie und werden geprägt durch die benachbarte gewerbliche Nutzung.

(8) Die „*Wohnquartiere der peripheren Einzelhausbebauung*“ zeichnen sich aus durch eine mittlere bis größere Entfernung zur Innenstadt. Sie haben eine eher niedrige Bebauung, zumeist bestehend aus freistehenden niedrigen Gebäuden. Diese Wohnquartiere sind in der Regel reine Wohngebiete für das breite Spektrum der Mittelschichten.
Der Suburbanisierung vergleichbar sind die im kleinstädtischen, ländlichen Raum entstandenen Siedlungen von Ein- und Zweifamilienhäusern für eine in die „Natur“ expandierende mittlere Mittelschicht.

(9) Die „*Villenviertel*“ weisen oft eine relativ periphere Lage auf, die in der Regel hinter dem Gürtel der „Mietskasernen“ beginnt. Sie sind durch eine niedrige, freistehende Bebauung für 1 bis 2 Haushalte pro Gebäude ausgewiesen. Die Gebäude werden fast ausschließlich zu Wohnzwecken genutzt. Dieser Wohnquartierstyp beherbergt die alten oberen Schichten. Er tritt pro Stadtregion nicht sehr häufig auf.

(10) An der *städtischen Peripherie* beginnt der „*ländliche Bereich*“ mit niedriger, aufgelockerter Bebauung für Wohnen und in der Regel hoch spezialisierter Landwirtschaft in kleingliedriger Struktur. Über die Stadtflucht und eine zunehmende Ausdehnung der Pendlerzone in der Phase der Desurbanisierung werden immer mehr ehemals ländlich geprägte Dörfer im Einzugsbereich der Städte zu neuen „*Vorstädten*“. Dominiert werden diese neuen Vorstädte von mittleren sozialen Schichten von Städtern.

4 Wohnquartierstypische Bewohner

Erhebt man in einer Umfrage die Wohnquartiersmerkmale in Verbindung mit durchgeführten Interviews, so kann man für einzelne Merkmale der Befragten analysieren, in welchem Typ Wohnquartier welches Merkmal besonders hervor tritt. 1995 wurde die Wohnquartiersbeschreibung in eine allgemeine Bevölkerungsumfrage, den Sozialwissenschaften-Bus 2/1995 eingeschaltet (Random-Route-Stichprobe; befragt wurden Deutsche im Alter ab 18 Jahre in Privathaushalten) und ergab folgendes Bild: In den Großstädten der alten Bundesländer unterschieden sich die Wohnquartiere hinsichtlich der folgenden Befragtenmerkmale signifikant (auf dem 1%-Niveau): Haushaltsgröße, Haushaltseinkommen, Alter und Bildung des/der Befragten und dessen/deren Status zum Erwerbsleben (Hoffmeyer-Zlotnik 2000: 193).

Bildet man einen Index des sozio-ökonomischen Status aus den klassischen drei Variablen „Bildung“, „Stellung im Beruf“ und „Einkommen“ (abgefragt gemäß den Demografischen Standards 1999), so ergibt sich folgendes Bild für

die Verteilung der Statusgruppen über die Wohnquartierstypen (vgl. Hoffmeyer-Zlotnik 2000: 191):

Im Zentrum erster Ordnung dominiert die untere bis mittlere Mittelschicht. Im nachgeordneten Zentrum dominiert die untere Mittelschicht. In der „Zone im Übergang" sind sowohl die untere als auch die obere Mittelschicht die sichtbaren Gruppen. In den „Mietskasernen" findet man das ganze Spektrum der Mittelschichten, mit einer Tendenz zur unteren Mitte. In dem Quartierstyp der „Reihen und Zeilen" dominiert die untere Mittelschicht und in den „peripheren Einzelhausgebieten" findet man wieder das ganze Spektrum der Mittelschichten mit einer Tendenz zur oberen Mitte.

Zur Identifikation der Struktur der dominanten Bewohner in den jeweiligen Wohnquartierstypen wurden die Befragten des Sozialwissenschaften-Bus 2/1995[1], über zentrale sozio-demographische Merkmale typisiert. Durchgeführt wurde diese Typisierung mittels einer Clusteranalyse mit dem Programm CLUSE, hier exemplarisch berichtet für die alten Bundesländer inklusive West-Berlin (alle Ergebnisse sind bei Hoffmeyer-Zlotnik 2004: 94 ff. nachlesbar). Der Vorteil einer Clusteranalyse besteht darin, dass spezifische sozio-demographische Profile der Befragten zu Typen von Personen gebündelt werden. Betrachtet wurde sowohl das Konzept des sozio-ökonomischen Status, SES (siehe u.a. Friedrichs 1995) als auch das Konzept des Lebenszyklus (u.a. Lichtenberger 1998; Falk 1998).

4.1 Wohnquartierstypische Bewohner nach dem Konzept des SES

In die Clusteranalyse wurden als Variablen einbezogen: der höchste allgemeinbildende Schulabschluss, erweitert um einen möglichen Hochschulabschluss, sowie die „Autonomie in der beruflichen Tätigkeit" (siehe Hoffmeyer-Zlotnik 1998), jeweils für die befragte Person, und zusätzlich auf Haushaltsebene das Haushaltseinkommen.

Zur genaueren Beschreibung der SES-Typen wurden die Cluster kreuztabelliert mit der Haushaltsgröße (Anzahl der Personen im Haushalt); dem Status im Erwerbsleben, unterteilt in erwerbstätig, noch nicht erwerbstätig (Schule, Ausbildung, Studium), Hausfrau, nicht mehr erwerbstätige Person (Rentner/Pensionär); dem Alter der befragten Person, unterteilt in vier Altersgruppen; und der Anzahl der Personen, die zum Haushaltseinkommen beitragen.

1 Random-Route-Stichprobe nach dem ADM-Stichproben-Design, mit 2.114 Befragten in den alten und 1.095 Befragten in den neuen Bundesländern. Die Grundgesamtheit der Befragten sind Personen mit deutscher Staatsangehörigkeit im Alter ab 18 Jahren, lebend in Privathaushalten.

In den alten Bundesländern war die optimale Lösung eine mit vier Clustern:

SES 1: *Hoher sozio-ökonomischer Status:* zeichnet sich aus durch Befragte mit hohem Bildungsabschluss, hoher Autonomie in der beruflichen Tätigkeit und hohem Haushaltseinkommen.
Dahinter verbergen sich Erwerbstätige im mittleren Alter mit im Durchschnitt zwei Einkommensbeziehern und keinem oder maximal einem Kind im Haushalt.

SES 2: *Mittlerer sozio-ökonomischer Status*: zeichnet sich aus durch Befragte mit im Durchschnitt mittlerem Bildungsabschluss, mittlerer Autonomie in der beruflichen Tätigkeit und einem gehobeneren Haushaltseinkommen.
Dahinter verbergen sich kleine Haushalte mit ein bis zwei Personen und Befragten im mittleren bis hohem Alter, die vor allem von Hausfrauen und Rentnern repräsentiert werden. Zum Haushaltseinkommen tragen ein bis zwei Personen bei.

SES 3: *Niedriger sozio-ökonomischer Status*: zeichnet sich aus durch Befragte mit mittlerem bis höherem Bildungsabschluss, niedriger Autonomie in der beruflichen Tätigkeit und niedrigem Haushaltseinkommen.
Dahinter verbergen sich junge, alleinlebende Menschen, in starkem Maße auch Studenten.

SES 4: *Niedriger sozio-ökonomischer Status mit hohem Haushaltseinkommen*: zeichnet sich aus durch Befragte mit im Durchschnitt mittlerem Bildungsabschluss, niedriger Autonomie in der beruflichen Tätigkeit. Das Haushaltseinkommen ist hoch.
Dahinter verbergen sich in großer Anzahl große Haushalte mit Schülern und Hausfrauen. In den großen Haushalten tragen auch viele Personen (drei bis vier) zum Haushaltseinkommen bei.

Wie Tabelle 4 zeigt, tendieren in den alten Bundesländern die mittleren bis hohen Statusgruppen eher dazu, in zentralen Bereichen der Großstadt zu siedeln. Die Zone im Übergang wird, betrachtet man nur die Wohnbevölkerung mit deutscher Staatsangehörigkeit, von den mittleren Schichten präferiert. Da der Mietskasernengürtel auch als ein innerstädtischer Teil der Großstadt zu betrachten ist, wundert es nicht, dass auch hier verstärkt hohe und mittlere Statusgruppen anzutreffen sind. Reihen- und Zeilenbauweise sprechen die mittleren und bei den unteren Statusgruppen die einkommensstarken Haushalte an. In der peripheren Einzelhausbebauung sind alle Statusgruppen anzutreffen, soweit sie sich dieses vom Einkommen her leisten können.

Tabelle 4: Verteilung der Befragten aus den alten Bundesländern inklusive West-Berlin über die Wohnquartierstypen, geclustert nach SES. In Zeilenprozent.

SES *Wohnquartierstyp*[x)]	**hoch**	**mittel**	**niedrig**	**niedrig/ Eink. hoch**	**N**
Zentrum erster Ordnung	14	33	33	19	21
Zentrum nachgeordnet	29	29	19	23	140
Zone im Übergang	13	52	15	20	54
Mietskasernen	28	30	23	18	109
Periph. Hochhäuser	10	40	10	40	20
Reihen und Zeilen	18	34	19	29	170
Periph. Einzelhäuser	29	36	9	26	396
N	229	319	137	225	910

[x)] Es werden nur jene Wohnquartierstypen berücksichtigt, die mehr als 10 Fälle aufweisen.
Hoffmeyer-Zlotnik 2004: 96

4.2 Wohnquartierstypische Bewohner nach dem Konzept des Lebenszyklus (LZ)

In die Clusteranalyse wurden als Variablen einbezogen: das Lebensalter der Zielperson; Anzahl der Kinder im Haushalt im Alter unter 18 Jahren; Zielperson im Ruhestand befindlich; Zielperson alleinlebend.

Zur genaueren Beschreibung der LZ-Typen wurden die Cluster kreuztabelliert mit der Bildung und der Autonomie in der beruflichen Tätigkeit der Zielperson in der Definition, wie sie beim SES-Konzept benutzt wurde (siehe oben).

In den alten Bundesländern war die optimale Lösung eine mit fünf Clustern:

LZ 1: *Erste Stufe – niedriges Befragten-Alter, wenige Kinder*: Dieses Cluster zeichnet sich aus durch junge Menschen, zum größten Teil ohne Kind im Haushalt, etwa zur Hälfte in Einpersonenhaushalten lebend; keine Rentner.
Dahinter verbirgt sich an Statusmerkmalen eine mittlere Bildung und eine eher niedrige Autonomie in der beruflichen Tätigkeit.

LZ 2: *Zweite Stufe – niedriges Befragten-Alter, viele Kinder*: Dieses Cluster zeichnet sich aus durch junge Menschen, zum größten Teil mit wenigstens einem Kind pro Haushalt; die Haushalte sind in der Regel Mehrpersonenhaushalte.
Die in diesem Cluster ausgewiesenen Befragten sind unauffällig hinsichtlich Bildung und Autonomie in der beruflichen Tätigkeit.

LZ 3: *Dritte Stufe – mittleres Befragten-Alter*: Die Befragten dieses Clusters gehören den mittleren Altersgruppen an und leben nur zu einem sehr geringen Anteil in Einpersonenhaushalten; etwa die Hälfte von deren Haushalten beherbergt Kinder; der Rentneranteil in diesem Cluster ist gleich null.
Hinter den Mitgliedern dieses Clusters verbergen sich an Statusmerkmalen eine Tendenz zu mittlerer oder höherer Bildung und zu mittlerer bis höherer Autonomie in der beruflichen Tätigkeit.

LZ 4: *Vierte Stufe – hohes Befragten-Alter, dominant sind Hausfrauen*: Dieses Cluster zeichnet sich aus durch ältere Menschen und einen geringen Anteil von Haushalten mit Kind; die Haushalte sind in der Regel Mehrpersonenhaushalte; die sichtbare Gruppe sind die Hausfrauen.
Die in diesem Cluster ausgewiesenen Befragten sind unauffällig hinsichtlich Bildung und Autonomie in der beruflichen Tätigkeit.

LZ 5: *Fünfte Stufe – hohes Befragten-Alter, dominant sind Rentner und Pensionäre*: Dieses Cluster zeichnet sich aus durch alte Menschen, in der Regel im Status von Rentnern und Pensionären; Kinder in den Haushalten sind nicht vorhanden; die Haushalte sind etwa zur Hälfte Mehrpersonenhaushalte.
Die in diesem Cluster ausgewiesenen Befragten weisen eine niedrige Bildung und eine niedrige Autonomie in der (zuletzt ausgeübten) beruflichen Tätigkeit auf.

Tabelle 5: Verteilung der Befragten aus den alten Bundesländern inklusive West-Berlin über die Wohnquartierstypen, geclustert nach LZ. In Zeilenprozent.

LZ (Stufe) Wohnquartierstyp[x)]	erste	zweite	dritte	vierte	fünfte	N
Zentrum erster Ordnung	48	6	30	3	9	33
Zentrum nachgeordnet	33	6	34	8	20	186
Zone im Übergang	35	4	31	6	24	78
Mietskasernen	43	7	30	4	16	139
Periph. Hochhäuser	25	4	42	4	25	24
Reihen und Zeilen	23	6	36	8	27	224
Periph. Einzelhäuser	20	9	35	10	25	569
N	337	95	428	101	292	1253

[x)] Es werden nur jene Wohnquartierstypen berücksichtigt, die mehr als 10 Fälle aufweisen.
Hoffmeyer-Zlotnik 2004: 100

Tabelle 5 zeigt, dass in den alten Bundesländern die unter Lebenszyklus im ersten Cluster anzutreffende Gruppe, die jungen Einpersonenhaushalte, dazu tendiert, in der inneren Stadt zu wohnen. Die im zweiten Cluster anzutreffende Gruppe, die jungen Befragten mit wenigstens einem Kind im Haushalt, verteilt sich ziemlich gleichmäßig über alle Wohnquartierstypen. Die im dritten Cluster anzutreffende Lebenszyklusgruppe, Mehrpersonenhaushalte mit Befragten im mittleren Alter und Kindern, präferiert die aufgelockerten, eher peripheren Quartierstypen. Die im vierten Cluster anzutreffende Gruppe, die dominiert wird von älteren Hausfrauen in Mehrpersonenhaushalten, ist eher in den nachgeordneten Zentren und an der städtischen Peripherie zu finden. Auch die im fünften Cluster anzutreffende Lebenszyklusgruppe, dominiert von Rentnern und Pensionären, ist verstärkt in den zentrumnahen Bereichen der Stadt oder an der städtischen Peripherie anzutreffen. Betrachtet man die fünf Cluster unter Lebenszyklus, so wird eine Tendenz zur Wanderung im Laufe des Lebens von den inneren städtischen Quartierstypen zur städtischen Peripherie sichtbar. Eine Rückwanderung der Alten in die innerstädtischen Quartierstypen wird bei der ersten Generation derer, die das Eigenheim im Grünen gebaut haben, allerdings noch nicht sichtbar.

5 Wohnquartier als Raum lokaler Netzwerke

Beginnend in den 60er Jahren (Pfeil 1965; Gans 1962) und intensiv seit den 80er Jahren (Fischer 1982; Wellman, Carrington & Hall 1988; Hoffmeyer-Zlotnik 1990) beschäftigen sich Stadtsoziologen und Netzwerkforscher damit nachzuweisen, welchen Stellenwert lokale Netzwerke haben und wie diese aussehen.

Im Folgenden soll die Situation in der deutschen Großstadt anhand von zwei Fallstudien diskutiert werden. Die erste Fallstudie wurde 1986 in Mannheim durchgeführt und erhob ego-zentrierte Netzwerke in vier voneinander unterschiedenen, vorab begangenen und abgegrenzten Wohnquartierstypen. Die zweite Fallstudie wurde 2000 in Gießen durchgeführt und kann ein wenig den Zusammenhang zwischen Wohnquartierstyp und Netzwerk erhellen.

5.1 Fallstudie Mannheim

Betrachtet man die Ergebnisse regionaler Netzwerkstudien, so zeigt sich:

- Die Größe eines Netzwerkes steigt mit dem sozialen Status einer Person;
- der Anteil der Verwandten im Netzwerk sinkt mit steigendem sozialen Status.

Die Mannheimer Studie von 1986 hat diesen Befund generell bestätigt, zeigt allerdings, dass man Netzwerke differenzierter betrachten muss: Es ist schichtab-

hängig, wie groß der Anteil der mit ego (der Zielperson) im selben Wohnquartier lebenden Netzpersonen (alteri) ist. Und dieser Anteil variiert von Quartierstyp zu Quartierstyp. Dieses kann überprüft werden, da die Mannheimer Studie unter anderem in vier vom Typ her unterschiedlichen Wohnquartieren jeweils etwa 100 Interviews über ego's Netzwerke durchgeführt hat.

Wohnquartier 1 ist ein altes, gewachsenes Quartier am Stadtrand, mit bester Luftqualität, das in der Selbstwahrnehmung der hier Befragten von der mittleren bis oberen Mittelschicht dominiert wird.

Wohnquartier 2 ist ein in den 70er Jahren neu erbautes Quartier am Innenstadtrand, in der Selbstwahrnehmung der hier Befragten für das ganze Spektrum der Mittelschichten offen.

Wohnquartier 3 ist ein im Konzept der Gartenstadt errichtetes älteres Wohnquartier an der Peripherie, das in der Selbstwahrnehmung der dort Befragten heute von der mittleren bis unteren Mittelschicht dominiert wird.

Wohnquartier 4 ist ein der Industrie benachbartes Arbeiterwohngebiet, das in der Einschätzung der hier Befragten ein Gebiet der unteren Mittelschicht darstellt.

Der Unterschied zwischen den quartierstypischen Netzwerken zeigt sich schon bei der Frage, wie häufig ego, also die befragte Person mindestens eine ihrer wichtigsten Kontaktpersonen im Wohnquartier selbst wohnen hat.

Im Wohnquartier 1, dem Quartier der mittleren bis oberen Mittelschichten, haben 39% der im Quartier lebenden Befragten wenigstens eine Netzperson im Wohnquartier wohnen. Im Wohnquartier 2, dem innerstädtischen Neubauquartier der 70er Jahre, welches das ganze Spektrum der Mittelschichten abdeckt, ist dieses bei 50% der dort lebenden Befragten der Fall. In der peripheren Gartenstadt der mittleren bis unteren Mittelschichten geben 68% der dort lebenden Befragten, und im Arbeiterwohnquartier 100% der dort lebenden Befragten an, dass mindestens eine ihrer zentralen Netzpersonen im Viertel wohne.

Abhängig ist die Einbindung des Netzwerkes ins Quartier von der Stellung im Lebenszyklus und von der Wohndauer im Quartier: Am stärksten verankert in ihrem Quartier sind die Altersgruppen der unter 30-jährigen und der über 60jährigen. Die mittlere Altergruppe der 30- bis 60-jährigen unterscheidet sich, außer im Arbeiterwohngebiet, von den Jungen und Alten erheblich: Im Arbeiterwohngebiet weisen auch die Befragten der mittleren Altersgruppe eine 100%ige Vernetzung im Quartier auf. In den Wohnquartieren der mittleren bis oberen Mittelschichten hat die mittlere Altersgruppe eine deutlich geringere Vernetzung im Quartier. Die zentralen Kontaktpersonen dieser Altersgruppe leben eher in größerer räumlicher Entfernung zum Wohnquartier – diese sind über den Job vernetzt.

5.2 Fallstudie Gießen

Die Gießener Fallstudie basiert nicht mehr auf Interviews in ausgewählten, bestimmten Wohnquartierstypen zugeordneten Stadtteilen. Befragt wurden 620 Bürger der Stadt Gießen im Alter ab 18 Jahre, ausgewählt über eine Zufallsstichprobe. Das konkrete Wohnquartier, in dem interviewt wurde, ist bei dieser Fallstudie unbekannt. Statt dessen liefert die befragte Person eine Beschreibung ihres Wohnquartiers über die drei Variablen: „Entfernung zum Stadtzentrum", „überwiegender Gebäudetyp der Nachbarwohngebäude" und „städtische Nutzung in unmittelbarer Nähe des Wohngebäudes" (siehe oben).

In Gießen ergaben sich 7 unterschiedliche Typen von Wohnquartieren, die sich, sortiert nach der baulichen Dichte, wie folgt beschreiben lassen:

1. der Zentrale Geschäftsbezirk und dessen unmittelbar benachbarte Umgebung;
2. ein Typ hoher baulicher Konzentration, verbunden mit Service und Handel. Hierbei handelt es sich in der Regel um innerstädtische Wohnquartiere;
3. ein Typ, in dem Wohngebäude neben Produktionsstätten liegen;
4. ein Typ, der als Stadtteilzentrum zu bezeichnen ist. Hier sind Wohngebiete mit Läden des Warenangebotes für den täglichen Bedarf durchsetzt;
5. ein Typ mit dichter Bebauung durch Mehrfamilienhäuser;
6. ein Typ mit aufgelockerter Bebauung durch Einfamilienhäuser;
7. der Typ der landwirtschaftlich geprägten Peripherie.

Betrachtet man die Verteilung der Befragten über die Quartierstypen, so ergibt sich folgendes Bild:

- Der Anteil der Befragten, die über 5 Jahre im jetzigen Quartier wohnen, nimmt mit zunehmender baulicher Dichte ab.
- Ein enges Verhältnis zur Nachbarschaft nimmt mit sinkender baulicher Dichte zu.
- Der selbstwahrgenommene soziale Status nimmt vom Zentrum zur Peripherie zu.
- Die Haushaltsgröße nimmt von der Peripherie zum Zentrum kontinuierlich ab.
- Nach dem Alter der Befragten betrachtet, ergibt sich eine U-förmige Verteilung: Die jungen Altersgruppen dominieren die innerstädtischen Wohnquartiere, die Altersgruppe der 45- bis 59-jährigen die peripheren Einfamilienhausgebiete und die Alten die kompakten Mehrfamilienhausgebiete.
- Befragte mit niedriger Bildung sind in den kompakt bebauten innerstädtischen Gebieten (Typen 2 und 5) anzutreffen, Befragte mit Abitur leben im

Zentrum (Gießen ist eine Universitätsstadt) oder an der niedrig bebauten Peripherie der Einfamilienhausgebiete.

Die Netzwerke wurden in der Gießener Studie auf die einfachste Art erhoben: Es wurde nach den drei Personen gefragt, mit denen ego „am besten befreundet" ist. Was unter „befreundet" zu verstehen sein sollte, definierten die Befragten selbst. Eingruppiert wurden die alteri (Netzpersonen) in die Kategorien: gehört (1) zur Kernfamilie, (2) zur erweiterten Familie, (3) zum Kreis der „Bekannten", (4) zum Kreis der Nachbarn, (5) zum Kreis der Kollegen.

Bei genau einem Drittel (33,4%) der Befragten stammt die Mehrheit der genannten Netzpersonen aus der Familie, bei zwei Drittel besteht die Mehrheit der genannten Netzpersonen aus Nicht-Familienmitgliedern.

Betrachtet man die Verteilung der Befragten nach der Zuordnung der erstgenannten Netzperson, so zeigt sich folgendes Bild:

- Der Anteil der Nachbarn als erstgenannte Netzperson steigt mit einer positiven Beurteilung enger nachbarschaftlicher Kontakte im Wohnquartier überhaupt.
- Ist der Nachbarschaftskontakt sehr eng, so ist die Wahrscheinlichkeit hoch, dass ein Nachbar als erste Netzperson genannt wird.
- Je höher der Bildungsabschluss ist, desto geringer wird die Wahrscheinlichkeit, dass die erstgenannte Netzperson zur „Kernfamilie" zählt, aber desto höher wird die Wahrscheinlichkeit, das die erstgenannte Netzperson in den Kreis der „Bekannten" fällt.
- Betrachtet man die erstgenannte Netzperson nach dem Alter der befragten Person, so sind die unter 30jährigen auf „Bekannte" und „Kollegen" fixiert, die 30- bis 45-jährigen präferieren die „Bekannten", die 45- bis 60-jährigen die „Nachbarn" und die „erweiterte Familie", die ab 60-jährigen die „Nachbarn" und die „Kernfamilie".

Tabelle 6 stellt vereinfacht eine Kreuztabellierung dar, bei der die beiden zentralen Variablen „Wohnquartierstyp" und „Netzwerktyp", vorgegeben durch die zuerst genannte Netzperson, zueinander in Beziehung gesetzt werden. „+" bedeutet, dass der Netzwerktyp in einem Quartierstyp überdurchschnittlich oft angetroffen wird, „-" bedeutet, dass der Netzwerktyp im Wohnquartierstyp unterdurchschnittlich oft angetroffen wird. Die Anzahl der Zeichen bezeichnet eine Intensität der Beziehung. „/" bedeutet, dass die Beziehung weder positiv noch negativ ist. Die erste Netzperson wird als Referenz für das Netzwerk gesetzt, da in drei Viertel aller Fälle die Mehrheit der genannten Netzpersonen der erstgenannten entspricht.

Tabelle 6: Verbindung von Netzwerk und Wohnquartier

Wohnquartierstyp		Kernfamilie	erweit. Familie	Bekannte	Nachbarn	Kollegen
1	zentraler Geschäftsbezirk	+	-	-	-	++
2	Service und Handel	-	-	++	/	-
3	Produktion	+	++	+	-	-
4	Stadtteilzentrum	/	+	+	/	-
5	Mehrfamilienhäuser	/	/	-	++	-
6	Einfamilienhäuser	/	+	-	+	+
7	ländliche Peripherie	++	—	—	—	—

Quelle: Hoffmeyer-Zlotnik, eigene Berechnung

Betrachtet man die Verbindung von Netzwerk und Quartierstyp für die zuerst genannte Netzperson, so zeigt sich, dass man jeden Typ Netzwerk in einem bestimmten Quartierstyp verorten kann, dass aber nicht jeder Quartierstyp eindeutig einem Netzwerktyp zuzuordnen ist. Dieses fällt besonders bei jenem Quartierstyp auf, der bei seiner Erstbesiedlung, die in der Regel innerhalb eines kurzen Zeitraumes stattfand, vom Lebenszyklus und sozio-ökonomischem Status gesehen als homogen besiedelt betrachtet wird: den peripheren Einfamilienhäusern. Das Problem mit diesem Quartierstyp ist, dass er bei der Erstbesiedlung in den 60er bis 80er Jahren zwar jeweils einheitlich von jungen Familien besiedelt wurde, dass aber die Besiedlungszeitpunkte einen Zeitraum von über 20 Jahren ausfüllen. Damit verliert dieser Quartierstyp in einer zeitlich punktuellen Querschnittsuntersuchung die ihm eigene Homogenität, vor allem, da die Erstbesiedeler nicht, sobald ihre Kinder das Elternhaus verlassen haben, wieder in die innere Stadt zurückgekehrt sind.

Abschließend wurde mit Hilfe einer Clusteranalyse eine Typisierung von Wohnquartier und Netzwerk hergestellt. Es wurde eine 5-Cluster-Lösung gewählt.

Die Beschreibung der Cluster gibt Muster wieder, die neben dem Wohnquartier sowohl ego als auch dessen Netzwerk beschreiben. Allerdings wurde nicht erfasst, ob die Netzperson in unmittelbarer Nachbarschaft zu ego wohnt.

Cluster 1:

Netzwerk:	geringer Anteil von Familie im Netzkernbereich; intensive Nachbarschaftskontakte;
Lebenszyklus:	die Kinder sind aus dem Haus;
SES:	niedrige bis mittlere Bildung; Autonomie im Job eher niedrig;
Wohnquartier:	eher Innenstadtrand;

Cluster 2:
Netzwerk: hoher Anteil von Familie im Netzkern; erst mit der 3. Netzperson werden Personen außerhalb der Familie gewählt; geringe Nachbarschaftskontakte;
Lebenszyklus: junge Menschen, die noch bei den Eltern leben;
SES: Studenten und Jungakademiker;
Wohnquartier: eher peripher, Einfamilienhausgebiet;

Cluster 3:
Netzwerk: hoher Familienanteil im Netzwerk; geringe Nachbarschaftskontakte;
Lebenszyklus: alte Personen in kleinen Haushalten;
SES: Bildung eher niedrig bei in der Vergangenheit höherer Jobautonomie;
Wohnquartier: eher Innenstadtrand;

Cluster 4:
Netzwerk: geringer Anteil von Familienmitgliedern; geringe Nachbarschaftskontakte;
Lebenszyklus: jung, aber nicht mehr bei den Eltern lebend; noch ohne eigene Kinder;
SES: gebildet, jedoch noch mit geringer Jobautonomie;
Wohnquartier: Innenstadt;

Cluster 5:
Netzwerk: geringer Anteil von Familienmitgliedern; geringe Nachbarschaftskontakte;
Lebenszyklus: Personen in der Phase: „junge Familie mit Kind“;
SES: Bildung eher hoch; Autonomie im Job im Befragtendurchschnitt eher mittel;
Wohnquartier: Innenstadtrand, kompakte Bebauung und peripher.

Die fünf Gießener Cluster zeigen, dass es eine Beziehung zwischen Typ des Netzwerks, Position im Lebenszyklus, sozio-ökonomischem Status (u.a. Hoffmeyer-Zlotnik 1990) und dem Wohnquartierstyp gibt. Allerdings ist Gießen eine sehr stark von der studentischen Population geprägte Mittelstadt, womit einige Muster ein wenig gegen die landläufige Meinung der Wohnpräferenz verlaufen, z.B. Cluster 2.

6 Schlussbemerkungen

Sozial-räumliche Differenzierung und statusadäquates Siedeln führt zu gruppendominant besiedelten Wohnquartieren. Diese Wohnquartiere lassen sich mit

wenigen Variablen beschreiben. Im Minimum sind drei Variablen hierzu erforderlich: die Lage innerhalb der Stadt, die Bebauungsdichte und der Nutzungsmix eines Quartiers. Sollen zusätzlich ethnische Kolonien identifiziert werden, muss mit einer vierten Variable der ethnische Mix erfasst werden. Diese drei bis vier Variablen zur Identifizierung von Nonrespondenten lassen sich über ein Kontaktprotokoll durch die Interviewer mit wenig Aufwand zusätzlich erfassen. Bedingung ist allerdings, dass eine Person vor Ort die Beschreibung des Wohnquartiers durchführt. Bei nicht face-to-face-Verfahren muss entweder eine Begehung vor Ort stattfinden oder die entsprechenden Informationen müssen in einem Geographischen Informationssystem gespeichert sein. Personen mit aufeinanderfolgenden Telefonnummern sind in der Regel keine Nachbarn und können damit nicht stellvertretend das Quartier eines Nonrespondenten beschreiben.

Gelingt eine Beschreibung des Wohnquartiers, so ist der Rückschluss auf die in diesem Quartier dominante Gruppe möglich. Allerdings muss beachtet werden, dass der Nonrespondent nicht zwangsläufig zur dominanten Gruppe gehören muss. Je kleinräumiger man ein städtisches Teilgebiet betrachtet, z.B. gebäudeweise, desto heterogener werden sich dessen Bewohner darstellen (vgl. Zapf 1969). Dennoch stellt das Wohnquartier einen zentralen Teil des menschlichen Aktionsraumes dar. Man unterliegt den Einflüssen aus dem Quartier und akzeptiert diese generell, auch wenn man nicht zur dominanten Gruppe gehört. Andernfalls würde man das Quartier verlassen.

Bedingung hierfür ist allerdings, dass der Wohnungsmarkt dem Individuum Wahlmöglichkeiten bietet, dass die zu charakterisierende Person aus eigener Entscheidung in diesem Quartier lebt. Wie gezeigt wurde, ist diese Wahl des Wohnquartiers nicht immer frei, da sie beeinflusst wird über Angebot, Nachfrage und Segmentierung des Wohnungsmarktes, worin durchaus Zugangsbeschränkungen für ethnische und soziale Gruppen beinhaltet sein können. Dennoch muss der Einzelne die Möglichkeit haben, seinen Bedürfnissen und seinem Status adäquat seine Wohnung wählen zu können, sofern die Bedürfnisse nicht den Status überschreiten. Bei Wohnungszwangsbewirtschaftung ist eine Charakterisierung von Bewohnern über regionale Merkmale nicht möglich.

Eine Wohnquartiersbeschreibung und -typisierung allein beschreibt allerdings noch keinen Nonrespondenten, auch wenn dieser zur dominanten Gruppe gehören sollte. Hierzu sind Zusatzinformationen notwendig. Zu beachten ist, dass ein Wohnquartierstyp Rückschlüsse auf demographische Merkmale nur in einem verschwommenen Bild ermöglicht. Der Wohnquartierstyp ermöglicht, befördert und erlaubt Rückschlüsse auf Merkmale von Lebensstilen. Hinter diesen Lebensstiltypen stehen bestimmte Personengruppen oder Haushaltstypen in demographischer und/oder sozialer Zusammensetzung. Daher muss die Nutzung von regionalen Merkmalen zur Charakterisierung von Personen einhergehen mit der Kenntnis von gruppenspezifischen Wohnpräferenzen und den dazu-

gehörigen Lebensstilen. Nützlich ist noch weiteres Zusatzwissen über den Stadttyp, wenn möglich sogar über die konkrete Stadt. So ist, um bei den genannten Beispielen zu bleiben, eine Stadt wie Mannheim anders strukturiert als eine Stadt wie Gießen. Die eine wird dominiert von Industrie, Handel und Wissenschaft, die andere von Universität und Verwaltung. Damit unterscheidet sich sowohl die Bewohner- als auch die Nutzungsstruktur beider genannter Städte.

Dennoch bleibt der zentrale Punkt eine Typisierung von Quartierstypen, die einen bestimmten Lebensstil befördern oder Personen mit einem bestimmten Lebensstil anziehen. Trotz großer Datenmengen in sehr kleinräumiger Zuordnung, teils bis auf 10 Adressen herunter, verbleiben auch die Anbieter der Marktforschung, die innerstädtische Milieus mit raumbezogenen Konzepten von Zielgruppenanalysen z.B. für die Werbung verfolgen, bei einer national anwendbaren Typisierung von Milieus oder Nachbarschaften/Wohnquartieren. Und die Marktforschung hat mit einer national anwendbaren kleinräumigen Regionalisierung zur Zielgruppenanalyse Erfolg. Die Sozialforschung sollte sich diesem Weg anschließen.

Betrachtet man die Literatur, so zeichnet sich allmählich ab, dass eine kleinräumige Regionalisierung nicht nur für eine Zielgruppenanalyse in der Marktforschung und im Marketing und nicht nur zur besseren Interpretation von Befragtenaussagen in der Sozialforschung sinnvoll ist, sondern auch für die Charakterisierung von Nonrespondenten. Seit den ersten deutschen Versuchen, über eine Wohnquartiersbeschreibung Nonrespondenten zu charakterisieren (Hoffmeyer-Zlotnik 1981), sind fast 25 Jahre vergangen. In der Zwischenzeit hat nicht nur der Unit-Nonresponse in Surveys drastisch zugenommen, sondern es werden international Wege gesucht, die Nonrespondenten über eine Regionalisierung zu charakterisieren. Dieses wird in den USA allerdings mit Census-Daten (z.B. Zanutto & Zaslavsky 2002; Korinek, Mistiaen & Ravallion 2004); als auch in Europa, hier vor allem bei der amtlichen Statistik (z.B. György 2004), angewandt.

Literatur

Blasius, J., 1993: Gentrification und Lebensstile. Wiesbaden: Deutscher Universitäts-Verlag.

Boustedt, O., 1966: Stadtregionen; in: ARL (Hrsg.): Handwörterbuch der Raumforschung und Raumordnung. Hannover: Gebr. Jänecke: 1916-1932.

Burgess, E.W., 1925: The Growth of the City: An Introduction to a Research Project; in: Park, R.E., Burgess, E.W. & R.D. McKenzie: The City. Suggestions for Investigation of Human Behavior in the Urban Environment. Chicago, London: The University of Chicago Press. Reprint 1967: 47-62.

Burgess, E.W., 1929: Urban Areas; in: Smith, T.V. & L.D. White (Hrsg.): Chicago, an Experiment in Social Science Research. Chicago: The University of Chicago Press: 113-138.

Carter, H., 1980: Einführung in die Stadtgeographie. Übersetzt und herausgegeben von F. Vetter. Berlin, Stuttgart: Gebr. Borntraeger.

Dangschat, J.S., 1994: Lebensstile in der Stadt. Raumbezug und konkreter Ort von Lebensstilen und Lebensstilisierungen; in: Dangschat, J.S. & J. Blasius (Hrsg.): Lebensstile in den Städten. Konzepte und Methoden. Opladen: Leske + Budrich: 335-354.

Dangschat, J.S., 1997: Armut und sozialräumliche Ausgrenzung in den Städten der Bundesrepublik Deutschland; in: Friedrichs, J. (Hrsg.): Die Städte in den 90er Jahren. Demographische, ökonomische und soziale Entwicklungen. Opladen, Wiesbaden: Westdeutscher Verlag: 167-212.

Demografische Standards 1999. www.geis.org/Methodenberatung/Untersuchungsplanung/Standarddemografie/dem_standards/demsta99.pdf

Denton, N.A. & D.S. Massey, 1988: Residential Segregation of Blacks, Hispanics, and Asians by Socioeconomic Status and Generation; in: Social Science Quarterly 69: 797-817.

Eekhoff, J., 1987: Wohnungs- und Bodenmarkt. Tübingen: Mohr.

Falk, W., 1998: Wohnen im Lebenslauf. Die Wirkungen der deutschen Wohnungspolitik. Amsterdam: G+B Verlag Fakultas.

Fischer, C.S., 1982: To Dwell Among Friends. Personal Networks in Town and City. Chicago: The University of Chicago Press.

Friedrichs, J., 1977: Stadtanalyse. Soziale und räumliche Organisation der Gesellschaft. Reinbek: Rowohlt.

Friedrichs, J., 1995: Stadtsoziologie. Opladen: Leske + Budrich.

Friedrichs, J., 1998: Ethnic Segregation in Cologne, Germany, 1984-94; in: Urban Studies 35/1998: 1745-1763.

Friedrichs, J. & R. Kecskes (Hrsg.), 1996: Gentrification. Theorie und Forschungsergebnisse. Opladen: Leske + Budrich.

Gans, H.J., 1961: Planning and Social Life; in: Journal of the American Institute of Planners 27.

Gans, H., 1962: The Urban Villagers. New York: Free Press.

Glasauer, H., 1986: Sozialpolitische Hoffnungen und die Logik des Marktes. Die Relevanz des Filtering-Modells für den städtischen Wohnungsmarkt. Arbeitsbericht Fachbereich Stadt- und Landschaftsplanung. Heft 70. Kassel: Gesamthochschule.

György, E., 2004: Analysing Unit Nonresponse Using Matched Census-Survey Records. Experiences from the Hungarian Labour Force Survey. Paper presented at 15th International Workshop on Household Survey Nonresponse. Maastricht/NL, 23.-25. August 2004.

Häußermann, H., 1996: Von der Stadt im Sozialismus zur Stadt im Kapitalismus; in: Häußermann, H. & R. Neef (Hrsg.): Stadtentwicklung in Ostdeutschland. Soziale und räumliche Tendenzen. Opladen: Westdeutscher Verlag: 5-47.

Hamm, B., 1982: Einführung in die Siedlungssoziologie. München: C.H. Beck.

Hard, G., 1973: Die Geographie. Eine wissenschaftstheoretische Einführung. Berlin, New York: de Gruyter.

Herlyn, I. & U. Herlyn, 1976: Wohnverhältnisse in der BRD. Frankfurt/Main, New York: Campus.

Hoffmeyer-Zlotnik, J., 1977: Gastarbeiter im Sanierungsgebiet. Das Beispiel Berlin-Kreuzberg. Hamburg: Christians.

Hoffmeyer-Zlotnik, J., 1981: Wohnquartiersbeschreibung als Mittel zur Messung soziologischer Merkmale von Ausfällen; in: ZUMA-Nachrichten 8: 5-24.

Hoffmeyer-Zlotnik, J. H.P., 1982: Community Change and Invasion: The Case of Turkish Guest Workers; in: Friedrichs, J. (Hrsg.): Spatial Disparities and Social Behavior. A Reader in Urban Research. Hamburg: Christians: 114-126.

Hoffmeyer-Zlotnik, J. H.P., 1984: Zur Beschreibung von Wohnquartieren - Die Entwicklung eines Instruments. ZUMA-Arbeitsbericht 84/05.

Hoffmeyer-Zlotnik, J. H.P., 1986: Eingliederung ethnischer Minoritäten - unmöglich?; in: Hoffmeyer-Zlotnik, J. H.P. (Hrsg.): Segregation und Integration. Die Situation von Arbeitsmigranten im Aufnahmeland. Mannheim: Forschung Raum und Gesellschaft: 15-55.

Hoffmeyer-Zlotnik, J. H.P., 1990: The Mannheim Comparative Network Research; in: Weesie, J. & H. Flap (eds.), 1990: Social Networks Through Time. Utrecht: ISOR: 265-279.

Hoffmeyer-Zlotnik, J. H.P., 1995: Welcher Typ Stadtbewohner dominiert welchen Typ Wohnquartier? Merkmale des Wohnquartiers als Hintergrundmerkmale zur Regionalisierung von Umfragen; in: ZUMA-Nachrichten 37: 35-62.

Hoffmeyer-Zlotnik, J. H.P., 1998: „Beruf“ und „Stellung im Beruf“ als Indikatoren für soziale Schichtung; in: Ahrens, W., Bellach, B.-M. & K.-H. Jöckel (Hrsg.): Messung soziodemographischer Merkmale in der Epidemiologie. RKI-Schriften 1/98: 54-64.

Hoffmeyer-Zlotnik, J. H.P., 2000: Regionalisierung sozialwissenschaftlicher Umfragedaten. Siedlungsstruktur und Wohnquartier. Wiesbaden: Westdeutscher Verlag.

Hoffmeyer-Zlotnik, J. H.P., 2004: Wohnquartiersbeschreibung: Ein Instrument zum Erfassen von Nachbarschaften; in: Kecskes, R., M. Wagner & C. Wolf (Hrsg.): Angewandte Soziologie. Wiesbaden: VS Verlag für Sozialwissenschaften: 77-102.

Ipsen, D., 1980: Wohnungsteilmärkte. Kassel: Gesamthochschule.

Ipsen, D., 1984: Die Auswirkungen des sozialen Wohnungsbaus auf den örtlichen Wohnungsmarkt. Eine Untersuchung von Umzugsketten. Arbeitsbericht des Fachbereichs Stadt- und Landschaftsplanung. Heft 48. Kassel: Gesamthochschule.

Ipsen, D., Glasauer, H. & W. Heinzel, 1981: Teilmärkte und Wirtschaftsverhalten privater Miethausbesitzer. Arbeitsbericht des Fachbereichs Stadt- und Landschaftsplanung. Heft 9. Kassel: Gesamthochschule.

Kecskes, R., 1997: Sozialräumlicher Wandel in westdeutschen Großstädten. Ursachen, Folgen, Maßnahmen; in: Friedrichs, J. (Hrsg.): Die Städte in den 90er Jahren. Demographische, ökonomische und soziale Entwicklungen. Opladen, Wiesbaden: Westdeutscher Verlag: 213-244.

Knauss, E., 1981: Räumliche Strukturen als Bedingungen der Bevölkerungsverteilung. Hamburg: Christians.

Korinek, A., Mistiaen, J.A. & M. Ravallion, 2004: Survey Nonresponse and the Distribution of Income. Working Paper. World Bank.

Kreibich, V., 1985: Wohnversorgung und Wohnstandortverhalten; in: Friedrichs, J. (Hrsg.): Die Städte in den 80er Jahren. Demographische, ökonomische und technologische Entwicklungen. Opladen: Westdeutscher Verlag: 181-195.

Leitner, H., 1983: Gastarbeiter in der städtischen Gesellschaft. Segregation, Integration und Assimilation von Arbeitsmigranten. Am Beispiel jugoslawischer Gastarbeiter in Wien. Frankfurt/Main, New York: Campus.

Lichtenberger, E., 1998: Stadtgeographie. Band 1. Begriffe, Konzepte, Modelle, Prozesse. Teubner Studienbücher Geographie. 3., neubearb. und erw. Aufl. Stuttgart, Leipzig: B.G. Teubner.

Pfeil, E., 1965: Die Familie im Gefüge der Großstadt. Zur Sozialtopographie der Stadt. Hamburg: Christians.

Ratcliff, R.U., 1949: Urban Land Economics. New York, Toronto, London: McGraw-Hill.

Regionale Standards 2005. www.gesis.org/Methodenberatung/ Untersuchungsplanung/Regionalisierung/reg_standards/regsta05.pdf

Schäfers, B., 1981: Sozialstruktur und Wandel der Bundesrepublik Deutschland. Ein Studienbuch zu ihrer Soziologie und Sozialgeschichte. 3. überarb. und erw. Aufl. Stuttgart: Enke.

Schelsky, H., 1961: Die Bedeutung des Klassenbegriffs für die Analyse unserer Gesellschaft; in: Schelsky, H., 1965: Auf der Suche nach der Wirklichkeit. Gesammelte Aufsätze. Düsseldorf, Köln: 352-388. Neuabdruck in: Seidel, B. & S. Jenker (Hrsg.), 1968: Klassenbildung und Sozialschichtung. Darmstadt: Wissenschaftliche Buchgesellschaft: 398-446.

Schinz, A., 1964: Berlin. Stadtschicksal und Städtebau. Braunschweig: Georg Westermann.

Shevky, E. & W. Bell, 1955: Social Area Analysis. Theory, Illustrative Application and Computational Procedures. Stanford: Stanford University Press.

Simon, M., 1990: Das Ring-Sektoren-Modell: Ein Erfassungsinstrument für demografische und sozio-ökonomische Merkmale und Pendlerbewegungen in gleichartig definierten Stadt-Umland-Gebieten. Grundlagen, Methodik, Empirie. Geographica Bernensia G 36. Bern: Geografisches Institut der Universität Bern.

Wellman, B., Carrington, P.J. & A. Hall, 1988: Networks as Personal Communities; in: Wellman, B & S.D. Berkowitz (eds.), 1988: Social Structures. A Network Approach. Cambridge: Cambridge University Press: 130-184.

Zanutto, E. & A. Zaslavsky, 2002: Using Administrative Records to Improve Small Area Estimation: An Example from the U.S. Decennial Census; in: Journal of Official Statistics, Vol.18, No. 4, 2002: 559-576.

Zapf, K., 1969: Rückständige Viertel. Frankfurt/Main: Europäische Verlagsanstalt.

Zinn, H., 1978: Sozialräumliche Segregation der Bevölkerung; in: Mühlich, E., Zinn, H., Kröning, W. & I. Mühlich-Klinger: Zusammenhang von gebauter Umwelt und sozialem Verhalten im Wohn- und Wohnumweltbereich. Schriftenreihe „Städtebauliche Forschung“ des BMBau, Heft 03.062/1978. Bonn: BMBau: 105-112.

Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen

Jürgen Krause, Maximilian Stempfhuber

1 Bisherige Vorschläge auf den Wiesbaden-Workshops

Die Thematik Integration sozialwissenschaftlicher Informationsangebote stand bereits zwei Mal auf der Agenda der Wiesbadener Workshops[1]. Im Kontext der Umsetzung der KVI–Projekte (s. KVI 2001) wurde sie in die Diskussion um Standards für Metadaten und die informationstechnologischen Probleme der Dokumentation empirischer Daten eingebracht (DDI-Format als aussichtsreichster Kandidat). Als unbefriedigend wurde diagnostiziert, dass die BMBF–Projekte, die zu Forschungsdatenzentren bei den statistischen Ämtern und Servicezentren im wissenschaftlichen Umfeld führten, die Heterogenität des Angebots eher verstärkt haben. Die Frage sowohl nach der technologischen als auch der konzeptuellen Integration habe keine Fortschritte gezeigt.

Entscheidend bleibt bis heute, dass die Ursachen dieser Defizite sehr prinzipieller Art sind und die gesamte Fachinformation betreffen, gleich ob Daten oder textuelle Dokumente die Basis des wissenschaftlichen Informationsservice bilden.

1.1 Heterogenität und Publizieren im Web

Es wurde analysiert, dass die alten Rezepte der wissenschaftlichen Fachinformation nicht mehr greifen und in vielen Teilbereichen zu wenig vertrauenserweckenden Ergebnissen führen, die weder ihre fachwissenschaftlichen Kunden befriedigen noch adäquate Modelle für zukünftige Entwicklungen bereitstellen. Generell ist einer der Tragpfeiler bisherigen Handelns, die Standardisierung als Basis der Interoperabilität von Informationsangeboten im Web differenzierter zu sehen, was zu dem Leitsatz führte: „Die Standardisierung ist von der verbleibenden Heterogenität her zu denken“. Was ist damit gemeint?

1 Jürgen Krause, Probleme der Integration und Heterogenität bei der Recherche textueller Dokumente. vascoda - infoconnex – SOWIPORT. 2. Konferenz für Sozial- und Wirtschaftsdaten, Wiesbaden Juni 2004; Gemeinsamkeiten und Grundlagen der KVI-Projekte: Verbleibende Möglichkeiten einer projektübergreifenden Zusammenarbeit auf der Basis des „Publizierens im Web“, Workshop „Metadaten II“, Statistisches Bundesamt, Wiesbaden, 12. Dezember 2003, erster KVI-Workshop; zu den Inhalten siehe Krause 2004c).

Will man Literaturinformationen – und später Fakteninformationen und multimediale Daten – aus verteilten und inhaltlich unterschiedlich erschlossenen Datenbeständen, die sich in miteinander nicht verbundenen, heterogenen Organisationsstrukturen und Zugänglichkeitskontexten befinden (Sondersammelgebiete der Universitätsbibliotheken, wissenschaftliche Spezialbibliotheken, Referenzdatenbanken, digitale Volltexte, Datenarchive), mit einer Anfrage integriert recherchieren, müssen die Probleme des inhaltlichen Zugriffs auf verteilte Dokumentenbestände gelöst werden. Ein vom Benutzer gewähltes Stichwort (Deskriptor) X kann in den verschiedenen Dokumentenbeständen die unterschiedlichsten Bedeutungen annehmen. Auch im engen Bereich der Fachinformation kann ein Term X, der aus einem hochrelevanten, mit viel Aufwand qualitativ hochwertig ermittelten Dokumentenbestand stammt, nicht mit dem Term X gleichgesetzt werden, den z. B. eine automatische Indexierung auf der Basis von Titeln aus einem Randgebiet liefert. Deshalb genügt eine rein technologische Verknüpfung verschiedener Dokumentenbestände und die formale Integration unter einer Benutzungsoberfläche allein nicht. Sie führt zum fehlenden Nachweis relevanter Dokumente und zu einer Fülle von irrelevanten Treffern.

Diese Feststellung wurde im Hinblick auf die konstatierte mangelnde Abgestimmtheit der KVI-Projekte mit dem Web-Paradigma des „Publizierens im Netz“ verbunden. Dessen Quintessenz liegt in der Grundforderung an jede Web-Aktivität: Jedes neue Angebot „... is designed to fit into a wider data input and output environment“ (Musgrave 2003: 5). Frühere Systementwicklungen mussten sich „nur“ darum kümmern, dass ihr System für sich genommen effizient und schnell Anfragen zuließ und die Benutzerbedürfnisse umsetzte. Heute genügt dies nicht. Niemand arbeitet mehr isoliert für sich und seine Benutzergruppe. Jeder ist Teil eines globalen Angebots und erfüllt in diesem fachwissenschaftlichen Informationskontext nur eine kleine, spezielle Aufgabe. Die Benutzer einer speziellen Datenbank werden sich nicht auf dieses eine Angebot beschränken, sondern auf viele vergleichbare Bestände integriert zugreifen wollen. Einige dieser Cluster sind bei Beginn der Entwicklung eines neuen Angebots schon bekannt. Wichtiger ist jedoch, dass mit Sicherheit in den nächsten Jahren nach Fertigstellung des eigenen Angebots neue Informationssammlungen im Web hinzukommen werden, auf die der Benutzer integriert zugreifen will.

Und weil man dies weiß, liegt die eigentliche Schwierigkeit nicht in der konkreten Systemprogrammierung, sondern in der Modellierung des Systems als passfähige Teileinheit, ein Denken, für das bei den KVI-Aktivitäten Handlungsbedarf gesehen wurde.

Im Folgenden geht es um einen Ausschnitt aus dieser Gesamtproblematik, um die konzeptuellen Probleme der von vielen Benutzern gewünschten integrierten Recherche in Umfragedaten (Beispiel Kinderpanel des DJI) und textuellen Daten (Beispiel die Literaturdatenbanken des Informationszentrum So-

zialwissencshaften (IZ): SOLIS, FORIS und infoconnex im Rahmen von vascoda).[2]

1.2 Genereller Ansatz Text-Fakten-Integration

Trotz der heterogenen Ausgangslage auf der Angebotsseite soll der Benutzer z.B. nicht gezwungen werden, sich zuerst in das Erschließungssystem einer Bibliothek einzuarbeiten, um dann bei einer Erweiterung seiner Suchintention auf unselbständige Publikationen ein zweites System der Inhaltserschließung lernen und in geeignete Suchstrategien umsetzen zu müssen, und wiederum ein anderes, wenn er weitere fachwissenschaftliche Datenbanken, auch solche zur Erschließung von Umfragedaten, ergänzend durchsuchen möchte.

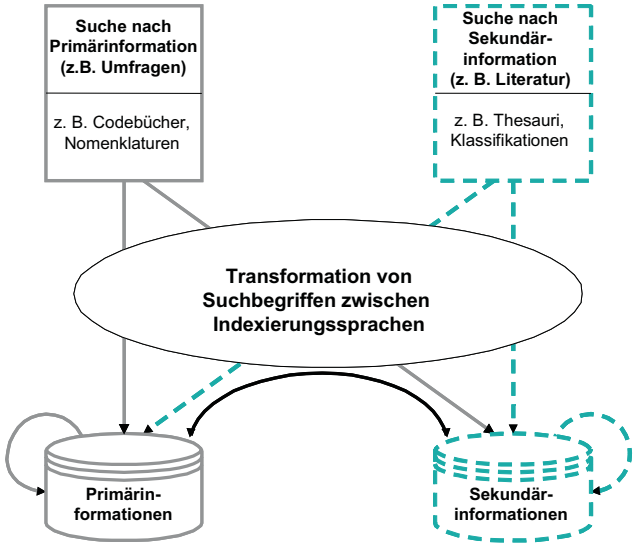


Abb. 1: Generelles Integrationsmodell für die Suche nach Primär- und Sekundärinformationen

Deshalb muss die bestehende Heterogenität der verschiedenen Inhaltserschließungssysteme durch geeignete Maßnahmen aufeinander bezogen werden. Dabei stellt die Integration von Fachdatenbanken und Bibliotheksbeständen nur eine erste Stufe dar. Sie ist durch Internetquellen und Faktendaten (z. B. Zeitrei-

2 http://www.vascoda.de/; http://www.infoconnex.de/; http://www.gesis.org/Information/SOLIS/ und http://www.gesis.org/Information/FORIS/ unter http://www.gesis.org/

hen zu Umfragen) zu ergänzen, generell um alle Datentypen, die wir heute bei virtuellen Fachbibliotheken auf den unterschiedlichsten Fachportalen oder elektronischen Marktplätzen finden.

2 Fortschritte Heterogenitätsbehandlung und Integration textueller Dokumente

Im Gegensatz zum eingangs genannten Informationsausschnitt der KVI-Projekte hat sich bei den textuellen Dokumenten in Deutschland in den letzten Jahren viel getan. Entscheidender Motor der Entwicklung ist das Wissenschaftsportal vascoda,[3] bei dem derzeit etwa 45 Informationsanbieter über alle Fächer hinweg vertreten sind. In diesem Kontext setzte sich auch das im Folgenden kurz skizzierte Modell zur Heterogenitätsbehandlung durch, das auch die Grundlage für das im Aufbau befindliche Fachportal Sozialwissenschaften ist (s. Krause/Schmiede 2004). Hilfestellung bei der Realisierung leistet dabei das Kompetenzzentrum Modellbildung und Heterogenität des IZ[4], das von allen Partnern von vascoda und Projekten der eScience-Initiative in Anspruch genommen werden kann.

2.1 Crosskonkordanzen, statistisch–quantitative und deduktive Verfahren

Das Modell der Heterogenitätsbehandlung in einer Welt polyzentrischer Informationsversorgung stellt einen allgemeinen Rahmen dar, in dem sich bestimmte Klassen von Dokumenten mit unterschiedlicher Inhaltserschließung analysieren und algorithmisch aufeinander beziehen lassen. Zentral sind intelligente Transferkomponenten zwischen den verschiedenen Formen der Inhaltserschließung, die den semantisch-pragmatischen Differenzen Rechnung tragen. Sie interpretieren die technische Integration zwischen den einzelnen Datenbeständen mit unterschiedlichen Inhaltserschließungssystemen zusätzlich konzeptuell. Die Begriffswelt der fachspezifischen und generellen Thesauri, Klassifikationen, eventuell auch thematische Begriffsfelder und Abfragestrukturen begrifflicher Datensysteme usw. sind aufeinander zu beziehen. Deshalb sind Transfermodule zwischen jeweils zwei Informationsbeständen unterschiedlichen Typs zu entwickeln, die den Übergang nicht nur technisch, sondern konzeptuell gestalten (s. Krause/Niggemann/Schwänzl 2003 und Krause 2004b; genaueres zum „bilateralen“ Transfer in Krause 2004a).

3 http://www.vascoda.de/; Pianos 2005.

4 http://www.gesis.org/Forschung/Informationstechnologie/KoMoHe.htm s. auch Mayr/Stempfhuber/Walter 2005.

Generell gibt es drei Verfahrensweisen, die in Bezug auf ihre Wirksamkeit im Einzelfall zu überprüfen und zu implementieren sind. Keines der Verfahren trägt die Last des Transfers allein. Sie sind ineinander verschränkt und wirken zusammen.

- Crosskonkordanzen zu Klassifikationen und Thesauri
  Die verschiedenen Begriffssysteme werden im Anwendungskontext analysiert und der Versuch gemacht, ihre Begrifflichkeit intellektuell aufeinander zu beziehen. Das Konzept darf nicht mit dem der Metathesauri verwechselt werden. Es wird keine neue Standardisierung bestehender Begriffswelten angestrebt. Crosskonkordanzen enthalten nur die partielle Verbindung zwischen bestehenden Terminologiesystemen, deren Vorarbeit genutzt wird. Sie decken damit den statisch bleibenden Teil der Transferproblematik ab.
  Bei der Recherche bieten solche Verzeichnisse die Möglichkeit, Terme des einen Begriffssystems auf die des anderen auszuweiten, im einfachsten Fall im Sinne einer Synonymie- oder Ähnlichkeitsrelation, aber auch als deduktive Regelbeziehung.
  Im Rahmen von vascoda und infoconnex wurden bisher fünf Crosskonkordanzen entwickelt.
- Quantitativ-statistische Ansätze
  Das Transferproblem lässt sich allgemein als Vagheitsproblem zwischen zwei Inhaltsbeschreibungssprachen modellieren. Für die im Information Retrieval behandelte Vagheit zwischen den Termen der Benutzeranfrage und denen des Datenbestandes sind verschiedene Verfahren vorgeschlagen worden (probabilistische Verfahren, Fuzzy-Ansätze, rough set theory und neuronale Netze (Mandl 2001, Zhang 2005), die sich auf die Transferproblematik anwenden lassen.
- Qualitativ-deduktive Verfahren
  Hierbei geht es darum, deduktive Zusammenhänge offen zu legen, die mit Techniken aus dem Bereich der Expertensysteme zu behandeln sind. Sie wurden bisher im Rahmen von vascoda und infoconnex noch nicht realisiert.

2.2 vascoda – infoconnex, ViBSoz und sowiport

An dem übergreifenden Ziel eines integrierten Fachportals bezogen auf die Sozialwissenschaften, eingebettet in ein allgemeines Wissenschaftsportal arbeiten in den letzten Jahren eine Reihe von Forschungs- und Entwicklungsstellen (siehe Krause/Schmiede 2004 und Pianos 2005). Sie alle befassen sich mit verschiedenen Teilaspekten, die in Zukunft integriert und aufeinander abgestimmt obige Zielsetzung erreichen sollen.

Thematisiert das sozialwissenschaftliche Fachportal sowiport[5] die intelligente und qualitativ hochwertige gemeinsame Recherche in polyzentrisch an verschiedenen Orten angebotenen Informationssammlungen, so überschritt infoconnex bereits diese Grenze. Gerade für die Sozialwissenschaften sind einerseits Fachgrenzen nur schwer zu ziehen, da es viele Überschneidungsbereiche mit anderen Wissenschaften gibt, andererseits hat interdisziplinäre Forschung, die über diese Überschneidungsbereiche hinausgeht, einen besonderen Stellenwert. Deshalb ist der nächste notwendige Schritt die Clusterbildung mit eng benachbarten Fächern. Im Informationsverbund infoconnex werden seit Mitte 2003 bereits die Datenbanken in den Bereichen Sozialwissenschaften, Pädagogik und Psychologie auf der Basis von Crosskonkordanzen integriert angeboten. Hinzukommen sollen die Wirtschaftswissenschaften und eine sukzessive Ausdehnung auf alle im fachwissenschaftlichen Portal integrierten Dokumententypen, wo immer dies sinnvoll erscheint.

Das Wissenschaftsportal vascoda thematisiert den Schritt vom Cluster zur frei durch den Benutzer definierbaren Kombination der Suchgrundlage über alle Fächer hinweg (beliebige interdisziplinäre Recherche). vascoda ist fachbezogen organisiert und bietet ausschließlich Dokumente an, „deren wissenschaftlicher Wert verifiziert ist". Derzeit ist der Zugriff auf Fachdatenbanken (einschließlich Ausgabe des Volltextes, falls vorhanden) und auf die Fachinformationsführer der virtuellen Fachbibliotheken möglich, allerdings - technisch bedingt - noch ohne Heterogenitätskomponenten. In Kürze werden mehrere zusätzliche Fächer (Sportwissenschaft, Medizin usw.) Crosskonkordanzen einsetzen und ihre Fachbestände untereinander und über die SWD[6] mit den Bibliotheks-Opacs verbinden. Sobald vascoda die notwendige Technologie einer Clusterbildung im Sinne von infoconnex bereitstellt, wird dieser Mehrwertdienst durch die entsprechenden vascoda-Dienste ersetzt werden.

2.3 Empirische Ergebnisse Transferkomponenten

Mittlerweile liegen erste Ergebnisse zur Evaluation der semantischen Transformationen vor (Marx 2005, Mayr/Stempfhuber/Walter 2005). Gemessen wurden u.a. der Term–Recall und die Term–Precision auf der Basis von zwei Parallelcorpora (USB = 16.914 Datensätze, SWD = 33.328).

5 Die virtuelle Fachbibliothek ViBSoz kann als Vorstufe zu sowiport angesehen werden (s. http://www.vibsoz.de und Müller 2003).

6 SWD ist die von den deutschen wissenschaftlichen Universalbibliotheken kooperativ aufgebaute Schlagwortnormdatei auf der Basis des Regelwerks RSWK „Regeln für den Schlagwortkatalog".

Tab. 1: Verbesserung durch intellektuell erstellte Crosskonkordanzen (s. Marx 2005, S. 39 und 49)

Verbesserung	Term Recall	Term Precision
IZ → SWD	USB: + 45,1%	USB: + 30,2%
Crosskonkordanz, intellektuell	DDB: + 47,4%	DDB: + 28,6%
SWD → IZ	USB: + 41,0%	USB: + 44,2%
Crosskonkordanz, intellektuell	DDB: + 44,5%	DDB: + 45,9%

Auch die Ergebnisse der statistischen Komponenten (s. Marx 2005) und der rough set-basierte Transfer (s. Zhang 2005) zeigen positive Ergebnisse, so dass von einer generellen Wirksamkeit des vorgeschlagenen Modells ausgegangen werden kann. Der nächste Schritt wird eine Kombination der statistischen mit den auf den intellektuell erstellten Crosskonkordanzen sein und die Ausweitung der Evaluation auf Benutzertest.

3 Heterogenitätskomponente Umfragedaten – textuelle Dokumente

Aus Sicht der Entwicklung von vascoda und SOWIPORT ist die Erweiterung um Faktenangebote nach dem in Abschnitt 2.1 skizzierten Modell (s. Abb. 2) zwingend und selbstverständlich.

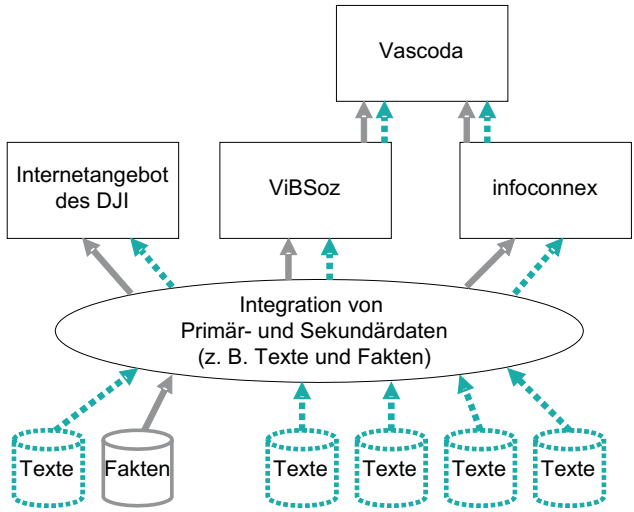


Abb. 2: Integration von Texten und Fakten in Fachinformationsangeboten

Deutlich wurde aber auch, dass es - neben klaren Befürwortern eines solchen Modells - eine Gruppe von Sozialwissenschaftlern gibt, die Projekten zur Ausweitung der Transferkomponenten auf Umfragedaten einen deutlichen Widerstand entgegensetzen. Ein Komplex der in diesem Kontext genannten Gegenargumente bezieht sich auf die semantische Tiefe der Analyse, die den Heterogenitätskomponenten aus dem Bereich textueller Dokumente zugrunde liegt. Es werden Gegenbeispiele angeführt, bei denen komplexere Zusammenhänge zwischen den Suchtermen bestehen, als sie die bisher realisierten Crosskonkordanzen enthalten. Daraus wird auf einen zu hohen Ballast bzw. auf das Verfehlen relevanter Einträge geschlossen.

Dieser Argumentation gehen wir im Folgenden auf der Basis von Beispielanfragen nach, die am IZ als Kundenwünsche eingingen und explizit sowohl auf Fakten (v.a. Umfragedaten) als auch textuelle Dokumente aller Art gerichtet waren.

3.1 Beispielgenerierung

Vier Anfragen von IZ Kunden, die sich sowohl auf Texte als auch auf Umfragedaten bezogen, wurden durch einen IuD-Spezialisten des IZ neu recherchiert, der Rücksprache mit Kollegen aus dem Zentralarchiv für empirische Sozialforschung (ZA) nahm. Die Themen waren:

1. „Befinden von Kindern berufstätiger Mütter“
2. „Umweltbewusstsein in Deutschland“
3. „Welche Faktoren beeinflussen bei potentiellen Erben von Familienbetrieben die Bereitschaft zur Betriebsübernahme?“
4. „Soziale Mobilität, speziell Bildungsmobilität in Bayern (Herkunft der Schüler und Studenten)“

Gleichzeitig wurden einige Sozialwissenschaftler um Beispiele zur Problematik der Text-Fakten-Integration gebeten und ihnen Fremdbeispiele mit Interpretation mit der Bitte um Kommentierung vorgelegt[7]. Das DJI nahm hierbei eine Sonderstellung ein. Es erarbeitet mit dem IZ einen DFG–Antrag zur Realisierung eines speziellen integrierenden Fachportals für das DJI-Kinderpanel. Abschnitt 3.3 basiert auf diesem gemeinsam erarbeiteten Antrag.

7 Unser Dank gilt den Kollegen H.M. Artus, F. Bauske, W. Bien, H. Dorn, W. Jagodzinski, R. Schnell, R. Siepmann, W. Sodeur, E. Weichselgartner und M. Zimmer.

3.2 1:1-Modellübertragung auf den Ebenen technologischer Verknüpfung und Crosskonkordanzen

Umfragen bestehen im Idealfall aus den erhobenen Daten auf Fragebogenebene und der zugehörigen Dokumentation – u. a. der Studienbeschreibung und dem Codebuch. Die Fragebögen bestehen aus den gestellten Fragen (Frageformulierung in natürlicher Sprache; in Abbildung 3 gekennzeichnet durch F_1 bis F_n) und den entsprechend gemessenen Variablen (bezeichnet durch einen kurzen Namen; in Abbildung 3 gekennzeichnet durch V_1 bis V_n). Hinzu kommen textuelle Bausteine in Kommentierungen oder Begleittexten.

Damit steht eine Fülle von textuellen Fundstellen für die Suche nach der Begrifflichkeit zur Verfügung, die für den Übergang von textueller zur Faktensuche erprobt werden kann.

- Der Übergang zwischen der Präsentation von Daten und textuellen Quellen war schon immer fließend.
  - Die Web-Suche zu Frage 1 ergab z.B. auf dem BMU-Server einen Pressehinweis zur Studie Umweltbewusstsein 2004.
  - Der Datenreport 2004 des Statistischen Bundesamtes bettet die Faktentabellen in einen kommentierenden Text ein. Bei Frage 2 führte die Kapitelüberschrift „17 Soziale Mobilität“ zu den gewünschten Fakten.
- Die Studienbeschreibungen der Datenbestandskataloge beim Zentrum für Umfragen, Mthoden und Analysen (ZUMA) enthalten eine Reihe von Metadatenfelder, die nach Suchtermen durchsucht werden können.
  - Der *Titel* enthielt bei den Beispielfragen mehrfach relevante Terme.
  - Das Feld *Inhalt* besteht aus semistrukturierten Termen, die an Thesaurusbegriffe erinnern, aber frei vergeben werden. Es findet sich fast alles: von Termen aus Unterüberschriften, die Hinweise auf das Thema geben, bis zu Begriffen aus den Suchfragen.
  - Auch hier fanden sich bei den Beispielanfragen relevante Terme.
  - Das Feld *Veröffentlichungen* (nur mit direktem Bezug zur Studie) enthielt ebenfalls relevante Terme.
- Studienbeschreibungen enthalten eine grobe *Kategorisierung* (keine Treffer bei den Beispielen)
- Die Fragen aus den *Fragebögen* und die dort gegebene Zuordnung der *Variablen*.

  Bei den Termen aus den Fragebögen ist von Anfang an klar, dass der semantische Abstand der dort zur Verfügung stehenden Begriffe zu denen der Themen wie Umweltbewusstsein oder Postmaterialismus in der Regel zu groß ist. Sie kommen somit meist nicht für die Themensuche in Betracht.

  Die Variablenlisten sind zusätzlich zum alphanumerischen Code oft mit textuellen Lesehilfen (Label) verbunden. Ihr Status für die Suche bleibt vor-

erst offen (siehe Abschnitt 3.3.1). Bei den Beispielsuchen führten sie nicht zu relevanten Termen.

Der einfachste Fall der Integration ist die technologische Vernetzung, wie sie heute jedem Wissenschaftler offen steht, der z.B. seine Daten Google öffnet. Bereits hier gibt es Treffer bei allen vier Beispielen.

Tab. 2: Beispiel Frage 1, Treffer ohne semantische Transferkomponenten (kursiv = Suchterme)

Beispielanfrage	**Datenbestandskatalog Studienbeschreibungen ZA**	**Weitere Datenquellen/Methoden**
Umweltbewusstsein in Deutschland	Studiennr. 3902	ZIS: Quellen
	Titel: *Umweltbewusstsein* in Deutschland 2002	Wingerter, C. (2005). Allgemeines *Umweltbewusstsein*. In ...
	Inhalt: Umweltbewusstsein ... *Umweltschutz*	Za&ZUMA (2005) *Umweltbelastung* (ALLBUS). In ...

Nimmt man einfache Regeln von Crosskonkordanzen hinzu wie „Umweltbewusstsein → Umweltschutz, Umweltbelastung“ erhöht sich die Trefferquote. Gleichzeitig gibt es jedoch andere Beispiele, bei denen die Zusammenhänge komplexer erscheinen.

3.3 Vom Thesaurus zur Ontologie

Beim Beispiel „*Befinden* von *Kindern berufstätiger Mütter*“ zeigt sich bereits bei der Suche textueller Dokumente eine klassische Schwierigkeit. Thesauri geben Relationen nur sehr eingeschränkt wieder (v.a. Oberbegriffe, Unterbegriffe, Ähnlichkeit, Synonymie).

Die meisten Nachweise mit obigen Suchbegriffen (z.B. bei der Datenbank FIS Bildung und beim Bildungsserver) ergeben Dokumente zur Befindlichkeit der (berufstätigen) Mutter, nicht der des Kindes. Ausschließen ließen sich diese irrelevanten Treffer nur durch eine umfangreichere Explizierung von komplexeren Relationen zwischen den Begriffen, als sie in Thesauri verwendet werden. In Thesauri wird bewusst auf sie verzichtet, da damit der Aufwand für die Indexierung deutlich steigen würde. Die These: Der zusätzliche Gewinn an Präzision in Einzelfällen rechtfertigt den Mehraufwand nicht. Der konzeptuelle Hintergrund: Beim Information Retrieval bleibt immer ein Teil der Semantik unanalysiert in den Suchtermen des Benutzers und den Deskriptoren der Inhaltserschließung erhalten. Die Intelligenz des Menschen ergänzt bei der Mensch-Maschine-Interaktion, die in der Regel iterativ mehrere Frageformulierungen zur gleichen Suchintention aneinanderreiht, diese nicht analysierten Teile (s. Kuhlen 2004). Die Frage ist somit nicht die, ob Teile der Semantik unanalysiert blei-

ben. Zu fragen ist, ob der (intelligente) Benutzer bei der Recherche diese Teile mit akzeptablem Aufwand ergänzen kann. Im einfachsten Fall geschieht dies durch das Übergehen von Ballastdokumenten. Es können aber auch komplexe deduktive Folgerungen die Voraussetzung für eine erfolgreiche Suche sein. Konsens im Information Retrieval ist, dass sich diese Frage nur empirisch klären lässt.

Konzeptuell gibt es seit langem Vorschläge zur tieferen semantischen Erschließung. Gerade als Gegenreaktion auf die unvollkommenen Verfahren der generellen Web-Suchmaschinen entwickelte sich in den letzten Jahren erneut eine breite Diskussion unter den Stichworten „semantic web", Ontologien und „topic maps" (siehe Fensel 2001 und Staab/Studer 2004 als Überblick).

> „Ontologien in Bibliotheks-, Informationswissenschaft und Informatik sind Thesauri, in denen die grundlegenden Bedeutungen von Wortfeldern und ihre Relationen zueinander in Computern abgebildet werden." (Umstätter/Wagner-Döbler 2005: 54).

Der Unterschied zum Thesaurus liegt nicht im konzeptuellen, sondern im Umfang. Die semantische Fundierung der Inhaltserschließung verbessert sich, womit sich die Hoffnung auf bessere Rechercheergebnisse verbindet. Knorz/Rein 2005:1 verdeutlicht das Dilemma dieser Ansätze:

> „Diese Hoffnung lässt sich begründen, aber wenig belegen. Anwendungen von Ontologien sind weitgehend zwei Enden eines Spektrums zuzuordnen ... diese Ontologie-Anwendungen
> - decken entweder ein weites Gebiet ab und arbeiten mit semantisch kaum differenzierten Relationen oder aber
> - sie arbeiten im Hinblick auf Attribute und Relationen sehr differenziert bei gleichzeitiger Beschränkung auf eine Miniwelt.
>
> Im ersten Fall geht das Ergebnis kaum über das hinaus, was ein konventioneller Thesaurus leistet, im zweiten Fall ... [ist] das Ergebnis im Wesentlichen auch durch konventionelle Datenbanken abzubilden".

Stammt das Unbehagen einiger Sozialwissenschaftler mit den in Abschnitt 2.3 diskutierten Heterogenitätskomponenten somit aus der Vorstellung, dass diese Art der Text-Fakten-Integration einer stärkeren semantischen Fundierung bedarf? Oder anders ausgedrückt: Befürchtet man, dass die Ergänzung der nichtanalysierten Semantik im Falle der Umfragedaten durch die Intelligenz des Benutzers in der Recherchesituation nicht nur in Einzelfällen fehlschlägt, sondern dies beim Übergang von den Texten zu den Umfragedaten eher die Regel als die Ausnahme ist?

Man kann das einfach mit den Verfahren aus Abschnitt 2.3 ausprobieren und dann empirisch testen (der bisherige Vorschlag). Andererseits kann man auch vorab die Analyse tiefer anlegen. Hierfür müssten umfragespezifische Ontolo-

gien entwickelt und mit den Inhaltserschließungsverfahren der textuellen Dokumente verknüpft werden. Das Modell erscheint auch plausibel, wenn man sich das Hochschulinformationssystem auf der Basis einer Ontologie in Knorz/Rein 2005 näher ansieht. Entscheidend war hier gerade, dass sowohl nach Fakten (Anzahl der Studenten, Gehaltssummen etc.) und gleichzeitig die klassische Literatursuche (Arbeitsberichte, Publikationslisten) im Rahmen eines einheitlichen Modells ermöglicht werden sollte.

Ontologien lassen sich sowohl flach als auch sehr stark differenziert anlegen. Deshalb muss sich der Aufwand für Ontologien nicht von vornherein als Kostenbarriere einer Integration herausstellen.

Hinweise darauf, dass dieser Weg Erfolg verspricht, ergaben sich aus einigen Beispielen zum Kinderpanel des DJI.

3.3.1 Beispiel Kinderpanel (DJI)

Die drei großen Längsschnittuntersuchungen des DJI[8], der Familiensurvey, der Jugendsurvey und das Kinderpanel, werden in der Abteilung „Social Monitoring“ inhaltlich, methodisch und forschungsorganisatorisch integriert. Das Kinderpanel untersucht je eine Alterskohorte mit fünf- bis sechsjährigen und neun- bis zehnjährigen Kindern. Dabei werden die Lebenslagen von Kindern im Sinne einer Sozialberichterstattung über Kinder differenziert beschrieben und die Einflüsse unterschiedlicher Lebenslagen auf die Persönlichkeitsentwicklung von Kindern nachgezeichnet. Die Ergebnisse der Surveys sind Teile des Informationsangebots des DJI, das neben Umfragedaten u. a. auch Literaturnachweise, Internetquellen und Projektinformationen enthält. Es wendet sich an eine breite Zielgruppe, die über die wissenschaftliche community hinausgeht.

Das DJI geht davon aus, dass seine Längsschnittstudien eine gute empirische Grundlage für politische und wirtschaftliche Entscheidungen sind und auf ein breiteres öffentliches Interesse stoßen, wenn es gelingt, sie integriert und benutzerfreundlich anzubieten. Deshalb soll es möglich sein, z. B. von einem relevanten Literaturnachweis aus damit zusammenhängende Umfragedaten oder von einer in der Umfrage erhobenen Variablen aus zugehörige Forschungsprojekte oder Publikationen zu finden, in denen die Thematik dieser Variable ebenfalls erhoben, wissenschaftlich ausgewertet oder in generellen Kontexten analysiert wurde[9]. Ein postulierter Standardfall ist, dass der Informationssuchende von ei-

8 http://www.dji.de

9 Das DJI bietet die Daten des Kinderpanel als SPSS-Dateien zum Download an, ein standardisiertes Format für die Weiterverarbeitung der Rohdaten, das aber für die inhaltliche Dokumentation der Daten weniger gut geeignet ist. Gleichzeitig werden sie fortlaufend in einer MySQL-Datenbank erfasst und sind über eine webbasierte Schnittstelle zugänglich. Über die webbasierte Schnittstelle besteht Zugriff sowohl auf die Rohdaten der Surveys im SPSS-Format als auch auf die einzelnen Variablen, deren Fragetexte und deren Ergebnisse in Form von statistischen Kennwerten. Das ZA archiviert die Daten und erstellt zusammen

nem bestimmten (sozialwissenschaftlichen) Thema ausgehend herauszufinden versucht, ob und wie ein Survey dieses Thema empirisch erhoben hat.

Die informationelle Aufbereitung und Integration der sozialwissenschaftlichen Surveys erfordert zunächst den automatischen Zugriff auf die einzelnen Surveyfragen. Der Informationssuchende kann von dort dann weiter auf die (statistisch aufbereiteten) Antworten und/oder auf thematisch verwandte Fachinformationen wie Literaturangaben, Web–Quellen, einschlägige Projekte usw. verwiesen werden.

Dass diese Zuordnung in vielen Fällen nicht mit Crosskonkordanzen oder automatisch erfolgen kann, zeigen die folgenden Beispiele.

i) Wie lange spielst du täglich am Computer?

ii) Wie oft triffst du dich zu Hause mit deinen Schulfreunden?

Frage i) kann zu einem Fragenkomplex gehören, der die Mediennutzung von Jugendlichen untersucht, wohingegen Frage ii) der Ermittlung ihres Sozialverhaltens dient. Das Freizeitverhalten von Jugendlichen ist bei beiden Fragen als übergeordnetes Thema denkbar. Die einzelnen Themen sind aber formal nicht aus den sprachlichen Ausdrücken der Frage und auch nicht aus den Ausdrücken des sprachlichen Fragekontextes verlässlich ableitbar.

Auch die Variablenbezeichnungen helfen nicht weiter. In einer Umfrage unter Kindern könnten Fragen z. B. lauten „Geschlecht des Kindes“, „Ich bin manchmal ängstlich“ oder „Ich kann mir gut vorstellen, wie sich andere Kinder so fühlen“ und die entsprechenden Variablen mit „a1001“, „a1003_07“ oder „a1052_4“ benannt sein. Es ist offensichtlich, dass weder die Frageformulierung selbst noch der Name der Variablen[10] direkt zur inhaltlichen Beschreibung der erhobenen Daten verwendet werden oder automatisch auf entsprechende Schlagwörter (Deskriptoren; in Abbildung 3 gekennzeichnet durch D1 bis Dn) eines Thesaurus abgebildet werden können. Die sozialwissenschaftliche Intention eines Informationssuchenden, der z. B. Daten über das Selbstwertgefühl von Mädchen im Alter von 8-9 Jahren finden möchte und mit Begriffen wie „Mädchen“ und „Selbstwertgefühl“ sucht, kann daher nicht direkt auf die Daten und nicht auf ihre originäre Inhaltserschließung in den Metadatenfeldern abgebildet werden.

Variablen einer Erhebung können einen viel komplexeren semantischen Zusammenhang repräsentieren als sie die Deskriptoren eines Thesaurus für Text-

---

mit dem DJI die Studienbeschreibung und die restlichen Dokumentationsunterlagen, die beim ZA standardmäßig zum Browsen und zur Termrecherche in den einzelnen Feldern der Metadaten angeboten werden
(s. http://www.gesis.org/Datenservice/Themen/53-CD-ROM/DJI-Jugendsurvey/ und http://www.gesis.org/Datenservice/Themen/53-CD-ROM/DJI-Familiensurvey/).

10 Beispiel für eine Variablenbezeichnung (Label), die alleine nicht aussagekräftig ist: „Jahre“ anstatt der eigentlichen Bedeutung „Strafmaß“.

dokumente vorsehen (z. B. „Kind“, „Geschlecht“ oder „Angst“ als unverbundene semantische Konzepte“). Das folgende Beispiel macht die mögliche semantische Unschärfe der Erschließung mit einem konventionellen Thesaurus deutlich:

> Eine Studie, die sich mit Auswirkungen der Fernsehberichterstattung über Jugendkriminalität auf das Verhalten von älteren Menschen mit unterschiedlichem Bildungshintergrund beschäftigt, könnte mit Hilfe eines sozialwissenschaftlichen Thesaurus z. B. wie folgt verschlagwortet sein: „Wirkung, Fernsehen, Berichterstattung, Jugendlicher, Kriminalität, alter Mensch, Bildungsniveau“. Bei einer 1:1 Übertragung auf die Ebene der Umfragedaten ergäben sich *missverständliche* Interpretationen:
>
> - Jugendkriminalität ist nur ein Teil des Gesamtkomplexes Kriminalität. Es ließe sich ohne eine tiefere semantische Erschließung ein missverständlicher Zusammenhang zwischen dem Gesamtkonzept Kriminalität und der Variable Alter (z.B. Alterskriminalität) oder der Variable Bildung (z.B. Korrelation zwischen deviantem Verhalten und Bildungsniveau) ableiten.
> - Die Verwendung von Jugendlicher + Kriminalität kann ohne den Begriffshintergrund Jugendkriminalität zu einem Missverständnis in der Ausprägung von Kriminalität führen: Jugendliche mit kriminellem Verhalten oder Jugendliche als Kriminalitätsopfer?

Der Nutzer einer Surveydatenbank muss bei der Suche somit die sozialwissenschaftliche Intention der Fragen erfassen, um die statistischen Ergebnisse der Fragen geeignet interpretieren zu können. Dies erfordert eine vertiefte Fachkenntnis und eine tiefere Auseinandersetzung mit einem Survey, als sie viele Nutzer vor dem Hintergrund der heutigen Grundsituation des „information overload“ bereit sind aufzubringen. Geht man somit davon aus, dass die oben diskutierten textuellen Felder der Studienbeschreibung wie Titel, Variablenlabel, Inhalt oder Kategoriezuordnung nur in Einzelfällen einen automatischen 1:1-Transfer zur Begrifflichkeit des Themas zulassen oder über die einfachen Relationierungen der Crosskonkordanzen erfasst werden können, bietet sich die Entwicklung einer Ontologie an, die die Terminologie der verschiedenen Begriffsebenen semantisch reichhaltiger miteinander verbindet als in den Thesauri.

3.3.2 Ontologiebasiertes integriertes Informationssystem zum DJI-Kinderpanel

Die Integration von Primär- und Sekundärinformationen soll somit mit Hilfe einer am sozialwissenschaftlichen Thesaurus des IZ orientierten Ontologie reali-

siert werden[11]. Die Ontologie soll die inhaltliche Beschreibung von Umfragen, der darin enthaltenen Variablen samt ihres Fragetextes und der zugrunde liegenden sozialwissenschaftlichen Intention miteinander verknüpfen. In dieses Netzwerk werden auch die Inhalte der Studienbeschreibungsfelder wie Titel, Inhalt und Klassifikation sowie ein Großteil der anderen textuellen Beschreibungen eingehen. Da es Ontologien erlauben, semantische Konzepte (in Abbildung 3 gekennzeichnet durch K1 bis K4) durch vielfältige Relationen zu verknüpfen (z.B. ist-Teil-von, beeinflusst, ist-Ergebnis-von), lassen sich komplexe Zusammenhänge präzise wiedergeben.

Die Ontologie dient zunächst der direkten Recherche nach Primärinformationen. In einem zweiten Schritt soll diese Ontologie so mit dem sozialwissenschaftlichen Thesaurus des IZ verknüpft werden, dass sich die Suche in den Umfragedaten integriert auf andere mit dem sozialwissenschaftlichen Thesaurus erschlossenen Informationen (z. B. Literatur, Forschungsprojekte, Internetquellen usw.) ausdehnen lässt.[12]

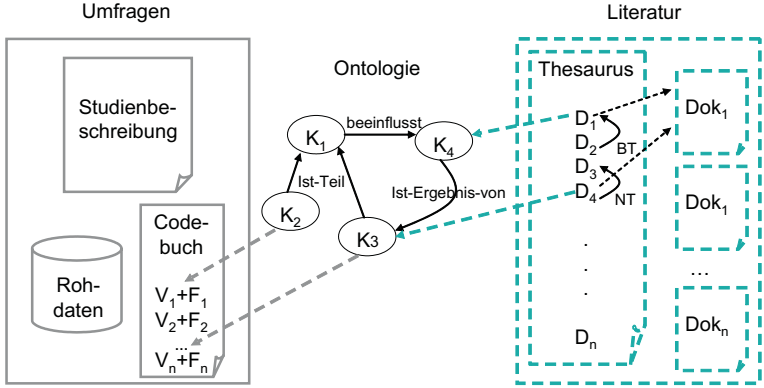


Abb. 3: Verknüpfung der Erschließung von Literatur- und Umfragedaten

Frageformulierung in natürlicher Sprache = F_i bis F_n; die entsprechend gemessenen Variablen (bezeichnet durch einen kurzen Namen) = V_i bis $V_{n;}$ Deskriptoren = D_i bis D_n

11 Für die Daten des DJI-Surveys existiert bislang noch kein entsprechendes standardisiertes Vokabular.

12 Innerhalb der Surveys stellt sich das spezielle Problem, von einer bestimmten für die eigene Forschung relevanten Surveyfrage ausgehend, ähnlich formulierte oder semantisch ähnliche Fragen in anderen Umfragen zu finden. Diese zusätzliche Ableitung kann mit computerlinguistischen oder statistischen Textverfahren realisiert werden.

4 Fazit

Es spricht viel dafür, dass sich die derzeitige Diskussion um die Text-Fakten-Integration durch den Ansatz einer ontologiebasierten Inhaltserschließung auf Seiten der Surveydaten sinnvoll fortentwickeln lässt.

Wie tief diese Ontologie sein muss oder ob nicht doch in vielen Bereichen die kostengünstigeren Lösungen der Heterogenitätsbehandlung textueller Dokumente ausreichen, lässt sich auf solch einer experimentellen Grundlage dann empirisch prüfen. In einem Wiederholungstest würden die Anfragen an das Ontologie basierte System erneut gestellt, nachdem die Relationen schrittweise entfernt wurden. Zeigt sich eine deutliche Ergebnisverschlechterung, wäre der zusätzliche Aufwand gerechtfertigt. Damit ließe sich auch die Tiefe der Ontologie experimentell optimieren und so bei zukünftigen Systementwicklungen der Aufwand reduzieren.

Auch für den Zugriff auf die textuellen Dokumente ließe sich durch den Ontologieeinsatz ein Mehrwert erzielen. Auf der Basis der in der Ontologie codierten Semantik kann versucht werden, komplexere und zugleich präzisere Textsuchen zu formulieren. Moderne Textretrievalsysteme erlauben z.B. die Verwendung von Näheoperatoren (zwei oder mehrere Begriffe müssen benachbart, in einem Satz oder in einem Absatz vorkommen), Reihenfolgeoperatoren (Begriffe müssen in einer bestimmten Reihenfolge vorkommen) oder unterschiedliche Gewichtung einzelner Suchbegriffe. In den mit Abstract und teilweise auch Volltext versehenen Fachdatenbanken kann somit mit einer Mischung aus boolescher Schlagwortsuche und boolescher oder statistischer Freitextsuche versucht werden, die Suchanfrage zu präzisieren.

Der Hauptvorteil liegt jedoch zum gegenwärtigen Zeitpunkt in dem Potential dieses Vorgehens zur Konfliktauflösung. Die Ontologie basierte Entwicklungshypothese scheint die notwendige Akzeptanz einer Anfangsplausibilität in den Sozialwissenschaften eher zu erreichen als die – wenn auch weniger arbeitsaufwändige – der Übertragung der Vorgehensweise von Abschnitt 2.3.

Literatur

Fensel, Dieter (2001): Ontologies. A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Berlin et al.

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.) (2001): Wege zu einer besseren informationellen Infrastruktur. Nomos, Baden-Baden.

Knorz, Gerhard; Rein, Birgit (2005): Semantische Suche in einer Hochschulontologie. In: Information, Wissenschaft & Praxis 56 Nr. 5 (erscheint).

Krause, Jürgen (2004a): Konkretes zur These, die Standardisierung von der Heterogenität her zu denken. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie 51, Nr. 2, S. 76 - 89.

Krause, Jürgen (2004b): Kapitel D 16: Standardisierung und Heterogenität. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation: Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. völlig neu gefasste Ausgabe 2004. München: Saur. (DGD-Schriftenreihe), S. 635 - 641.

Krause, Jürgen (2004c): Standardization, Heterogeneity and the Quality of Content Analysis: a key conflict of digital libraries and its solution. In: IFLA Journal: Official Journal of the International Federation of Library Associations and Institutions 30, No. 4, S. 310 – 318.

Krause, Jürgen; Niggemann, Elisabeth; Schwänzl, Roland (2003): Normierung und Standardisierung in sich verändernden Kontexten: Beispiel: Virtuelle Fachbibliotheken. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie, 50, Nr. 1, S. 19 - 28.

Krause, Jürgen; Schmiede, Rudi (2004): Auf dem Weg zu einem Fachportal Sozialwissenschaften. In: Soziologie - Forum der Deutschen Gesellschaft für Soziologie 33, Nr. 3, S. 22 - 38.

Kuhlen, Rainer (2004): Kapitel A 1: Information. In: Kuhlen, Rainer; Seeger, Thomas; Strauch, Dietmar (Hrsg.): Grundlagen der praktischen Information und Dokumentation: Band 1: Handbuch zur Einführung in die Informationswissenschaft und -praxis, 5. völlig neu gefasste Ausgabe 2004. München: Saur. (DGD-Schriftenreihe), S. 3 - 20.

Mandl, Thomas (2001): Tolerantes Information Retrieval. Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Konstanz: UVK, Univ.-Verl. (Schriften zur Informationswissenschaft; Bd. 39).

Marx, Matthias (2005): Empirische Ergebnisse zur Evaluation semantischer Transformationen. IZ-Arbeitsbericht. Juli 2005 (erscheint).

Mayr, Philipp; Stempfhuber, Maximilian; Walter, Anne-Kathrin (2005): Auf dem Weg zum wissenschaftlichen Fachportal – Modellbildung und Integration heterogener Informationssammlungen. In: Ockenfeld, Marlis (Hrsg.): 27. DGI-Online-Tagung. Frankfurt am Main: DGI. S. 29 - 43.

Müller, Matthias N.O. (2003): Integration unterschiedlich erschlossener Datenbestände am Beispiel der Virtuellen Fachbibliothek Sozialwissenschaften. In: Schmidt, Ralph (Hrsg.): Competence in Content. 25. Online-Tagung der DGI, Frankfurt am Main, 03. bis 05. Juni 2003. Frankfurt am Main: DGI. S. 335 – 345.

Musgrave, Simon (2003): NESSTAR Software Suite. http://www.nesstar.org (January 2003)

Pianos, Tamara (2005): Was macht eigentlich vascoda? Vision und Wirklichkeit. In: ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie, 52, S. 67 -78.

Staab, Steffen; Studer, Rudi (Hrsg.) (2004). Handbook on Ontologies. Berlin: Springer.

Umstätter, Walther; Wagner-Döbler, Roland (2005): Einführung in die Katalogkunde - Vom Zettelkatalog zur Suchmaschine. Stuttgart: Hiersemann.

Zhang, Xueying (2005): Concept integration of document databases using different indexing languages. In: Information Processing & Management 41 (erscheint).

Fusion und Integration von Daten: Datenschutz und Standesregeln

Erich Wiegand

1 Einleitung

Die Fusion von Datenbanken, das Zuspielen von Forschungsergebnissen in Datenbanken und das Anreichern von Forschungsergebnissen aus Datenbanken haben für die Markt- und Sozialforschung erheblich an Bedeutung gewonnen. Die Verbände stehen deshalb vor der Aufgabe, die allgemein anerkannten berufsständischen Verhaltensregeln auf die Fusion und Integration von Daten anzuwenden und damit diese Forschungstechniken in das System der Selbstregulierung der Markt- und Sozialforschung einzubeziehen.

Mit der kürzlich verabschiedeten und gerade im Internet unter www.adm-ev.de veröffentlichten „Richtlinie zum Umgang mit Datenbanken in der Markt- und Sozialforschung" haben die Verbände der deutschen Markt- und Sozialforschung – also der ADM, die ASI, der BVM und die D.G.O.F. – diesbezüglich einen wichtigen Schritt getan. Ich werde deshalb meinen Vortrag im Einzelnen an den Inhalten dieser Richtlinie ausrichten.

Zuvor werde ich allerdings das System der Selbstregulierung und die Grundprinzipien des berufsständischen Verhaltens in der Markt- und Sozialforschung darstellen. Sie sind der Kontext, in dem die berufsständischen Verhaltensregeln für den wissenschaftlichen Umgang mit Datenbanken bzw. die Fusion und Integration von Daten in der Markt- und Sozialforschung zu sehen sind.

2 Selbstregulierung in der Markt- und Sozialforschung

Die Markt- und Sozialforschung betreibt im Gegensatz zu vielen anderen Branchen schon seit Jahrzehnten eine ernsthafte Selbstregulierung des berufsständischen Verhaltens. Das gilt nicht nur für Deutschland, sondern auch im internationalen Kontext: Bereits seit dem Jahr 1948 gibt es den „IHK/ESOMAR Internationalen Kodex für die Praxis der Markt- und Sozialforschung", der in den nunmehr 57 Jahren seines Bestehens natürlich mehrfach überarbeitet wurde. Dieser Kodex regelt die grundlegenden berufsständischen Pflichten von Markt- und Sozialforschern gegenüber den Befragten sowie den Auftraggebern und der Öffentlichkeit. Er wird inzwischen weltweit von über hundert nationalen Verbänden der Markt- und Sozialforschung in mehr als fünfzig Ländern anerkannt.

Die deutschen Verbände der Markt- und Sozialforschung haben den „ESOMAR-Kodex“ mit einer vorangestellten „Erklärung für das Gebiet der Bundesrepublik Deutschland“ versehen und ihm verschiedene Richtlinien zur Seite gestellt. In diesen Richtlinien werden die grundlegenden berufsständischen Verhaltensregeln des Kodex im Hinblick auf verschiedene Methoden und Techniken der empirischen Markt- und Sozialforschung konkretisiert.

Die Einhaltung der Berufsgrundsätze und Standesregeln wird in Deutschland durch den Rat der Deutschen Markt- und Sozialforschung e.V. gewährleistet. Dabei handelt es sich um eine verbandsübergreifende Beschwerdestelle, die im Jahr 2001 von den Verbänden der deutschen Markt- und Sozialforschung gemeinsam gegründet wurde. Ihre Arbeitsweise und Sanktionsmöglichkeiten sind ähnlich denen des Deutschen Presserats und des Deutschen Werberats. Die Selbstregulierung der deutschen Markt- und Sozialforschung ist also kein „zahnloser Tiger“.

In der Vergangenheit standen die Verbände verschiedentlich mit den Aufsichtsbehörden für den Datenschutz wegen der Selbstregulierung der Markt- und Sozialforschung in Kontakt. Sie sind dazu bereit, diese Kontakte zu intensivieren und zu systematisieren, um die bestehende Selbstregulierung in Richtung einer Koregulierung weiter zu entwickeln. Bedingung für den Erfolg einer solchen Koregulierung der Markt- und Sozialforschung ist allerdings eine auf Interessenausgleich gerichtete Rechtsgüterabwägung durch die daran beteiligten Institutionen.

Folglich müssen die aus den Artikeln 1 und 2 des Grundgesetzes für die Bundesrepublik Deutschland ableitbaren Persönlichkeitsrechte – insbesondere das Recht auf informationelle Selbstbestimmung – und die durch Artikel 5 garantierte Freiheit der Wissenschaft und Forschung durch konkrete berufsständische Verhaltensregeln so miteinander in Einklang gebracht werden, dass sowohl die Verbände der deutschen Markt- und Sozialforschung als auch die zuständigen Aufsichtsbehörden mit dem Ergebnis ihrer Verhandlungen zufrieden sein können.

3 Grundprinzipien des berufsständischen Verhaltens

Zu den sowohl forschungsethisch und methodisch als auch rechtlich begründeten Grundprinzipien des berufsständischen Verhaltens in der wissenschaftlichen Markt- und Sozialforschung gehören die Wahrung der Anonymität der befragten Personen und die Trennung zwischen wissenschaftlicher Forschung und anderen Tätigkeiten, insbesondere der Werbung und Verkaufsförderung. Beide berufsständischen Verhaltensregeln gelten grundsätzlich auch für die Fusion und Integration von Daten bzw. für den wissenschaftlichen Umgang mit Datenbanken.

Die Wahrung der Anonymität bedeutet, dass die befragten Personen sicher sein können, dass die bei ihnen erhobenen Daten nur in anonymisierter Form für Forschungszwecke verarbeitet und an den Auftraggeber der Umfrage übermittelt werden. Das schließt zugleich aus, dass die erhobenen Daten in personenbezogener Form gegenüber den befragten Personen gezielt für Marketingzwecke genutzt werden können.

Damit garantiert die Wahrung der Anonymität zugleich die Trennung zwischen wissenschaftlicher Forschung und forschungsfremden Tätigkeiten. Diese Trennung bedeutet natürlich nicht, dass die anonymisierten Forschungsergebnisse im Marketing, in der Werbung und in anderen Bereichen keine Anwendung finden. Wenn das der Fall wäre, hätte die Marktforschung – und auch ein Teil der empirischen Sozialforschung – ihre Funktion als angewandte wissenschaftliche Forschung verfehlt.

4 Arten des Umgangs mit Datenbanken in der Markt- und Sozialforschung

In der Markt- und Sozialforschung sind fünf grundlegende Arten des wissenschaftlichen Umgangs mit Datenbanken zu unterscheiden:

1. das Ziehen von Stichproben aus Datenbanken,
2. die mathematisch-statistische Analyse von Datenbanken,
3. die Fusion von Datenbanken,
4. das Zuspielen von Forschungsergebnissen in Datenbanken,
5. das Anreichern von Forschungsergebnissen aus Datenbanken.

4.1 Ziehen von Stichproben aus Datenbanken

Für das Ziehen von Stichproben aus Datenbanken zum Zwecke wissenschaftlicher Umfragen der Markt- und Sozialforschung ist aus datenschutzrechtlicher Sicht vor allem von Bedeutung, dass die zur Befragung ausgewählten Personen bei der Bitte um ein Interview gegebenenfalls auf ihr Widerspruchsrecht gemäß § 28 BDSG hingewiesen werden.

4.1.1 Exkurs: Definition des Begriffs „Kennziffern“

Bevor ich auf die anderen Arten des wissenschaftlichen Umgangs mit Datenbanken in der Markt- und Sozialforschung zu sprechen komme, muss ich zuvor den Begriff „Kennziffern“ definieren, denn sie spielen sowohl bei der mathematisch-statistischen Analyse von Datenbanken als auch beim Zuspielen von Forschungsergebnissen in Datenbanken eine entscheidende Rolle: In der „Richtli-

nie zum Umgang mit Datenbanken in der Markt- und Sozialforschung“ bezeichnen „Kennziffern“ alle Ergebnisse mathematisch-statistischer Operationen, die aus den in Datenbanken bereits gespeicherten Merkmalen oder aus durch Umfragen erhobenen Daten in Form von Indizes, Scores, Typen oder ähnlichem berechnet und individuellen Datensätzen zugeordnet werden.

4.2 Mathematisch-statistische Analyse von Datenbanken

Bei der mathematisch-statistischen Analyse von in Datenbanken bereits gespeicherten Daten werden Datenbanken bzw. daraus gezogene Stichproben mittels wissenschaftlicher Verfahren auf mögliche und bisher unbekannte Strukturen und Zusammenhänge der in der Datenbank enthaltenen Merkmale untersucht. Dabei werden die besagten Kennziffern berechnet und als statistische Erwartungswerte den einzelnen Fällen in der Datenbank zugeordnet. Das geschieht mittels mathematisch-statistischer Klassifikations- oder Zuordnungsverfahren. Die wissenschaftliche Analyse von Datenbanken erfordert keine Verarbeitung personenbezogener Daten. Sie ist deshalb datenschutzrechtlich ohne Einschränkung zulässig.

4.3 Fusion von Datenbanken

Bei der Fusion von Datenbanken werden den in einer (empfangenden) Datenbank enthaltenen Merkmalen die in einer anderen (spendenden) Datenbank enthaltenen Merkmale mittels mathematisch-statistischer Klassifikations- oder Zuordnungsverfahren als weitere Merkmale fallweise zugespielt. Technische Voraussetzung dafür ist, dass eine Anzahl von gemeinsamen Merkmalen in beiden Datenbanken enthalten sind, um auf der Grundlage ähnlicher Kombinationen der Merkmalsausprägungen – also auf der Grundlage so genannter „statistischer Zwillinge“ – die Zuordnung vornehmen zu können. Der wissenschaftliche Zweck der Fusion von Datenbanken besteht hauptsächlich darin, die mathematisch-statistischen Analysen oder die Auswahl von zu befragenden Personen auf einer breiteren Grundlage von Merkmalen vornehmen zu können.

Die Fusion von Datenbanken ist zulässig, wenn keine personenbezogenen Daten verarbeitet werden oder wenn wegen der Art der Daten oder der Art ihrer Nutzung keine schutzwürdigen Belange der Betroffenen beeinträchtigt werden. Davon kann in der wissenschaftlichen Markt- und Sozialforschung grundsätzlich ausgegangen werden.

4.4 Zuspielen von Forschungsergebnissen in Datenbanken

Beim Zuspielen von Forschungsergebnissen in Datenbanken werden den in der Datenbank bereits gespeicherten Merkmalen in Form von berechneten Kennziffern fallweise neue Merkmale hinzugefügt. Dabei können die den Forschungsergebnissen – d.h. den berechneten Kennziffern – zugrunde liegenden Daten sowohl bei den in der Datenbank erfassten Personen als auch bei darin nicht erfassten Personen erhoben worden sein. Das Zuspielen von Forschungsergebnissen in Datenbanken ist auf unterschiedliche Weise möglich:

1. durch Übermittlung der Zuspielungsregeln,
2. durch Zuspielen in pseudonymisierte Datenbanken,
3. durch Zuspielen in personenbezogene Datenbanken,
4. durch Zuspielen in personenbezogener Form.

4.4.1 Übermittlung der Zuspielungsregel

Bei der Übermittlung der Zuspielungsregeln werden dem Auftraggeber von der forschenden Stelle die mathematischen Regeln übermittelt, nach denen die Kennziffern berechnet und als statistische Erwartungswerte fallweise zugeordnet werden. Die Berechnung und Zuordnung nimmt der Auftraggeber selbst vor. Der forschenden Stelle werden keine personenbezogenen Daten übermittelt.

4.4.2 Zuspielen in pseudonymisierte Datenbanken

Beim Zuspielen in pseudonymisierte Datenbanken übermittelt der Auftraggeber die Datenbank, der die Forschungsergebnisse zugespielt werden sollen, der forschenden Stelle ohne Identifikationsmerkmale. Die forschende Stelle nimmt die Berechnung und fallweise Zuordnung der Kennziffern vor und übermittelt dem Auftraggeber die solcherart ergänzte Datenbank. Die forschende Stelle hat auch bei dieser Vorgehensweise keinen Zugriff auf die in der Datenbank enthaltenen personenbezogenen Daten.

4.4.3 Zuspielen in personenbezogene Datenbanken

Beim Zuspielen in personenbezogene Datenbanken enthält die Datenbank, die der Auftraggeber der forschenden Stelle zum Zwecke des Zuspielens der Forschungsergebnisse übermittelt, nicht anonymisierte oder pseudonymisierte, sondern eben personenbezogene Daten. Die forschende Stelle nimmt wieder die Berechnung und fallweise Zuordnung der Kennziffern vor und übermittelt dem Auftraggeber die solcherart ergänzte Datenbank. Diese Vorgehensweise ist nur

zulässig, wenn eine Einwilligung der Betroffenen in die Übermittlung ihrer personenbezogenen Daten an die forschende Stelle vorliegt.

4.4.4 Zuspielen in personenbezogener Form

Das Zuspielen der in der wissenschaftlichen Markt- und Sozialforschung erhobenen Daten in personenbezogener Form in Datenbanken des Auftraggebers ist nach den Standesregeln der Markt- und Sozialforschung in jedem Fall unzulässig, auch wenn eine entsprechende Einwilligung vorläge. Trotzdem will ich auch diese theoretisch mögliche Form der Zuspielung der Vollständigkeit halber hier erwähnt haben.

4.5 Anreichern von Forschungsergebnissen aus Datenbanken

Beim Anreichern von Ergebnissen der Markt- und Sozialforschung aus Datenbanken werden den durch Befragung erhobenen Daten in einer Datenbank gespeicherte Merkmale fallweise zugespielt. Häufig handelt es sich dabei um beim Auftraggeber der Untersuchung gehaltene Datenbanken. Diese Art des Umgangs mit Datenbanken ist sozusagen das Gegenteil des Zuspielens von Forschungsergebnissen in Datenbanken. Das Anreichern von Forschungsergebnissen aus Datenbanken ist zulässig, wenn dadurch bei der Rückübermittlung der Forschungsergebnisse an den Auftraggeber die Anonymität der befragten Personen nicht gefährdet wird.

5 Einwilligung der Betroffenen

Die dargestellten Arten des Umgangs mit Datenbanken in der Markt- und Sozialforschung sind datenschutzrechtlich nicht zu beanstanden, wenn die in der Regel notwendige Einwilligung der betroffenen Personen vorliegt. Dabei ist zu unterscheiden zwischen der Einwilligung der Personen, die an einer wissenschaftlichen Befragung teilnehmen sollen, und der Einwilligung der in einer Datenbank erfassten Personen in die wissenschaftliche Verarbeitung und Nutzung ihrer Daten.

Die Teilnahme an einer wissenschaftlichen Untersuchung der Markt- und Sozialforschung basiert immer auf einer Einwilligung der dafür ausgewählten Personen. Beim Einholen der Einwilligung durch die forschende Stelle wird neben dem Hinweis auf die Freiwilligkeit der Teilnahme und der Anonymisierung der erhobenen Daten auch über den allgemeinen Zweck der Untersuchung informiert. Und natürlich werden dabei die zu befragenden Personen gegebenenfalls auch auf ihr Widerspruchsrecht gemäß § 28 BDSG hingewiesen.

Für das Verarbeiten und Nutzen von personenbezogenen Daten in einer Datenbank ist eine Einwilligung der betroffenen Personen erforderlich. Dabei ist vom Halter der Datenbank zugleich eine Einwilligung für Zwecke der Markt- und Sozialforschung einzuholen, wenn spätere mathematisch-statistische Analysen der Datenbank oder das Zuspielen von Forschungsergebnissen geplant oder wahrscheinlich sind.

6 Selbstregulierung und Koregulierung

Lassen Sie mich zum Schluss meines Vortrags nochmals auf das Prinzip der Selbstregulierung bzw. der Koregulierung zurückkommen. In einer Zeit des raschen Wandels der Informationstechnologien, der natürlich auch vor den Forschungstechniken und vor allem den Möglichkeiten der Fusion und Integration von Daten nicht halt macht, sind gesetzliche Vorschriften immer weniger geeignet, berufsständisches Verhalten angemessen zu regeln. Deshalb sehe ich in einer ernsthaften Selbstregulierung bzw. einer vernünftigen Koregulierung für die wissenschaftliche Markt- und Sozialforschung den „Königsweg“, um den Schutz der Persönlichkeitsrechte und zugleich das Recht auf Freiheit der Forschung sicher zu stellen, denn beide führen zu einer entsprechenden Selbstverpflichtung mit konkreten berufsständischen Verhaltensregeln.

Der Begriff „vernünftig“ in Bezug auf die Koregulierung beinhaltet, dass sie in einem fairen Diskussionsprozess der beteiligten Institutionen und Verbände zu Verhaltensregeln führt, die das Resultat einer auf Interessenausgleich gerichteten Rechtsgüterabwägung darstellen und deshalb von den verschiedenen Interessengruppen oder Stakeholdern gleichermaßen akzeptiert werden können.

Verzeichnis der Autorinnen und Autoren

Haluk Akinci M.A.
Universität zu Köln, MLFZ
Greinstr. 2, 50939 Köln
Tel.: 0221/470-4232, Fax: 0221/470-5169
E-Mail: akinci@wiso.uni-koeln.de

Uwe Czaia
CZAIA Marktforschung GmbH
Kleiner Ort 1, 28357 Bremen
Tel.: 0421/207-1300, Fax: 0421/207-1330
E-Mail: u.czaia@czaia-marktforschung.de

Dr. Jörg Hagenah
Universität zu Köln, MLFZ
Greinstr. 2, 50939 Köln
Tel.: 0221/470-6163, Fax: 0221/470-5169
E-Mail: hagenah@wiso.uni-koeln.de

Johann Hahlen
Statistisches Bundesamt
Gustav-Stresemann-Ring 11, 65189 Wiesbaden
Tel.: 0611/75-2100, Fax: 0611/75-3183
E-Mail: johann.hahlen@destatis.de

PD Dr. Jürgen H. P. Hoffmeyer-Zlotnik
Zentrum für Umfragen, Methoden und Analysen (ZUMA)
B 2, 1, 68159 Mannheim
Tel.: 0621/1246-175, Fax: 0621/1246-100
E-Mail: hoffmeyer-zlotnik@zuma-mannheim.de

Dr. Hans Kiesl
Institut für Arbeitsmarkt- und Berufsforschung (IAB)
Regensburger Str. 104, 90478 Nürnberg
Tel.: 0911/179-3084, Fax: 0911/179-3258
E-Mail: Hans.Kiesl@iab.de

Prof. Dr. Jürgen Krause
Informationszentrum Sozialwissenschaften
Lennéstr. 30, 53113 Bonn
Tel.: 0228/2281-145, Fax: 0228/2281-4
E-Mail: krause@bonn.iz-soz.de

Dr. Stefan Tuschl
Head of the EX-A-MINE Centre
TNS Infratest
Landsberger Str. 338, 80687 München
Tel.: 089/5600-1107, Fax: 089/5600-1611
E-Mail: stefan.tuschl@tns-infratest.com

Prof. Dr. Heiner Meulemann
Institut für angewandte Sozialforschung der Universität zu Köln (IfAS)
Greinstr. 2, 50939 Köln
Tel.: 0221/470-5658; Sekr. -5714, Fax: 0221/470-5169
E-Mail: meulemann@wiso.uni-koeln.de

PD Dr. Susanne Rässler
Institut für Arbeitsmarkt- und Berufsforschung (IAB)
Regensburger Str. 104, 90478 Nürnberg
Tel.: 0911/179-3084, Fax: 0911/179-3258
E-Mail: susanne.raessler@iab.de

Hartmut Scheffler
TNS Emnid GmbH & Co. KG
Stieghorster Str. 66, 33605 Bielefeld
Tel.: 0521/92 57-328, Fax: 0521/92 57-250
E-Mail: hartmut.scheffler@tns-emnid.com

Hans-Gerd Siedt
Statistisches Bundesamt
Gustav-Stresemann-Ring 11, 65189 Wiesbaden
Tel.: 0611/75-2845, Fax: 0611/75-4000
E-Mail: hans-gerd.siedt@destatis.de

Dr. Maximilian Stempfhuber
Informationszentrum Sozialwissenschaften
Lennéstr. 30, 53113 Bonn
Tel.: 0228/2281-145, Fax: 0228/2281-4
E-Mail: stempfhuber@bonn.iz-soz.de

Michael Wiedenbeck
Zentrum für Umfragen, Methoden und Analysen
B 2, 1, 68159 Mannheim
Tel.: 0621/1246-279, Fax: 0621/1246-100
E-Mail: wiedenbeck@zuma-mannheim.de

Erich Wiegand
Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. (ADM)
Langer Weg 18, 60489 Frankfurt
Tel.: 069/97 84 31 36, Fax: 069/97 84 31 37
E-Mail: wiegand@adm-ev.de

Dr. Raimund Wildner
GfK AG
Nordwestring 101, 90319 Nürnberg
Tel.: 0911/395-2573, Fax: 0911/395-4041
E-Mail: raimund.wildner@gfk.de

Teilnehmerverzeichnis

A

Abele, Franz; *Statistisches Amt der Landeshauptstadt Stuttgart*
Akinci, Haluk; *Universität zu Köln*
Alter, Hannah; *Statistisches Bundesamt*, Wiesbaden
Arikan, Sanyel; *Hessisches Statistisches Landesamt*, Wiesbaden

B

Bachmann, Thomas; *TNS Infratest Holding GmbH & Co. KG*, München
Bandilla, Dr. Wolfgang; *Zentrum für Umfragen, Methoden und Analysen (ZUMA)*, Mannheim
Baur, Nina; *Katholische Universität Eichstätt-Ingolstadt*
Behrens, Kurt; *BIK ASCHPURWIS + BEHRENS GMBH Markt-, Media- und Regionalforschung*, Hamburg
Berke, Paul; *Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen*, Düsseldorf
Berlin, Jan; *SKOPOS Institut für Markt- und Kommunikationsforschung GmbH*, Hürth
Bihler, Wolf; *Statistisches Bundesamt*, Wiesbaden
Birkigt, Holger; *Statistisches Bundesamt*, Wiesbaden
Blohm, Dieter; *Hessisches Statistisches Landesamt*, Wiesbaden
Brand, Dr. Ruth; *Statistisches Bundesamt*, Bonn
Breiholz, Holger; *Statistisches Bundesamt*, Bonn
Bruckert, Andreas; *Mafo-Institut GmbH & Co.KG*, Schwalbach a. Ts.
Buck, Dr. Peter W.; *Hessisches Statistisches Landesamt*, Wiesbaden

C

Chatzis, Martina; *MARPLAN Forschungsgesellschaft mbH*, Offenbach
Czaia, Uwe; *CZAIA Marktforschung GmbH*, Bremen

D

Demant, Brigitte; *Statistisches Bundesamt*, Bonn
Deutschmann, Marc; *Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen*, Düsseldorf
Dittrich, Stefan; *Statistisches Bundesamt*, Wiesbaden

E

Ebert, Michael; *Ebert + Grüntjes GbR*, Obertshausen
Ehling, Dr. Manfred; *Statistisches Bundesamt*, Wiesbaden
Enderer, Jörg; *Statistisches Bundesamt*, Wiesbaden

F

Faulbaum, Prof. Dr. Frank; *Universität Duisburg-Essen*, Campus Duisburg

Fender, Raimund; *Millward Brown Germany GmbH & Co. KG*, Frankfurt am Main
Fleck, Dr. Claudia; *Statistisches Bundesamt*, Wiesbaden
Förster, Hubert; *Media Markt Analysen GmbH & Co.KG*, Frankfurt am Main
Frietsch, Rainer; *Fraunhofer-Institut für System- und Innovationsforschung*, Karlsruhe
Fürnrohr, Dr. Michael; *Bayerisches Landesamt für Statistik und Datenverarbeitung*, München

G

Goebel, Jan; DIW Berlin, *Deutsches Institut für Wirtschaftsforschung*, Berlin
Gossler, Marc; *LINK Institut für Markt- und Sozialforschung GmbH*, Frankfurt am Main
Gräb, Christopher; *Statistisches Bundesamt*, Wiesbaden
Granath, Ralf-Olaf; *Bundesinstitut für Berufsbildung*, Bonn
Gruber, Stefan; *Statistisches Bundesamt*, Bonn
Grüning, Ingrid; *Arbeitsgemeinschaft Media-Analyse / Media-Micro-Census GmbH*, Frankfurt am Main
Grüntjes, Jens; *Ebert + Grüntjes GbR*, Obertshausen

H

Hagenah, Dr. Jörg; *Universität zu Köln*
Hahlen, Johann; *Statistisches Bundesamt*, Wiesbaden
Hein, Birgit; *Statistisches Bundesamt*, Bonn
Heitzig, Dr. Jobst; *Statistisches Bundesamt*, Wiesbaden
Hellenschmidt, Jens; *Bundesministerium für Verkehr, Bau- und Wohnungswesen*, Bonn
von der Heyde, Christian; *TNS Infratest Holding GmbH & Co. KG*, München
Hoffmann, Hermann; *Ipsos GmbH Medienforschung*, Hamburg
Hoffmeyer-Zlotnik, PD Dr. Jürgen H. P.; *Zentrum für Umfragen, Methoden und Analysen (ZUMA)*, Mannheim
Hofmann, Melanie; *ENIGMA GfK Medien- und Marketingforschung GmbH*, Wiesbaden
Houben, Henriette; *Universität Hannover*

J

Jahn, Andreas; *IWD Marktforschung*, Magdeburg

K

Kaiser, Melanie; *ENIGMA GfK Medien- und Marketingforschung GmbH*, Wiesbaden
Ketzmerick, Thomas; *Zentrum für Sozialforschung Halle e.V. an der Martin-Luther-Universität Halle-Wittenberg*

Kiesl, Dr. Hans; *Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit (IAB)*, Nürnberg
Klein, Rainer; *Statistisches Landesamt Rheinland-Pfalz*, Bad Ems
König, Christian; *Statistisches Bundesamt*, Wiesbaden
Kordsmeyer, Volker; *Statistisches Bundesamt*, Wiesbaden
Köster, Gabriele; *Statistisches Landesamt des Freistaates Sachsen*, Kamenz
Krause, Prof. Dr. Jürgen; *Informationszentrum Sozialwissenschaften*, Bonn
Krebs, Prof. Dr. Dagmar; *Justus-Liebig-Universität Gießen*
Kuchler, Carsten; *Statistisches Bundesamt*, Wiesbaden
Kusiak, Marzena; *SCHAEFER MARKTFORSCHUNG Institut für Markt-, Sozial- und Werbeforschung GmbH*, Hamburg

L

Lauterbach, Dr. Nora; *Statistisches Bundesamt*, Wiesbaden
Lorentz, Dr. Kai; *Statistisches Bundesamt*, Wiesbaden
Lukanow, Katja; *Zentrum für Sozialforschung Halle e.V. an der Martin-Luther-Universität Halle-Wittenberg*

M

Maiterth, Dr. Ralf; *Universität Hannover*
Menning, Sonja; *Deutsches Zentrum für Altersfragen*, Berlin
Metschke, Dr. Rainer; *Berliner Beauftragter für Datenschutz und Informationsfreiheit*, Berlin
Meulemann, Prof. Dr. Heiner; *Institut für angewandte Sozialforschung der Universität zu Köln (IfAS)*
Müller, Berthold; *Hessisches Statistisches Landesamt*, Wiesbaden

N

Neubarth, Wolfgang; *Zentrum für Umfragen, Methoden und Analysen (ZUMA)*, Mannheim
Neiestroy, Marek; *ENIGMA GfK Medien- und Marketingforschung GmbH*, Wiesbaden

P

Peper, Jürgen; *Niedersächsisches Landesamt für Statistik*, Hannover
Puhl, Achim; *Institut für Sozialarbeit und Sozialpädagogik e.V. (ISS)*, Frankfurt am Main

R

Radmacher-Nottelmann, Nils; *Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen*, Düsseldorf
Rässler, PD Dr. Susanne; *Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit (IAB)*, Nürnberg
Rengers, Dr. Martina; *Statistisches Bundesamt*, Wiesbaden

Rolland, Sebastian; *Statistisches Bundesamt*, Bonn
Rösch, Günther; *Büro für Erhebungsdesign und Datenanalyse*, Frauenberg

S

Sacher, Matthias; *Statistisches Bundesamt*, Wiesbaden
Schäfer, Dieter; *Statistisches Bundesamt*, Wiesbaden
Scheffler, Hartmut; *TNS Infratest Holding GmbH & Co. KG*, Bielefeld
Schwickerath, Marco; *Statistisches Bundesamt*, Wiesbaden
Siedt, Hans-Gerd; *Statistisches Bundesamt*, Wiesbaden
Siepmann, Dr. Rolf; *Deutsches Jugendinstitut e.V.*, München
Sodeur, Prof. Dr. Wolfgang; *Universität Duisburg-Essen*, Campus Essen
Sommer, Kay; *Statistisches Bundesamt*, Wiesbaden
Sopp, Gerd; *IFAK-Institut GmbH & Co. KG Markt- und Sozialforschung*, Taunusstein
Stahl, Matthias; *Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V.*, Bonn
Stock, Dr. Gerhard; *Statistisches Bundesamt*, Wiesbaden
Stock, Dr. Wilfried; *Institut für angewandte Verkehrs- und Tourismusforschung e.V. (IVT)*, Mannheim
Stralla, Dr. Heinz; *Statistisches Bundesamt*, Wiesbaden
Stroh, Astrid; *Statistisches Bundesamt*, Wiesbaden
Sturm, Roland; *Statistisches Bundesamt*, Wiesbaden

T

Tillmanns, Dr. Christoph; *GfK Fernsehforschung GmbH*, Nürnberg
Tuschl, Dr. Stefan; *TNS Infratest Holding GmbH & Co. KG*, München

V

Venema, Mathias; *MARPLAN Forschungsgesellschaft mbH*, Offenbach
Vorgrimler, Dr. Daniel; *Statistisches Bundesamt*, Wiesbaden

W

Walsemann, Ute; *Statistisches Bundesamt*, Bonn
Walther, Dr. Matthias; *Statistisches Bundesamt*, Bonn
Wiedenbeck, Michael; *Zentrum für Umfragen, Methoden und Analysen (ZUMA)*, Mannheim
Wiegand, Erich; *ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.*, Frankfurt
Wildner, Dr. Raimund; *GfK AG*, Nürnberg
Wilsdorf, Prof. Dr. Steffen H.; *Universität Leipzig*

Z

Zwick, Markus; *Statistisches Bundesamt*, Wiesbaden

Der vorliegende Tagungsband dokumentiert die Beiträge der wissenschaftlichen Tagung „Datenfusion und Datenintegration“, die am 30. Juni und 01. Juli 2005 gemeinsam vom Statistischen Bundesamt, dem ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. und der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI) in Wiesbaden durchgeführt wurde.

InformationsZentrum
Sozialwissenschaften

der Arbeitsgemeinschaft
Sozialwissenschaftlicher Institute e.V.

Lennéstraße 30 • D-53113 Bonn
Telefon 02 28 / 22 81 - 0
Telefax 02 28 / 22 81 - 120

Das IZ ist Mitglied der
Gesellschaft Sozialwissenschaftlicher
Infrastruktureinrichtungen e.V.
Die GESIS ist Mitglied der
Leibniz-Gemeinschaft.

ISBN 3-8206-0148-1
EUR 10,-