Entscheidungsbaumalgorithmen und ihre Anwendung in der Soziologie

Eine empirische Sekundäranalyse von PC-Nutzung am Beispiel von Theodor Geigers Konzepten und der methodische Vergleich von Entscheidungsbäumen, logistischer Regression und Diskriminanzanalyse

Inaugural - Dissertation zur Erlangung des Doktorgrades

der

Philosophisch-

Sozialwissenschaftlichen

Fakultät der

Universität Augsburg

vorgelegt von
Stefan Lebert (geb. Bauer)
aus Nürnberg
2007

Fürth 2006

Erstgutachter:	Prof. Dr. Helmut Giegler
Zweitgutachter:	Prof. Dr. Fritz Böhle
Tag der mündlichen Prüfung:	05. Juni 2007

I N H A L T S V E R Z E I C H N I S

T A B E L L E N V E R Z E I C H N I S

ABBILDUNGSVERZEICHNIS

KAPITEL I EINLEITUNG

Diese Arbeit verfolgt zwei Ziele: zum einen sollen die Ideen und Konzepte eines fast vergessenen soziologischen Klassikers, Theodor GEIGER, den theoretischen Teil begründen. Zum anderen liegt der methodische Schwerpunkt der Arbeit auf einem, in den Sozialwissenschaften kaum genutzten multivariaten Verfahren der Entscheidungsbäume.

GEIGERs Schichtungsbegriff eignet sich - wie diese Arbeit zeigen wird - auch heute noch, soziale Ungleichheit zu beschreiben. Interessanterweise ist er inzwischen - ohne dass sein Name explizit genannt wird - mit seinen Begriffen und Ideen in der Soziologie längst verankert, ohne dass man seinen Namen nennt. Die einzige Studie, die auch heute noch rezipiert wird, ist die 1932 erschienene „Soziale Schichtung des deutschen Volkes", eine Sekundäranalyse der Volkszählung von 1925. Daneben gibt es jedoch eine große Anzahl von Literatur, die sich mit Begriffen wie Individualisierung, Kultur oder Geschmack befassen. Themen, die auch heute noch im Mittelpunkt soziologischen Interesses stehen.

Ziel des Theorieteils ist es, die Konzepte GEIGERs darzustellen und mit den heutigen Sozialstrukturansätzen zu konfrontieren, um dadurch Möglichkeiten und Schwächen herauszuarbeiten. Dabei zeigt sich, dass es durchaus Parallelen zu den meisten heutigen Sozialstrukturansätzen gibt, sei es auf theoretischer (z. B. die deutlichen Gemeinsamkeiten zwischen Habitus bei BOURDIEU und Mentalität bei GEIGER) oder auf empirischer (z. B. die Verortung von Volksmusik bei GEIGER und SCHULZE) Ebene.

Das Erkenntnisinteresse der Arbeit ist es, an einem relativ einfachen Sozialstrukturmodell, das an die Ideen GEIGERs anknüpft, die Methoden der Entscheidungsbäume, die in der Soziologie kaum bekannt sind, praxisorientiert als Fallstudie darzustellen. Es wird weder mit der Arbeit intendiert, ein neues Sozialstrukturmodell zu generieren noch Methoden mathematisch darzustellen.

Vielmehr ist es Ziel, anwendungsorientiert aufzuzeigen, wie die eingesetzten methodischen Verfahren „funktionieren" - und wie sich die Ergebnisse gegebenenfalls unterscheiden. Im Rückgriff auf GEIGER soll gezeigt werden, dass seine Ideen auch heute noch bedenkenswert sind.

Neben der Frage, wie Entscheidungsbäume angewandt werden können, steht die Leistungsfähigkeit der Verfahren - das heißt, ob sie anderen, in der Soziologie gebräuchlichen Methoden wie der Regression oder der Diskriminanzanalyse ebenbürtig sind.

Die forschungsleitende Fragestellung lautet: läßt sich Personal Computer-Nutzung (PC-Nutzung) sozialstrukturell, aber auch durch Kultur- und Freizeitvariablen erklären? - Taugen GEIGERs Ansätze dazu, Themen, an denen er nicht geforscht hat, zu erklären? Sind also die GEIGERschen Theorien und Konzepte heute noch relevant und einsetzbar? - Gerade Konzepte, die über die Zeit hinweg soziale Tatsachen erklären können, sollten nicht unterbewertet bleiben. Somit ist es ein erklärtes Ziel dieser Arbeit, gerade auf weniger bekannte Arbeiten GEIGERs hinzuweisen.

Der quantitativ-methodische Schwerpunkt in Form von Entscheidungsbaumverfahren werden in einigen Wissenschaftsdisziplinen (Marketing, Medizin, etc.) eingesetzt, jedoch kaum in der Soziologie. Es stellt sich die Frage, ob diese Verfahren auch für die Soziologie fruchtbar sind.

Entscheidungsbäume segmentieren Samples. Anhand einer abhängigen Variablen (in dieser Arbeit: PC-Nutzung) und verschiedenen unabhängigen (sozialstrukturellen, Kultur- und Freizeit-) Variablen werden kleinere Gruppen herausgearbeitet, für die bestimmte Merkmale (z. B. bestimmte Berufsgruppen, Einkommen oder Bildungsgrad) aufweisen und sich somit typisieren lassen.

Hierbei geht es nicht um eine Simulation, wie sie BACHER et al. (2004) für die Clusteranalyse vorstellen, sondern um einen anwendungsorientierten

Beitrag zum Verständnis und Einsatz des Verfahrens. Aus diesem Grund wurde ein eher einfaches, aber für den Leser überschaubares Beispiel theoretisches Problem gewählt.

GEIGER hat Anfang der 30er Jahre des letzten Jahrhunderts in seiner Arbeit „Die soziale Schichtung des Deutschen Volkes" die Bevölkerung aufgrund der Berufszählung von 1925 mit einer Art Clusteranalyse typisiert: er ging von den kleinsten beruflichen Einheiten aus und faßte diese immer weiter zu größeren Gruppen zusammen („aszendierendes Verfahren"). Entscheidungsbäume gehen den umgekehrten Weg: von einer Gesamtpopulation, die nach einer bestimmten Variablen „gegliedert" ist, z. B. PC-Nutzer und Nichtnutzer, werden Untergruppen nach sog. unabhängigen Variablen (z. B. Alter, Geschlecht, ...) segmentiert. Voraussetzung ist hierbei, dass diese unabhängigen Variablen bekannt sind - auch hier kommt ein Verfahren GEIGERs, der dominanten und subordinierten Schichten zum Einsatz.

Ich hoffe, mit dieser Arbeit einen Beitrag dazu zu leisten, dass Entscheidungsbaumalgorithmen zusätzlich zu den bereits „bewährten" Verfahren der multivariaten Statistik in die sozialwissenschaftliche Forschung als gleichwertige Verfahren aufgenommen werden.

Ich danke allen, die mich bei dieser Arbeit unterstützt haben. Vor allem danke ich Herrn Prof. Dr. Giegler und Herrn Prof. Dr. Böhle für das Interesse, die Ratschläge und das Engagement, mit dem sie mich bei der Arbeit unterstützten.

Fürth, im Herbst 2006

Stefan Lebert

KAPITEL II THEORETISCHER HINTERGRUND

„Nicht immer ist Neues wichtiger. Oft Alterprobtes sogar neu" (SERNER (1981: 68))

„Es ist ein Fehler, jemanden, den man seinem Milieu entreißen will, zu kompromittieren. Kompromittiere sein Milieu vor ihm." (SERNER (1981: 80))

1 Theodor GEIGER als Ausgangspunkt sozialstruktureller Überlegungen

Wer sich heute mit Sozialstrukturanalyse befaßt, sieht sich mit einer scheinbaren Unübersichtlichkeit verschiedener Theorien, Ansätze und Konzepte konfrontiert. Bei genauerer Betrachtung lassen sich drei „Grundströmungen" erkennen: Ansätze, die sog. „alte", vertikale Ungleichheiten ablehnen und neue Wege der Sozialstrukturanalyse gehen (wie z. B. das SINUS-Institut in Heidelberg oder SCHULZE (vgl. FLAIG et. al. (1997), SCHULZE (1988, 1990, 1992)) und Konzepte, die versuchen, vertikale und neue, horizontale Ungleichheiten zu integrieren (vgl. VESTER (2001)). Der dritte Weg, die theoretisch entwickelten Ideen Ulrich BECKs (vgl. BECK (1983, 1986)), sind zu keinem Modell ausgebaut. Sie haben aber weitreichende Auswirkungen auf die beiden anderen Richtungen bzw. auf die aktuelle soziologische Diskussion.

In dieser Arbeit wird den Ansätzen der Vorzug gegeben, die alte (z. B. Beruf, Bildung, Einkommen) Ungleichheitsdimensionen heranziehen. Allerdings reichen heute - das zeigt die umfangreiche Forschung zu diesem Thema - diese Variablen nicht mehr aus, Sozialstruktur vollständig zu erklären. Sie sind die wichtigsten, jedoch nicht die einzigen Variablen zur Beschreibung unterschiedlicher Lebenslagen.

Daneben werden neue (z. B. Alter, Geschlecht, Freizeit- und Kulturvariablen) Ungleichheitsdimensionen berücksichtigt. Durch die sog. „Bildungsexpansion", Massenarbeitslosigkeit, eine steigende Zahl der Sozialhilfe- bzw. HARTZ IV-Empfänger verschwinden soziale Unterschiede, sog. „alte" Ungleichheiten nicht: sozial schwächeren Menschen wird tagtäglich, beispielsweise durch Werbung, deutlich gemacht, dass es gesellschaftliche Gruppen gibt, die sich finanziell wesentlich mehr leisten können. Auch auf dem Arbeitsmarkt wird Schlechterqualifizierten klar, dass für sie keine oder im besten Fall Stellen mit geringer Kompetenz und geringem Einkommen zur Verfügung stehen. Somit ist soziale Ungleichheit sehr wohl in den Köpfen der Menschen vorhanden - auch wenn sich dies nicht in einer Institutionalisierung wie z. B. in Form einer Arbeitslosen-Partei niederschlägt, sondern in begrenzten Konsum- und Freizeitmöglichkeiten.[1] Viele Ansätze der „Risikobiografie" oder der „Bastelbiografie" (vgl. BECK (1996: 97)) gehen über diese Tatsache (teilweise) hinweg und berücksichtigen (finanzielle) Ungleichheiten nicht. Freizeitaktivitäten und Konsum kosten Geld - und damit sind diese „Lebensstile" immer auch abhängige Variablen der alten Ungleichheiten (Einkommen bzw. Ressourcen).

Bildung ist heute eine grundlegende Voraussetzung für beruflichen Aufstieg. Diese Bildungsmöglichkeiten werden aber z. B. von Kindern un- und angelernter Arbeiter weniger genutzt. Deren Studierendenanteil ist 41mal geringer als die von selbständigen Akademikern (vgl. MEYER (2001b: 256)). Wären sog. „alte" Ungleichheiten keine Schichtungskriterien mehr, wie einige Ansätze behaupten, gäbe es z. B. auch keine intensiven Diskussionen über die Zusammenlegung von

1. Anm.: Viele Lebensstilkonzepte mißachten, dass Freizeitaktivitäten (z. B. Skifahren, Golf spielen, Tennis, etc.) nicht unbeträchtliche finanzielle Kosten nach sich ziehen. Somit ist der Lebensstil nicht frei wählbar oder „inszenierbar", sondern hängt eng mit sog. „alten" Ungleichheiten (Einkommen, Bildung, etc.) zusammen.

Arbeitslosengeld und Sozialhilfe: jeder könnte sich für ca. 350 Euro Regelsatz frei inszenieren, allen Kultur- und Freizeitaktivitäten nachgehen und gemäß seinem Lebens- und Konsumstil sich entfalten.

Andererseits reichen alte Ungleichheitsdimensionen (z. B. Einkommen, berufliche Stellung) heute nicht mehr aus, den Aufbau einer Gesellschaft sinnvoll zu beschreiben. Weitere Dimensionen (z. B. Alter, Geschlecht) können für Auf- oder Abstiege in einer Gesellschaft verantwortlich sein.

Dies ist ein Ansatzpunkt, an dem die Überlegungen des Soziologen Theodor GEIGER ansetzten: welche Schichtungslinien prägen heute hauptsächlich unsere Gesellschaft (= dominante Schichtungen)?

Es ist nicht Gegenstand dieser Arbeit, GEIGERs Sozialstrukturüberlegungen auf die heutige Zeit zu übertragen. Vielmehr sollen in der kritischen Diskussion seiner Gesellschaftskonzepte aktuelle Fragestellungen der PC-Nutzung untersucht werden.

Neue, „bunte" Lebensstilkonzepte mögen zu interessanten Einsichten einer Gesellschaft beitragen[2] - eine alleinige Erklärungskraft besitzen sie aber ebensowenig wie rein vertikale Schichtmodelle.[3]

Die Entwicklung des Sozialstaats mit individueller Absicherung im Krankheits-, Arbeitslosen-, Renten- und Pflegefall hat - zumindest bis in die 80er/90er Jahre des letzten Jahrhunderts - den Lebensstilforschern recht gegeben: durch die gute Absicherung sind vertikale Ungleichheiten etwas in den Hintergrund getreten - obwohl sie stets

2. Anm.: Große sozialstrukturelle „blinde Flecke" sieht MEYER (2001b: 265) vor allem im „unteren Segment der Sozialstruktur": „Feststeht, um es mit Manfred Garhammer zu sagen, dass die LIfestyle-Typologien neben der auffallenden Abstinenz der Geschlechterfrage ihren 'blinden Fleck' im unteren Segment der Sozialstruktur haben."

3. Zu den Problemen der Lebensstilforschung vgl. MEYER (2001b: 259ff.). Zur empirischen Situation in Deutschland vgl. Bundesministerium für Arbeit und Sozialordnung (2001: 76)

vorhanden waren. So stellt z. B. BECK fest, dass sich die Ungleichheitsrelationen erhalten haben - die Lebensbedingungen aller Deutschen hat sich jedoch verbessert („Fahrstuhl-Effekt“).

Auch bei SCHULZEs Erlebnisgesellschaft findet sich dies wieder: Grundannahme SCHULZEs ist die These, dass die Bundesrepublik sich von einer Knappheits- in eine Überflußgesellschaft verwandelt hat (vgl. KONIETZKA (1995: 87)). Kinder werden z. B. nicht mehr hauptsächlich der „Familienfortführung“ oder als Alterssicherung gesehen, sondern sollen ihren Eltern Freude machen.

Es soll hier keine neue Diskussion um „alte“ und „neue“ Ungleichheiten in der Soziologie aufgeworfen werden - Tatsache ist jedoch, dass es auch in den 80er/90er Jahren Teile der Gesellschaft gab, die von der Teilhabe an Prozessen ausgeschlossen waren - der von MEYER weiter oben beschriebene „blinde Fleck“ der Lebensstiltypologien.

In einem ersten Schritt werden die soziologischen Grundhaltungen GEIGERs skizziert und mit heutigen Sozialstrukturansätzen konfrontiert bzw. aktualisiert. Ziel ist die Beantwortung der Frage, wie sich PC-Nutzung sozialstrukturell erklären läßt.

Die Forschungen über Personal Computer-Nutzung zielen häufig darauf ab, Nutzersegmente zu identifizieren, die sich z. B. an bestimmten Anwendungen festmachen lassen - wie „Spieler“, „Downloader“, etc. (vgl. BÜHL (1999)). Erst dann werden die Gruppen - zumeist recht kurz - typisiert.

Diese Arbeit geht den umgekehrten Weg: sie versucht, anhand eines Repräsentativdatensatzes (EUROBAROMETER 56.0 aus dem Jahr 2001) die wichtigsten alten sozialstrukturellen Zusammenhänge (z. B. Bildung, Einkommen, Beruf) zu identifizieren, die in dieser Arbeit weiter als „dominante Schichtungen“ bezeichnet werden. In einem zweiten

Schritt werden Gruppen anhand dieser Kriterien gebildet, die sich weiter mit Kultur- und Freizeitvariablen beschreiben lassen (z. B. Besuch von Kino, Theater, Oper, Musikrichtungen wie Klassik, Pop/Rock, Volksmusik, etc.). Damit wird der Überlegung dieser Arbeit Rechnung getragen, dass Lebensstile niemals isoliert betrachtet werden können, sondern immer ein Spiegel ihres (finanziellen oder kulturellen) Backgrounds sind.

Es wird sich zeigen, ob die verbreiteten (hier bewußt übertriebenen) Klischees („PC-Nutzer sind jünger, männlich, sitzen den ganzen Tag vor dem PC und haben keinerlei Interesse, ihre Wohnung zu verlassen und an Kultur- und Freizeitaktivitäten teilzunehmen“) sich bestätigen oder nicht.

Methodisch wird - neben dem umfangreichen Einsatz der deskriptiven Statistik - multivariat auf die logistische Regression und die Diskriminanzanalyse zurückgegriffen. Diese Verfahren werden in den Sozialwissenschaften häufig eingesetzt.

Im Mittelpunkt des Methodeneinsatzes stehen jedoch Verfahren, die ursprünglich zwar aus den Sozialwissenschaften stammen, jedoch bis heute vor allem von anderen Wissenschaftsdisziplinen (z. B. Informatik, Medizin, Biologie, Marketing) eingesetzt werden: Entscheidungsbäume[4]. Ein erklärtes Hauptziel dieser Arbeit ist es, diese Verfahren in der Soziologie zu etablieren - oder zu verwerfen. Aufgrund bestimmter statistischer Kennzahlen gibt es direkte Vergleichsmöglichkeiten zwischen Logistischer Regression und Diskriminanzanalyse einerseits und den Entscheidungsbäumen andererseits, was den Vergleich deutlich erleichtert. Neben dem - teilweise - grafischen Verfahren der

4. Ein Grund für den mangelnden Einsatz könnte in der bis jetzt relativ teuren Anschaffung der Software liegen (erst ab Version 13 wurden Entscheidungsbaumalgorithmen auch in SPSS - allerdings als teures Zusatzmodul - implementiert, vorher wurde ein „Standalone-Programm“ von SPSS („Answertree“) angeboten - und auch in der fast ausschließlich hochmathematischen Darstellung der Algorithmen.

Entscheidungsbäume sollen weitere, neue Ansätze zur mehrdimensionalen Grafikdarstellung vorgestellt werden.[5]

Somit lassen sich folgende Hauptziele dieser Arbeit formulieren:

- Theoretisch: Lassen sich dominante sozialstrukturelle Schichtungen - und damit auch die Überlegungen GEIGERs - feststellen?'
- Theoretisch: (Wie) Lassen sich PC-Nutzer sozialstrukturell (und „lebensstiltypisch") beschreiben?
- Methodisch: Sind Entscheidungsbäume anderen, „bewährten" Verfahren aus den Sozialwissenschaften ebenbürtig?
- Methodisch: Eignen sich weitere, neu entwickelte grafische Verfahren für die Soziologie?
- Methodisch: Wie klassifizieren Entscheidungsbäume?

Ziel der Arbeit ist es nicht, ein neues Sozialstrukturmodell zu etablieren, sondern den Bereich der PC-Nutzung zu erklären. Deshalb sind die gefundenen dominanten Merkmale keine Schichtungsmerkmale für die Gesamtgesellschaft, sondern nur für diese Art der Techniknutzung - auch wenn sich sicherlich Parallelen finden lassen.

Eine größere Rolle kommt allerdings dem Methodeneinsatz zu: nachdem Entscheidungsbäume nicht als Standardverfahren in den Sozialwissenschaften eingesetzt werden, wird mit dieser Arbeit überprüft, ob sich diese Verfahren überhaupt für diese Wissenschaftsdisziplin eignen.

1.1 Soziologische Grundhaltungen GEIGERs

Theodor GEIGER (1891 - 1952) ist eher ein „kleiner Klassiker" der Soziologie (neben WEBER oder auch MARX). Sein bekanntestes Werk in der Soziologie ist „Die Soziale Schichtung des Deutschen Volkes", das 1932 erschien und als herausragende Sekundäranalyse seiner Zeit gilt.[6]

5. Anm.: Rainer SCHNELL hat sich ausführlich mit Grafikdarstellung in seinem Buch „Graphisch gestützte Datenanalyse" (1999) befaßt. Dort sind auch Grundlagen zu den eingesetzten Verfahren der Parallel-, Spine- und Mosaic-Plots zu finden.

In den frühen Arbeiten GEIGERs, die kurz nach dem Ersten Weltkrieg entstanden, werden zwei Grundzüge seines Denkens deutlich, die sich durch seine gesamte spätere soziologische Laufbahn ziehen: zum einen ein starker empirischer (quantitativer) Bezug, zum anderen ein Blick für soziale Ungleichheiten.

Anders als z. B. die Frankfurter Schule, die eher kulturpessimistisch ist, sieht GEIGER in einer empirischen „Aufklärungsarbeit", die sich Werturteilen zu enthalten hat, einen Fortschritt der Gesellschaft.

> „Vielleicht, lieber Leser, werden Sie hinter vieler Schärfe, manchem hart und kalt klingendem Worte ahnen, dass es mir darum geht, die Sache des Menschen zu führen, des ewig getretenen, gequälten, geschändeten. Ist es nicht endlich an der Zeit, ihn aus der Knechtschaft der Ismen und Systeme zu befreien und - leben zu lassen?" (GEIGER (1964: 7))

MEYER bemerkt hierzu:

> „Mit der Aufklärungslehre des Intellektuellen Humanismus versucht Geiger ... sein Wunschbild einer 'kritisch aufgeklärten Gesellschaft' ... auf den Begriff zu bringen." (MEYER (2001a: 220))

Ziel ist es für GEIGER, wie MEYER (vgl. MEYER (2001a: 220)) formuliert, Antworten zu finden, ob demokratisches Zusammenleben ermöglicht und gestärkt werden kann, damit es zu keinem Rückschritt in totalitäre Systeme (z. B. Nationalsozialismus) kommt: „Kritische Aufklärung des Menschen, ihre Intellektualisierung", so hofft GEIGER, „soll der Ideologie den Resonanzboden entziehen" (RODAX (1991: 42))[7]. Und:

> „Der Kitt, der die heutige Grossgesellschaft zusammenhalten kann, ist nicht ein Gemeinschaftsgefühl zwischen Mensch und Mensch *persönlich*, sondern die *sachliche* Einordnung in einen gemeinsa-

6. Zum Lebenslauf GEIGERs vgl. BACHMANN (1995), RODAX (1991), auch TRAPPE (1978, 1993). GEIGERs Lehrbuch der Soziologie, das 1939 im Exil entstand, ist nur in einer Rohübersetzung ins Deutsche übertragen (vgl. GEIGER (1939))

7. Somit steht GEIGER eher in der Tradition der späteren „Kölner Schule". Er präferiert quantitative Sozialforschung, wobei die qualitative Forschung bis zu GEIGERs Tod sich erst langsam etabliert - somit weiß man nicht, ob GEIGER nicht auch qualitativer Forschung aufgeschlossen wäre. Im Gegensatz zu vielen seiner Kollegen der 20er und 30er Jahre hält er (quantitative) Empirie jedoch für wichtig.

> men Daseinsrahmen. Der aufs höchste rationalisierte Lebenszuschnitt des technischen Zeitalters ist auf die Dauer nur möglich, wenn die Menschen selbst sich in ihren politisch-wirtschaftlichen Beziehungen mehr von intellektuellen als von sentimentalen Antrieben leiten lassen. Sie auf diesen Weg zu bringen, ist die wichtigste Aufgabe der Volkserziehung in der Gesellschaft von heute. Wo mit Kollektiv*gefühlen* geladene Menschenmassen und moderne Grosstechnik einander begegnen, dort lauert soziales Chaos oder brutale Diktatur am nächsten Kreuzweg." (GEIGER (1955: 79))

Bereits in seiner Hochschullehrertätigkeit in Braunschweig Ende der 20er Jahre wird seine oben erwähnte Haltung zu Werturteilsfragen deutlich:

> „Er ergreift - bei aller sonst geübter Zurückhaltung in Werturteilsfragen in der Lehre - gegen die früh als verhängnisvoll erkannte nationalsozialistische Bewegung entschieden Partei. In seinen Lehrveranstaltungen reagiert er auf die Agitation nationalsozialistischer Studenten sehr gelassen und erklärt in aller Ruhe, er glaube nicht, dass die völkische Idee noch eine Chance habe, weil die Menschen frei sein wollten." (RODAX (1991: 92))

Werturteile haben - GEIGER wird sich besonders in seinen späten Publikationen damit beschäftigen - in der Soziologie bzw. in der Wissenschaft nichts verloren - auch wenn man sich als Forscher dem nicht völlig entziehen kann. Hier zählen Fakten, gestützt durch Empirie, aber nicht zuletzt auch durch Theorie. Diese Überzeugung versucht er auch, auf die Sozialstrukturanalyse zu übertragen. Er arbeitet in der Weimarer Republik an seinem Begriff der Mentalität:

> „[Mentalität ist] ... geistig-seelische Disposition, ist unmittelbare Prägung des Menschen durch seine soziale Lebenswelt und die von ihr ausstrahlenden, an ihr gemachten Lebenserfahrungen." (GEIGER (1987: 77))

Der Mentalitätsbegriff eröffnet ihm wesentlich flexiblere Zugänge zur Sozialstruktur als der MARX´sche Klassenbegriff: Mit dem Begriff der Schichtung, den er neu belebt und dem Mentalitätsbegriff umgeht er den Klassenbegriff und dessen eher deterministische Deutung von sozialer Stellung und Bewußtsein. Seine Kritik gegen MARX findet sich an vielen Stellen seines Werkes (u. a. in der „Klassengesellschaft im Schmelztiegel"). Er spottet:

> „Die klassenlose Gesellschaft soll das letzte Wort der geschichtlichen Entwicklung sein. Eine solche Behauptung kann nur der aufstellen, der mit Sinn und Endziel der Weltgeschichte vertraut ist." (GEIGER (1948/49, S. 38))

Dabei ist für GEIGER der Schichtbegriff nicht ein unkritisches Pendant zum Klassenbegriff, sondern vielmehr eine Möglichkeit, mit einem nicht-deterministischen, „unbelasteten" und vor allem zeitgemäßen Begriff an das Problem der sozialen Strukturierung heranzugehen.

GEIGER nähert sich dem Begriff der Schicht folgendermaßen an:

> „Wir können einen Begriff der Schicht bilden, der beinahe ohne Urteilsinhalt ist. Definieren wir als *Schicht die Gesamtheit der Personen innerhalb einer Bevölkerung, denen irgendein Merkmal gemein ist,* so ist die Feststellung der Schichtstruktur eine Aufgabe der bloßen Merkmalsbestimmung, der Identifizierung von Merkmalsträgern und endlich der Zählung. Das ist Sozialstatistik, hat aber mit Soziologie wenig zu tun. Der Wahl von Merkmalen (Merkmalsreihen) für die Klassifikation einer Bevölkerung sind grundsätzlich keine Grenzen gezogen [sic!]. Man kann die gleiche Bevölkerung nach Belieben kreuz und quer klassifizieren - und tut das tatsächlich in hundert verschiedenen Frageabsichten. Nach Geschlecht und Alter, Beruf, Stellung und Einkommen, ..." (GEIGER (1962b: 189f.))

Er koppelt die oben beschriebenen Soziallagen an die „sozialen Haltungen, Willensrichtungen, Bewegungen, usw." und versucht, diese „Mentalitäten" bestimmten Soziallagen zuzuordnen (vgl. GEIGER (1962b: 194)). Hieraus erkennt er, dass die Soziallage nicht alleine ausschlaggebend ist (wie bei MARX), sondern dass neben die Soziallage Mentalitätszüge treten können, die sich widersprechen. So stellt er fest, dass die Einkommenshöhe der „Lohnproletarier" nicht wesentlich von den „Tagwerkern für eigene Rechnung" differiert. Somit könnte man sie einer Klasse (oder Schicht) zuordnen. Vom Eigentumsgedanken (Mentalitätsaspekt) hingegen unterscheiden sich die beiden Gruppen wesentlich (vgl. GEIGER (1987: 107f.)).

Es wird weiter deutlich, dass Soziallage und Mentalität nicht in einem deterministischen (wie z. B. bei MARX), sondern eher in einem statisti-

schen Zusammenhang stehen, d. h., dass bestimmte Mentalitäten in bestimmten Soziallagen vermehrt auftreten:

> „Wie wenig daraus [aus der Sozialstatistik, Anm. S. L.] für die Einsicht in die psychische Lage des Industriearbeiters zu gewinnen ist, zeigt die persönliche Bekanntschaft mit zwei beliebigen, im gleichen Betrieb, an gleicher Arbeitsstelle, seit längerer Zeit nebeneinander tätigen Arbeitern. Gemein ist ihnen die äußere Situation ... Was sich nach statistischer, das heißt quantenmessender Methode ermitteln läßt, ist der generelle und unpersönliche Rahmen, innerhalb dessen menschliche Schicksale und Leben ablaufen.“ (GEIGER (1962a: 154))

GEIGER ist nicht der positivistische Forscher, wie er von einigen Kollegen gesehen wurde: zwar hat die Empirie bei ihm einen hohen Stellenwert, wird aber durch Theorie gestützt. Seine Aussagen - z. B. zum Zusammenhang zwischen Mentalität und Schicht - sind niemals deterministisch, sondern dynamisch und offen, so dass man seine Überlegungen auch heute noch für sozialstrukturelle Untersuchungen heranziehen kann.

Betrachtet man verschiedene Studien GEIGERs, wird die Dynamik seines Vorgehens offensichtlich. Zieht er 1932 bei seiner Studie „Die Soziale Schichtung des Deutschen Volkes“ noch Berufs- und Mentalitätskriterien zur Erklärung der Sozialstruktur heran[8], schlägt er zwanzig Jahre später ein vierdimensionales Modell, bestehend aus Wirtschaftszweig, beruflicher Stellung, Einkommenshöhe und Bildungsgrad vor (vgl. GEIGER (1962b: 196)). Als Begründung für die unterschiedlichen Schichtungsdeterminanten führt GEIGER an, dass die Gesellschaft der 40er/50er Jahre eine andere als die der 20er Jahre ist und andere „dominante Schichtungen“ erfordert.

> „Fast jede Gesellschaft ist in mehrfachen Richtungen geschichtet. Von diesen Schichtungen sind einige von untergeordneter Bedeu-

8. Das von GEIGER entwickelte „aszendierende Verfahren“ erinnert stark an eine Clusteranalyse. Er geht von den kleinsten Merkmalsträgern (in diesem Fall: Berufe) aus und faßt diese bei Ähnlichkeit zusammen. Kommen neue Merkmalsträger hinzu, werden neue Kategorien gebildet bzw. die bereits bestehenden Kategorien überprüft.

> tung (s u b o r d i n i e r t e Schichtungen), andere aber entscheidend für die Sozialstruktur (d o m i n a n t e Schichtung)." (GEIGER (1948/49: 45))

Die Veränderungen der Sozialstruktur bezeichnet GEIGER als „Umschichtung":

> „Der Begriff der Umschichtung bezielt [sic!] die Änderung der Sozialstruktur selbst, sei es, dass man nur an Verlagerung im Verhältnis bestehender Schichten zueinander, sei es, dass man an eine radikale Umgruppierung der Bevölkerung nach anderen Schichtungsmerkmalen denkt." (GEIGER (1962c: 139))

Als weiteren Mobilitätsbegriff führt GEIGER den Begriff der „Fluktuation" ein, der sich heute mit dem Begriff der „sozialen Mobilität" gleichsetzen läßt (vgl. GEIGER (1962c: 114), BERGER (1998: 574ff.)). Damit tritt GEIGER für einen mehrdimensionalen, dynamischen Schichtungsbegriff ein.

Mit diesem Instrumentarium hat GEIGER ein differenziertes Modell sozialer Schichtung erarbeitet, das weit über die meisten Konzepte der Nachkriegszeit hinausgeht. Als Empiriker ist ihm klar, dass nicht alle Variablen, die zur Schichtung beitragen, berücksichtigt werden können, sondern nur die wichtigsten.

Hätte sich die deutsche Sozialstrukturanalyse der 60er und 70er Jahre an diese Maxime gehalten, wären die Begriffe der „Klasse" oder „Schicht" heute nicht derart diskreditiert. Neue „dominante" Schichtungen wie Alter oder Geschlecht hätten erkannt und umgesetzt werden müssen. So kommt es Ende der 70er Jahre eher zu einem „Kontinuitätsbruch": Klassen- und Schichtbegriffe stehen Lebensstilkonzepte gegenüber.

Das Schichtkonzept GEIGERs beinhaltet zusammenfassend:

- einen nicht-deterministischen Schichtbegriff
- einen stark empirischen Bezug, gestützt durch Theorie
- „dominante" und „subordinierte" Schichtungsmerkmale, die immer wieder für Gesellschaften und unterschiedliche Zeiten zu überprüfen sind,
- die Ablehnung von „Sozialstatistik", also Gruppenbildung reiner Merkmalsträger
- die Verknüpfung von sozialer Lage und Mentalität
- soziale Dynamisierungsprozesse („soziale Mobilität" und „Umschichtungen")

GEIGER war ein empirischer Soziologe, der versucht hat, auf theoretischer Grundlage möglichst vorurteilslos und wertfrei soziale Tatsachen zu erforschen. Für die Untersuchung der Sozialstruktur hat er ein Instrument vorgelegt, das heute modifiziert als Grundlage für Ungleichheitsanalysen herangezogen werden kann. Dies wird am Ende des Kapitels gezeigt. SCHROTH bemerkt:

> „In der Geigerschen Schichtungstheorie sind daher alle Aspekte der heutigen Diskussion bereits angesprochen. Geigers mehrdimensionaler Ansatz läßt, verglichen mit der heutigen Diskussion, einen umfassenderen Blickwinkel der Untersuchung der Sozialstruktur zu, gerade auch durch den Anspruch, den Zusammenhang zwischen sozialer Lage und Mentalität in den Mittelpunkt einer Schichtanalyse zu stellen. ... Die dynamische Sichtweise Geigers und der Begriff der Umschichtung belegen, dass dieses Modell der sozialen Schichtung offen für neue Formen sozialer Ungleichheit ist." (SCHROTH (1999: 36))

Diese „neuen Formen sozialer Ungleichheit" werden mit der vorliegenden Arbeit näher ausgeleuchtet. Halten GEIGERs Überlegungen einer Konfrontation mit heutigen Sozialstrukturmodellen stand oder sind seine Konzepte den heutigen Sozialstrukturansätzen sogar überlegen?

Bei einem Rückgriff auf Klassiker stellt sich stets die Frage, warum nicht aktuellere, der jeweiligen Zeit entsprechendere theoretische Ansätze für die Erklärung eines Problems herangezogen werden. In den nächsten Unterkapiteln werden deshalb die Unterschiede, aber auch die Gemeinsamkeiten zwischen ausgewählten heutigen Kon-

zepten und GEIGERs Überlegungen herausgearbeitet. Dies geschieht aber nicht mit dem Ziel, GEIGERs Ideen als den heutigen Ansätzen weit üblegene Theorie zu charakterisieren, sondern aufzuzeigen, dass GEIGERs Konzepte sehr aktuell - und in Teilen mit heutigen Theorien identisch sind. Zum anderen wird dargelegt, wie sehr sich Begriffe - wie z. B. der Schicht - von der ursprünglichen Definition GEIGERs entfernt haben. Eine Reformulierung scheint hier sinnvoll - und somit auch die Anknüpfung an das Original, den Urheber des modernen Schichtbegriffs.

1.2 GEIGERs Überlegungen zur „Individualisierung“

Aufgrund der umfangreichen Diskussion um den Individualisierungsbegriff von BECK (1983, 1986) soll kurz auf die Überlegungen GEIGERs zu diesem Thema zurückgegriffen werden. Der Individualisierungsbegriff ist nicht neu, er geht auf SIMMEL zurück. GEIGER bemerkt:

> „Darunter verstehe ich gewiß nicht krassen Egoismus oder ein Sichloslösen aus mitmenschlicher Gemeinschaft, sondern eine Steigerung der inneren Selbständigkeit. Eine Individualisierung in diesem Sinne ist in der Tat mit der zunehmenden Ausgliederung der Gesellschaft einhergegangen. Der Einzelne ist nicht mehr ein untergeordnetes Glied quasi-substanzieller gesellschaftlicher Ganzheiten, die ihn umfangen und verschlingen und durch deren Vermittlung allein er seinen Platz unter den Mitmenschen hat. Heute steht er weithin freizügig zwischen gesellschaftlichen Lebenskreisen oder in deren Schnittpunkt Er ist ihr Mitglied, nicht ein Teilchen von ihnen, er hat Anteil an ihnen, ohne an sie gefesselt zu sein. ... Allein schon diese schwebende Lage im grenzenlosen Gefüge gesellschaftlicher Beziehungen bedeutet eine soziale Freisetzung des Einers.“ (GEIGER (1964: 128))

Diese Komponenten - vor allem der Begriff der Freisetzung - findet sich später bei BECK wieder. Menschen sind - nach GEIGER - bereits in den 50er Jahren nicht mehr so stark an Familie oder Kirche gebunden - und können ihren eigenen Lebensweg festlegen (vgl. GEIGER (1964: 65)).

Zwar finden sich die Ideen GEIGERs bei BECK wieder - aber weder in der „Risikogesellschaft“ noch in seinem Aufsatz im Sammelband „Soziale Ungleichheiten“ existiert ein Literaturhinweis auf diese Quelle: in ersterem wird lediglich von GEIGER die „Klassengesellschaft im Schmelztiegel“ erwähnt, aber nicht „Demokratie ohne Dogma“ aus der obige Zitate stammen. Dies ist umso verwunderlicher, weil BECK sich explizit auf Klassiker bezieht - MARX und WEBER (vgl. BECK (1986: 130ff.)). Aus dieser Übereinstimmungen zwischen GEIGER und BECK scheint es sinnvoll, die Aussagen beider zu kontrastieren:

TABELLE 1 INDIVIDUALISIERUNG: GEMEINSAMKEITEN UND UNTERSCHIEDE ZWISCHEN GEIGER (1964) UND BECK (1986)

GEIGER	BECK
„Aber der einzelne wird innerhalb dieser gesellschaftlichen Zusammenhänge die innere Freiheit seiner Persönlichkeit wahren lernen. Er wird mit anderen zusammen leben und wirken, ohne seine eigenen Meinungen und Wertungen aufzuopfern und ohne von den andern [sic!] zu erwarten, dass sie denken und werten sollen wie er.“ (S. 130)	„Auf dem Hintergrund eines vergleichsweise hohen materiellen Lebensstandards und weit vorangetriebenen sozialen Sicherheiten werden die Menschen in einem historischen Kontinuitätsbruch aus traditionalen Klassenbedingungen und Versorgungsbezügen der Familie herausgelöst und verstärkt auf sich selbst und ihr individuelles Arbeitsmarktschicksal mit allen Risiken, Chancen und Widersprüchen verwiesen.“ (S. 116)
„Schwieriger freilich ist es, der angeborenen Klassenlage zu entrinnen. Aber auch hier sind die Hindernisse nur noch wirtschaftlich-faktischer, nicht mehr institutionell-rechtlicher Art. Armer Leute Kind kann nicht beschließen, Kapitalist zu werden. Davon abgesehen aber bietet eine Gesellschaft mit ziemlich unbehindertem Zugang zu höherer Ausbildung reiche Möglichkeit der freien Wahl einer Berufslaufbahn, damit aber eines Platzwechels im Klassenaufbau. Zudem ist die schicksalhafte Bedeutung der Klassenlage sichtlich im Schwinden begriffen ..“ (S. 65)	„Diese Tendenz zur 'Klassenlosigkeit' sozialer Ungleichheit tritt exemplarisch in der Verteilung der Massenarbeitslosigkeit hervor.“ (S. 117) Gesellschaftliche Krisen erscheinen als persönliche Krisen (vgl. S. 118)

TABELLE 1 INDIVIDUALISIERUNG: GEMEINSAMKEITEN UND UNTERSCHIEDE ZWISCHEN GEIGER (1964) UND BECK (1986)

GEIGER	BECK
„Alleine schon die schwebende Lage im grenzenlosen Gefüge gesellschaftlicher Beziehungen bedeutet eine soziale Freisetzung des Einers. Glaubensinhalte, Meinungen, Werturteile sind ihm nicht länger durch Überlieferung oder Autorität der Gruppen, die ihn umfangen, vorgeschrieben.“ (S. 128)	„Auf der anderen Seite tritt für das Handeln der Menschen die Bindung an soziale Klassen eigentümlich in den Hintergrund. Ständisch geprägte Sozialmilieus und klassenkulturelle Lebensformen verblassen“ (S. 116)
„Das Familienleben z. B. kann - unter im übrigen günstigen Bedingungen - heute viel intimer gestaltet werden als unter der patriarchalischen Ordnung. Teils schon deshalb, weil der Familienkreis enger ist, nur noch Eltern und unerwachsene Kinder umfaßt. Vor allem aber, weil die Familie nur noch Heim ist, keinerlei nach außen gerichtete Funktionen mehr hat. Der Familienvater ist nicht mehr Autorität und Obrigkeit, ein kleiner Despot zwischen seinen vier Wänden. Die Ehegatten sind Kameraden auf gleichem Fuße ... Liebe ist an die Stelle der Autorität, Vertraulichkeit an Stelle von Unterwürfigkeit getreten.“ (S. 64)	„Diese Freisetzung ... wird überlagert durch eine Freisetzung relativ zu *Geschlechtslagen*.“ (S. 118)
	„Mit der Durchsetzung der Industriegesellschaft wird insofern immer schon die Aufhebung ihrer Familienmoral, ihrer Geschlechtsschicksale, ihrer Tabus von Ehe, Elternschaft und Sexualität, ja die Wiedervereinigung von Haus- und Erwerbsarbeit betrieben. ... *Der oder die einzelne selbst wird zur lebensweltlichen Reproduktionseinheit des Sozialen*“ (S. 118)
	„... die Individuen werden innerhalb und außerhalb der Familie zum Akteur ihrer marktvermittelten Existenzsicherung und der darauf bezogenen Biographieplanung und -organisation.“ (S. 119)
„Die Individualisierung, von der hier die Rede ist, hat, wie man sieht, nichts mit einer Lösung der gesellschaftlichen Bande zu tun. Die Gesellschaft als ein Gefüge des Zusammenlebens und Zusammenwirkens bleibt davon unberührt.“ (S. 130)	„Individualisierung wird dementsprechend hier als ein historisch widersprüchlicher *Prozeß der Vergesellschaftung* verstanden.“ (S. 119)

Hier zeigen sich sowohl interessante Parallelen als auch Unterschiede: während GEIGER Anfang der 50er Jahre die Lockerung der sozialen Beziehungen bzw. der sozialen Herkunft begrüßt, weist BECK (1986) richtigerweise eher auf die Probleme hin (Stichworte: Frauenarbeitslosigkeit, „Individuelles Schicksal“). Während die Rollenverteilung der 50er Jahre überwiegend klar definiert war, ist das in den 80er Jahren nicht mehr so: verschiedene „Konzepte“ von Single-Dasein,

Lebenspartnerschaften (homosexuell und heterosexuell, verheiratet und nicht verheiratet, geschieden), Wohngemeinschaften, etc. koexistieren nebeneinander. Während es GEIGER in den 50er Jahren noch leicht fällt, Familie als Kernfamilie (Eltern und ihre nicht erwachsenen Kinder) zu definieren, fällt es heute schwer, diesen Begriff zu fassen. Zwar wandelten sich die Paarbeziehungen nach GEIGER vom Despotismus zu Gleichberechtigung, an Scheidung war jedoch nicht zu denken. Die Themen Arbeitslosigkeit und Emanzipation hatten in den 50er Jahren keine Bedeutung. Durch den Wiederaufbau und das „Wirtschaftswunder" gab es keine hohen Arbeitslosenzahlen; dieses Problem taucht erstmals Ende der 70er/Anfang der 80er Jahre in der Bundesrepublik auf.

Mit Ausnahme dieser Punkte, die zeitliche bzw. historische Aussagen wiederspiegeln, klingen die Aussagen BECKs und GEIGERs seltsam gleich; während GEIGER die Freisetzung der Menschen aus dem Blickwinkel der 50er Jahre eher positiv begrüßt, sieht BECK - aus dem Blickwinkel der 80er Jahre - größere gesellschaftliche Risiken. Aber auch das ist nicht neu und wurde bereits in den 60er Jahren von BAUDRILLARD festgestellt:

> „Auf die gleiche Weise hat auch die bürgerliche und industrielle Revolution, Schritt für Schritt, das Individuum von religiösen, moralischen, familiären Implikationen befreit und der einzelne erlangte als Mensch die rechtliche, als Arbeitskraft aber die faktische Freiheit, das heißt, die Freiheit, sich als solche zu verkaufen." (BAUDRILLARD (2001: 27))

Zudem halten die Thesen BECKs einer empirischen Überprüfung nicht immer stand. Problematisch an seinen Überlegungen ist, dass empirische Belege aus Globalstatistiken stammen, die bei näherem Betrachten empirische Probleme aufwerfen. So ist für BECK ein Beleg für die „zweite Moderne" (also die Zeit jenseits der „ersten Moderne", d. h. der Industriegesellschaft) dass heute Menschen aus ihren traditionellen Strukturen herausgelöst werden (vgl. BECK (1986: 115)). Seine

Argumentation zielt darauf ab, dass z. B. Studierende beim Studienbeginn viele neue Erfahrungen sammeln, die - vereinfacht ausgedrückt - die alten Vergemeinschaftungen (Herkunftsfamilie) überlagern, da die Studienorte weit vom Heimatort entfernt sind und somit auch die Kontakte keine Rolle mehr spielen.

Zieht man die 14. Erhebung des Deutschen Studentenwerks (1995: 383f.) heran, so zeigt sich, dass sich sowohl in den alten als auch in den neuen Bundesländern 24 % der Studierenden am Heimatort immatrikulieren. Zwar ziehen 3/4 der Studierenden vom Heimatort weg, aber:

> „Gleichwohl in welcher Wohnform sie dann am Studienort leben, bleiben mehr oder weniger starke Bindungen an den Heimatort bestehen. Für die Studierende [sic!] ist es deshalb geradezu charakteristisch, dass sie auch während der Semesterzeit vorwiegend übers Wochenende mehr oder weniger häufig vom Hochschulort zu Eltern oder Partner in den Herkunftsort fahren." (Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (1995: 382))

Der Anteil der „Elternwohner" ist seit Mitte der 70er Jahre sowohl für Uni- als auch FH-Studierende sehr stabil (vgl. Bundesminister für Bildung und Wissenschaft (1989: 342)) und weit davon entfernt, so dass die Aussage BECKs („Auszug aus dem Elternhaus löst Sozialisationserfahrungen") für einen Großteil der Studierenden zutreffend ist.

Ein zweites Beispiel: eine Grundannahme BECKs ist die Herauslösung aus traditionellen Zusammenhängen. Ein Indiz dafür ist die zunehmende Zahl der Einpersonen-Haushalte, die von 1961 (ca. 20 %) bis 1997 auf etwa 35 % anstieg (vgl. HRADIL (1998: 9)). Geht man allerdings - wie z. B. HRADIL - der Frage nach, was Singles ausmachen, so kommt dieser zu folgender Definition. Singles

- leben in mittlerem Lebensalter (25 - 55 Jahre)
- haben keinen festen Partner und
- geben an, alleine leben zu wollen (vgl. HRADIL (1998: 10))

Nach HRADIL leben „höchstens 3 Prozent der erwachsenen Bevölkerung“ (HRADIL (1998: 10)) als Single.

Damit wird die Bedeutung von Singles in Deutschland stark überschätzt. Die Differenz zwischen 35 % Single-Haushalten und 3 % eines „harten Kerns“, wie ihn HRADIL beschreibt, ergibt sich aus der Definition: entweder sind Singles verwitwet oder jünger als 25 Jahre (z. B. Studierende oder Auszubildende mit erster Wohnung). Die Frage der Partnerschaft bleibt bei BECK ebenso ungeklärt wie die Neigung, bewußt alleine leben zu wollen (Partner können in anderen Wohnungen leben, Menschen sind auf Partnersuche, ...).

Dies offenbart einen weiteren Unterschied zwischen GEIGER und BECK: während GEIGER versucht, durch den Mentalitätsbegriff menschliches Handeln transparenter zu machen, tritt dieses Moment bei BECK eigentümlich in den Hintergrund. An mehreren Stellen seines Werkes beschreibt er Folgen bestimmter gesellschaftlicher Auswirkungen auf bestimmte gesellschaftliche Gruppen. Die Frage, „was mit den Menschen passiert“, beantwortet BECK allerdings nicht, sondern geht zu einem anderen Thema über. LASH bemerkt:

> „Ich möchte einen Schritt zurücktreten und jene Frage stellen, die weder Beck noch Giddens mit ausreichendem Nachdruck stellen: Warum, so könnte man fragen, findet sich Reflexivität in manchen Bereichen und in anderen nicht? ... Gibt es vielleicht neben den erwähnten 'Reflexivitätsgewinnern' ganze Batallione von 'Reflexivitäts*verlierern*' in den heutigen Informationsgesellschaften, die zunehmend von Klassengegensätzen geprägt sind, denen jedoch jedes Klassenbewußtsein fehlt?“ (LASH (1996: 210))

Somit kommt BECK zu keiner abschließenden Klassifikation wie GEIGER oder auch VESTER („Modernisierungsgewinner“ - „Modernisierungsverlierer“), was sicherlich auch darauf zurückzuführen ist, dass BECK unsere Gesellschaft auf den Weg in die „Zweite Moderne“ sieht, in der wir aber noch nicht angekommen sind. Trotzdem ist die Frage, „was passiert mit den Menschen?“, aus meiner Sicht einer der

wichtigsten Fragen der Sozialstrukturanalyse (auch der Soziologie), denn es geht nicht um abstrakte oder theoretische Gebilde, sondern um eine realitätsgerechte Beschreibung menschlichen Zusammenlebens.

Geht man noch einen Schritt weiter und vergleicht die Aussagen BECKs mit denen LYOTARDs findet sich kaum etwas Neues, auch nichts Originelles in BECKs Werk:

> „Die Neuerung besteht darin, dass in diesem Zusammenhang die alten Attraktionspole, die Nationalstaaten, Parteien, Berufsverbände, Institutionen und historischen Traditionen, an Anziehungskraft verlieren. Sie müssen anscheinend auch nicht ersetzt werden, zumindest nicht auf dieser Ebene." (LYOTARD (1986: 53))

BECK formuliert das folgendermaßen:

> „Die gesellschaftlichen Institutionen - politische Parteien, Gewerkschaften, Regierungen, Sozialämter usw. werden zu *Konservatoren einer sozialen Wirklichkeit, die es immer weniger gibt*. (BECK (1986: 158))

Wo liegen hier noch die Unterschiede, wo läßt sich bei BECK etwas Neues entdecken? - Die Schlussfolgerungen BECKs sind marginal im Gegensatz zu den eben aufgezeigten Parallelen zu GEIGER, BAUDRILLARD und LYOTARD.

Die empirischen Belege zeigen, dass ein Ungleichheitsbegriff, wie ihn BECK anwendet, empirisch problematisch ist. Die Diskussion um den Sinn von Klassen- und Schichtkonzepten, die BECK und andere lostraten und die zu Lebensstilkonzepten führten, ist ebenfalls nur eingeschränkt nutzbar. SCHULZE greift z. B. bei seiner Milieubildung auf klassische eindimensionale Merkmale von Alter und Bildung zurück, die - würde man sie als Schichtanalyse begreifen - weit hinter GEIGER zurückfällt (siehe unten).

Ein weiterer Kritikpunkt stammt von VESTER: VESTERs Kritik gegen rein individualistische, Klassen- und Schichten leugnende Ansätze wie z.

B. die von BECK und SCHULZE zielt in die Richtung, dass er Individualisierung oder Pluralisierung nicht in Abrede stellt: er macht aber deutlich, dass diese - von allen Autoren nicht unkritisch gesehenen - neuen Möglichkeiten nicht in allen Teilen der Gesellschaft gleich wirken:

> „Unsere Grundhypothese war die der *Pluralisierung der Klassenmilieus,* d. h. wir nahmen an, dass die genannten Tendenzen der Individualisierung und Pluralisierung nicht überall gleich wirkten, sondern im Zusammenspiel mit gesellschafts- und lebensgeschichtlich bereits erworbenen und tradierten Gruppenidentitäten.“ (VESTER (1994: 137)).

Die Argumentation, dass - vereinfacht - unterschiedliche Sozialisationsfaktoren zu unterschiedlichen Mentalitäten und Chancen führen, ist leicht nachvollziehbar und in der Literatur auf breiter Ebene erforscht (vgl. u. a. HURRELMANN (1998), GEISSLER (1996)). Dieser Aspekt bleibt bei BECK weitgehend unberücksichtigt.

1.3 GEIGER und BOURDIEU - Mentalität und Habitus am Beispiel des Musikgeschmacks

Neben den vielen Veröffentlichungen GEIGERs zur Sozialstruktur gibt es einen kleinen, relativ unbekannten Aufsatz, in dem er ein Experiment zum Musikgeschmack beschreibt. Die Ergebnisse dieses Experiments sind - zumindest in deutschsprachigen Veröffentlichungen - zu keiner vollständigen Theorie ausgebaut worden. Sie widersprechen aber sowohl den Erkenntnissen der „Frankfurter Schule“ (z. B. ADORNOs Arbeiten zur Musiksoziologie) als auch den Ergebnissen von SCHULZE und BOURDIEU.

GEIGERs Überlegungen zielen darauf ab, dass jeder Mensch schon einmal mit klassischer Musik in Berührung gekommen ist:

> „Without exaggeration it can be said that in Denmark some kind of public prejudice exists to the effect that 'classical music is not for us, the plain people'. In consequence, a considerable number of radio listeners are fighting shy of classical music, admittedly without ever having given it a fair trial. The public attitude may be characterized

> by the fact that the question 'Have you ever heard a composition by Mozart?' is sometimes emphatically answered in the negative - which simply cannot be true. The respondent has in all probability heard classical music frequently, and enjoyed it, without being aware of its classical nature." (GEIGER (1950: 454))

Daraus schlußfolgert er, dass auch die breite Öffentlichkeit klassische Musik hören würde - wäre sie nicht mit einer - wie BOURDIEU sagen würde - „Abgrenzung nach unten" verbunden.

Um diese Hypothese zu testen, startete GEIGER mit dem Dänischen Rundfunk ein Experiment: er ließ an zwei aufeinanderfolgenden Samstagen zur Hauptsendezeit (19:40 Uhr), zu dem ausschließlich klassische Musik gesendet wurde, das gleiche musikalische Programm ausstrahlen. Dies bestand aus Werken von Haydn, Schubert, Mozart, Beethoven und Mendelssohn-Bartholdy. Am ersten Samstag wurden die Musikstücke ohne Ankündigung (Titel, Komponisten, Musiker) ausgestrahlt - und als „Populärmusik" bezeichnet, am darauffolgenden Samstag mit allen Angaben zu Komponisten und Musik („Klassische Musik").

GEIGER ermittelte die „Reichweiten" dieses Programms mit einem sog. „Programmeter", einem Gerät, das die Einschalt- bzw. Ausschalthäufigkeit der Rundfunkgeräte erfaßte. Es war ihm auch möglich, die Ergebnisse für Arbeiterviertel und „Well-to-do"-Area" zu trennen.

Die Grafiken der „Reichweiten" zeigten, dass die Kurven beider Gruppen bei der Ankündigung als „Populärmusik" wesentlich höher lagen als bei der Ankündigung als „E-Musik", wobei die Werte bei der Arbeitergruppe bei der ersten Ankündigung gegen Ende doppelt so hoch waren (über 80 %) als bei der zweiten Ausstrahlung.[9]

9. Anm.: Die gesamten Ergebnisse finden sich bei GEIGER (1950: 458ff.).

GEIGER zeigt somit, dass - und hier steht er nicht im Widerspruch zu BOURDIEU - bestimmte Sozialisationseffekte („inkorporiertes Kulturkapital“) das Hören klassischer Musik begünstigen.

Dieses Forschungsergebnis widerspricht auf dem ersten Blick ebenfalls den reinen Lebensstilforschern, zeigt sich doch, dass der Musikgeschmack auf „alten“ sozialen Ungleichheiten (Sozialisation) beruht. Wären Lebensstile (in diesem Falle: der Musikgeschmack) frei wählbar, müßten schon rein statistisch die Anteile der Klassik-Hörer deutlich ansteigen - und klassische Musik bzw. Fernseh- oder Radiosender (z. B. DeutschlandRadio Kultur, arte, 3sat) „popularisieren“. Andererseits sind Kulturschemata seit HERDER („Hochkultur“, „Trivialkultur“) vorgegeben und deklinieren sich als Selbstverständlichkeit durch alle Sender: es gibt sog. „Hochkultursender“, die einen sogenannten „kulturellen Anspruch“ mit ihrem Programm verknüpfen und es gibt „Trivialkultursender“, die eher auf Unterhaltung setzen. - Dies vermischt sich heute deutlich und wirft die provokative Frage auf, ob sich Privat- und öffentlich-rechtliche Sender wirklich noch tiefgreifend vom Anspruch unterscheiden - und damit, ob Rundfunk- und Fernsehgebühren noch zeitgemäß sind.[10]

Die Probleme eines derartigen Kulturbegriffs liegen auf der Hand: Klassische Musik oder Opern einerseits können als Hochkultur, Volksmusik und Operette andererseits als Trivialkultur definiert werden. Wie sieht es aber mit Musikrichtungen wie Jazz, Minimalmusik oder neuer elektronischer Musik aus, die sich teilweise ganz bewußt derartiger Klassifizierungen entziehen?

Der letzte Aspekt ist eine Gemeinsamkeit GEIGERs mit BOURDIEU - obwohl BOURDIEU der rein quantitativen Perspektive wohl eher ableh-

10. Es gibt Vorabendserien in ARD und ZDF, die durchaus auch von Privatsendern ausgestrahlt werden könnten (z. B. „Marienhof“, „Unser Charly“). Daneben gibt es z. B. auch ein recht breites Angebot an Volksmusiksendungen.

nend gegenüberstand. Der Begriff der Mentalität unterscheidet sich nur graduell von dem des Habitus:

> „Die Mentalität ... ist geistig-seelische Disposition, ist unmittelbare Prägung des Menschen durch seine soziale Lebenswelt und die von ihr ausstrahlenden, an ihr gemachten Lebenserfahrungen." (GEIGER (1932: 77))

> „... der Habitus ist *Erzeugungsprinzip* objektiv klassifizierbarer Formen von Praxis und *Klassifikationssystem* ... dieser Formen. In der Beziehung dieser beiden den Habitus definierenden Leistungen: der Hervorbringung klassifizierbarer Praxisformen und Werke zum einen, der Unterscheidung und Bewertung der Formen und Produkte (Geschmack) zum anderen, konstituiert sich die *repräsentierte soziale Welt*, mit anderen Worten *der Raum der Lebensstile*." (BOURDIEU (1997: 277f.))

SCHWINGEL vereinfacht die Habitus-Definition folgendermaßen:

> „Der Habitus ist *keine* angeborene und ein für allemal vorgegebene Instanz der Handlungsgenerierung, sondern vielmehr stellt er ein im Laufe des Sozialisationsprozesses *erworbenes* System von Dispositionen dar, das aufgrund der prinzipiellen Unabgeschlossenheit von Lernprozessen, grundsätzlich *historischer Natur* ist." (SCHWINGEL (1993: 65))

Somit liegen sowohl dem Mentalitäts- als auch dem Habitusbegriff Sozialisationsprozesse zugrunde, die sich hinsichtlich „alter" sozialstruktureller Variablen identifizieren lassen. Der fundamentale Unterschied zwischen dem Klassen- und Schichtbegriff bei BOURDIEU und GEIGER ist die Offenheit des Begriffs für Mobilität und sozialen Wandel, der bei BOURDIEU eingeschränkter, bei GEIGER teilweise sehr optimistisch gesehen wird.

GEIGER und BOURDIEU kommen beide zu dem Schluß, dass die Position innerhalb einer Gesellschaft durch ähnliche Erfahrungen (Sozialisation), ähnliche Ressourcen (Bildung, Beruf, etc.), aber auch durch eine zusätzliche „Generierung" dieser Erfahrungen zu „Mentalitäten" bzw. zum „Habitus" verdichtet werden.

1.4 GEIGER und SCHULZE - Mentalität vs. „Erlebnisgesellschaft“

Das Modell der sog. „Erlebnisgesellschaft“ von SCHULZE beruht auf der Grundannahme, dass sich die deutsche Gesellschaft von einer Knappheits- zu einer Überflußgesellschaft entwickelt hat (vgl. KONIETZKA (1995: 87)). Damit knüpft er teilweise an den „Fahrstuhl-Effekt“ BECKs an, dass die Gesellschaft insgesamt materiell „eine Etage höher“ gefahren sei.

Selbst wenn man nur einen flüchtigen Blick auf offizielle Statistiken wirft, so hat Deutschland 1998 knapp 2,5 Mio. Sozialhilfeempfänger im Westen und fast 3 Mio. Sozialhilfeempfänger im Osten (vgl. BUNDESMINISTERIUM FÜR ARBEIT UND SOZIALORDNUNG (2001: Band II, 124f.)). Addiert man die gemeldeten Arbeitssuchenden von ca. 4 Millionen Menschen dazu, so ergibt sich eine Summe von fast 10 Millionen Menschen, die offiziell arbeitslos sind bzw. Sozialhilfe beziehen. Das ist etwa ein Achtel der gesamten bundesrepublikanischen Bevölkerung (bei unberücksichtigter Dunkelziffer und bei Hinzurechnung aller nichterwerbsfähiger Personen, zum Beispiel Kinder und Rentner). Dass diese Gruppen „eine Etage höher“ gefahren sind und sich heute aufgrund ihres materiellen Wohlstandes „frei inszenieren“ können, ist wohl eher unwahrscheinlich. Durch die tiefgreifenden HARTZ-Reformen gibt es immer mehr Menschen, die ein sehr geringes Einkommen beziehen.

SCHULZEs Untersuchung, die er nur in Nürnberg durchführte, bezieht er auf die gesamte Bundesrepublik. Anhand aufwendiger multivariater Analysen gelangt er zu einem Modell, das er auf die Faktoren „Alter“ und „Bildung“ zurückführt - und somit eigentlich auf ein Modell, das große Ähnlichkeit zu sozialstatistischen Schichtmodellen der 60er Jahre („BOLTE-Zwiebel“) aufweist, vor denen GEIGER immer gewarnt hat. Die Altersvariable ist für ihn horizontal angesiedelt und trägt der

zunehmenden Bildungsexpansion Rechnung, die in den Modellen der (unmittelbaren) Nachkriegszeit keine Beachtung fanden.

Die Sekundäranalyse, die Untersuchungsgegenstand dieser Arbeit ist, widerspricht diesem Konzept: Alter ist - gerade was Technikfähigkeiten angeht - kein horizontales, sondern ein vertikales Merkmal: Ältere, die nicht so selbstverständlich mit Computer und Internet aufgewachsen sind, werden von den Jüngeren in diesem Punkt „überholt". Damit ändert sich auch die Rolle der Generationen. Zum ersten Mal geben die Jüngeren „den Ton an" (vgl. BRETON (2000)). Noch vor 20 Jahren war Alter an (berufliche) „Erfahrung" geknüpft und wurde eher positiv als „Kapital" eines Unternehmens gesehen. Vertikale sozialstrukturelle Merkmale werden von SCHULZE kaum wahrgenommen und schlagen sich eher in Lebensstilen denn in Lebenschancen nieder - auch wenn er die wichtige Variable Bildung als vertikale Dimension annimmt.

Auch GEIGER hat sich - interessanterweise aus Sicht des Expressionismus - mit dem Thema des Erlebnisses auseinandergesetzt:

> „Man wendet sich von der Welt der handgreiflichen Dinge ab, vertieft sich in die Wunder des eigenen Ich. Die Projektion wird umgekehrt: statt die Außenwelt im eigenen Innern sich spiegeln zu lassen, will der einzelne seine Seele nach außen projizieren. ... Solche Philosophie erstrebt nicht ein objektives Weltbild, sondern eine subjektive Weltanschauung. Ihr geht es nicht um die Welt der Dinge und deren Erkenntnis, sondern um das 'Erlebnis'." (GEIGER (1964: 16))

Aus GEIGERs Sicht wäre die sozialstrukturelle Einteilung, wie SCHULZE sie vornimmt, höchst problematisch, da „das Erlebnis" im Mittelpunkt der Untersuchung steht.[11] Es scheint auch so, dass die Entwürfe von BECK und SCHULZE nicht originäre Beträge sind, sondern schon von anderen Forschern aus unterschiedlichen Disziplinen behandelt wurden (GEIGER, „postmoderne" Philosophie).

SCHULZEs Verdienst ist die Herausarbeitung einer Kulturtypologie, die deutlich macht, dass sich neben Hochkultur- und Trivialschema heute zusätzlich auch noch ein „Spannungsschema“ nachweisen läßt.

Während das Hochkulturschema (u. a. Klassische Musik, „gute Literatur“ und Museumsbesuche) und das Trivialschema (u. a. Volksmusik, Schlager, Quizsendungen) den traditionellen Charakterisierungen folgt, enthält das Spannungsschema „action“-betonte Elemente (z. B. Rockmusik, Weggehen, ...) (vgl. SCHULZE (1997: 163)). Diesen „Grundorientierungen“ ordnet SCHULZE über die Variablen Alter und Bildung seine Milieus zu.

Damit trägt er dazu bei, die (bundesdeutsche) Kultur durch Milieus und sozialstrukturelle Merkmale charakterisiert zu haben.

1.5 Kritik an GEIGER

Es gibt eine ganze Reihe von Kritikpunkten am GEIGERschen Programm. Hier wäre an erster Stelle die Hinwendung zur reinen quantitativen Analyse zu nennen. Ich möchte hier nicht die Diskussion wiedergeben, die sich bis heute zwischen „qualitativen“ (z. B. narrative Interviews, offene Fragen) und „quantitativen“ (z. B. Fragebogen, standardisierte Interviews) Forschern zuträgt - die Argumente sind hinlänglich bekannt und in vielen Lehrbüchern zur empirischen Sozialforschung zu finden. Diese Position GEIGERs ist - nach meiner Auffassung - heute nicht mehr haltbar, wenn man bedenkt, dass dem Forscher völlig unbekannte Forschungsfelder zuerst einmal durch qualitative

11. GEIGER hat sich 1914 freiwillig im 1. Weltkrieg gemeldet. Damit ist - zumindest damals - eine gewisse Nähe zum Expressionismus nicht unwahrscheinlich, wenn man seine bürgerliche Herkunft berücksichtigt. Allerdings war er sehr schnell ernüchtert und hat sich seit Ende des 1. Weltkriegs gegen Krieg und auch die expressionistische Haltung ausgesprochen. Das Zitat GEIGERs spiegelt die Ernüchterung wider - und die daraus abgeleitete Hinwendung zur Empirie wird deutlich. Da es durchaus Parallelen des Expressionismus zur heutigen Zeit gibt (z. B. die Entfaltung des Individuums, Selbst-Erfahrung) wäre GEIGER diesem Thema gegenüber wahrscheinlich nicht sehr aufgeschlossen, da er die Gefahren selbst erlebt hat.

Erhebungen erschlossen werden müssen (z. B. Ethnologie) - und eher die Forschungsfrage über die Methode entscheiden sollte.

GEISSLER (1995: 291f.) bemerkt kritisch, dass die „Blüte des alten Mittelstandes" nach dem Zweiten Weltkrieg von GEIGER überschätzt wurde. Ebenso die „Machtbalance" zwischen Arbeitnehmern und Arbeitgebern, die sich heute zu einer Hegemonie der Arbeitgeber verschoben hat. Das kapitalistische System ist auch nicht, wie GEIGER unterstellt hat, durch „planwirtschaftliche Elemente" aufgelockert worden. Ganz im Gegenteil, wie wir heute feststellen: neben dem „Kapitalismus" sind eher Lobbyverbände zur Durchsetzung von Interessen (z. B. in der Landwirtschaft oder im Gesundheitswesen) getreten.

GEIGER ist es auch nicht gelungen, die Sozialstruktur der Nachkriegszeit auf eine „griffige Formel" zu bringen. Dies hat aber wohl eher mit seinem frühen Tod (1952) zu tun, wenn man bedenkt, dass die letzten Kriegsheimkehrer erst Mitte der 50er Jahre nach Deutschland zurückfanden und sich Deutschland bis dahin einer sozialstrukturellen Analyse „entzogen" hat. Andere sozialstrukturelle Analysen der 50er Jahre - wie z. B. von SCHELSKY - wurden von diesem später relativiert (vgl. SCHELSKY (1979a, 1979b, 1979c)). Diese Möglichkeit hatte GEIGER nicht mehr.

Ein weiterer Vorwurf an die GEIGERsche Sozialstruktur ist die vorsichtige Äußerung, dass Klassen und Schichten nicht mehr für das spätere Berufsleben prägend sind (vgl. Tabelle 1 auf Seite 17). Dies hat sich nur in Teilen bewahrheitet: zwar sind objektiv die Bildungswege für alle „offen", statistische Ergebnisse sprechen aber dagegen (vgl. u. a. Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (1995: 46)). Der Anteil der Studierenden aus Arbeiterfamilien stieg zwar seit 1953 deutlich an (bis Ende der 70er Jahre), nahm bis

Ende der 80er Jahre etwas ab und stieg seitdem wieder an. Heute liegt der Anteil der Studierenden aus dem Arbeitermilieu bei ca. 16 % (zum Vergleich: 41 % Studierende aus Angestelltenfamilien). Durch die zunehmende Bildungsexpansion und die dadurch einhergehende Entwertung von (niedrigeren) Bildungsabschlüssen sind somit einem Großteil von Kindern aus Arbeiterfamilien der „Weg nach oben" deutlich versperrt: nicht nur, dass „armer Leute Kind" nicht beschließen kann, Kapitalist zu werden (vgl. GEIGER (1964: 65)), Mobilitätsbarrieren sind von GEIGER in diesem Zusammenhang auch unterschätzt worden. Die Einführung von Studiengebühren wird diese Barrieren weiter erhöhen.

Festzuhalten bleibt, dass die GEIGERschen „dominanten Schichtungen" heute nicht mehr die gleichen sind wie vor fünfzig Jahren - die Gesellschaft hat sich in dieser Zeit grundlegend geändert. Auch die Empirie hat sich erheblich weiterentwickelt.

Andererseits muss man konstatieren, dass es - so provokativ es klingt - seit GEIGER nichts grundsätzlich Neues mehr in der Sozialstrukturanalyse gegeben hat: die Begriffe (Schicht, Lebensstil, Milieu, Habitus, Individualisierung, Mentalität, Mobilität, Freisetzung, etc.) sind bereits von GEIGER verwendet worden - ohne Anspruch der Urheberschaft. Nur die Zeit macht es notwendig, die Begriffe zu überdenken, sie gegebenenfalls zu refomulieren bzw. sie zu ergänzen und aktualisieren. Infolgedessen macht es keinen Sinn, ahistorisch z. B. am Freisetzungsbegriff von BECK anzuknüpfen, wenn er schon bei GEIGER verwendet wurde.

Es geht mir in dieser Arbeit nicht darum, GEIGER „in die heutige Zeit" zu übertragen, sondern ihn in seiner Zeit zu belassen und vielmehr an seine Ideen und Anregungen anzuknüpfen. Die Frage lautet nicht „Wie kann man die Gesellschaft mit der GEIGERsche Sozialstruktur-

analyse heute erklären?", sondern vielmehr „Was kann man an GEIGERs Theorie aufgreifen und weiterentwickeln, damit man ein Ergebnis mit aktueller und hoher Erklärungskraft erhält?" - Selbst GEIGER würde heute so verfahren und andere „dominante" bzw. „subordinierte Schichtungen wählen,[12] - wobei sich dominante bzw. subordinierte Schichtungen in dieser Arbeit rein auf den Teilaspekt der PC-Nutzung beziehen. Die Veränderungen der Begriffe seit GEIGER soll anhand der Kontinuitätsbrüche und deren Auswirkungen auf die bundesdeutsche Geschichte beleuchtet werden.

1.6 Zusammenfassung: GEIGERs Beitrag für die Untersuchung der Sozialstruktur heute, die Erweiterung durch die „Postmoderne" und für die PC-Nutzung

GEIGERs Ideen (vor allem die dominanten Schichten) sollen in dieser Arbeit herangezogen werden, um damit Einstellungen zum PC näher zu untersuchen. Diese Arbeit vertritt nicht den Anspruch, die Ideen GEIGERs zu einem „neuen" Gesellschaftskonzept auszubauen, sondern zielt darauf ab, die Ideen GEIGERs auf die Nutzung dieser Technologie zu übertragen und anzuwenden.

Die Konfrontation von GEIGERs Ansätzen mit heutigen Sozialstrukturmodellen hat gezeigt, dass die Ansätze von BECK und SCHULZE vertikale Ungleichheit unzureichend oder überhaupt nicht erfassen. Der Ansatz von BOURDIEU, in den 70er Jahren in Frankreich entwickelt, paßt ebenfalls nicht auf die Sozialstruktur Deutschlands heute: Ungleichheit wird als Klassenkonzept mit geringer sozialer Mobilität beschrieben.

12. GEIGER selbst gibt in der „Klassengesellschaft im Schmelztiegel" einen Hinweis darauf, dass das MARXsche System nur aufgrund des „Gegensystems" des Liberalismus bestehen kann (vgl. GEIGER (1948/49, S. 38)). Damit räumt er auch indirekt ein, dass jedes Gesellschaftssystem zu ganz bestimmten Zeiten bestimmte Sozialstrukturentwürfe entwickelt.

Der Vorteil von GEIGERs Schichtenkonzept liegt in der Offenheit - auch für heutige Verhältnisse. Durch die Verknüpfung allgemeiner, durch sozialen Wandel sich verändernder vertikaler Merkmale und Mentalitäten erhält man ein flexibles Instrument zur Untersuchung der Sozialstruktur.

Das Schichten-Konzept GEIGERS hat nichts mit Computerforschung zu tun, kann aber auch hier sinnvoll eingesetzt werden, wenn angenommen wird, dass Techniknutzung einen gewissen statistischen Zusammenhang mit sozialstrukturellen bzw. Kultur- und Freizeitstilen besitzt.

Der Begriff der vertikalen Ungleichheit wird ausdrücklich erweitert: nicht nur Beruf, Bildung oder Einkommen, auch Alter und Geschlecht sind heute eindeutig vertikale Merkmale, die über Teilhabe und Chancen entscheiden - vor allem, wenn es um Techniknutzung geht.

2 Der PC aus sozialwissenschaftlicher Sicht

Durch die technische Entwicklung der letzten fünfzig Jahre, insbesondere der letzten 20 Jahre ist die Computertechnologie in Form von Chips, Homecomputern, usw. nicht mehr aus unserem Leben wegzudenken: Elektronik findet sich in Autos, Haushaltsgeräten, in PCs, und anderen technischen Geräten.[13]

Der PC ruft heute sowohl große Zustimmung als auch große Ablehnung oder Angst hervor. Dies beruht auf einigen Charakteristika, die kein anderes technisches Gerät besaß und besitzt.

Nicht wenigen Arbeiten über Computer haftet entweder das Vorurteil an, dass der Computer das Leben komplett verändert und kon-

13. Der Begriff des „Computers" bezieht sich auf alle Geräte, die mittels Chips gesteuert werden (Autos, Waschmaschinen, Handys, etc.). Im Mittelpunkt der Arbeit steht jedoch der PC (Personal Computer).

trolliert, oder der Computer wird als Allheilmittel für alle heutigen und künftigen Probleme hingestellt, die technische Seite glorifiziert (Stichwort: papierloses Büro) und seine Vielfältigkeit gepriesen.

In dieser Arbeit wird der Personal Computer als wichtige Technologie verstanden, die aber weder den „Untergang des Abendlandes" herbeiführt, noch die Lösung aller Probleme darstellt. Er kann, sofern der Bediener versiert ist, viele Alltags- und Routineaufgaben erledigen, er ist universell programmierbar, aber er ist und bleibt ein Arbeitsgerät und entwickelt auch kein Eigenleben. Insofern ist meine Betrachtung eher nüchtern. In dieser Arbeit spielt der PC vor allem eine Rolle im Zusammenhang mit dem Umgang durch die Menschen. Es gibt keine „richtige" und „falsche" Nutzung des Rechners, es gibt viele Arten und Möglichkeiten, die aus der Sicht dieser Arbeit gleichberechtigt nebeneinanderstehen.[14]

Während alle anderen technischen Geräte zumeist für wenige Funktionen konzipiert sind, kann der PC universell eingesetzt werden: zur Texterfassung, Datenbankverwaltung, Modelleisenbahnsteuerung, als Videorecorder, usw.

Durch die vielfältigen Möglichkeiten des Rechners ist die „Kommunikation" per Betriebssystem komplex. Alle anderen technischen Geräte (auch die, die heute neu entwickelt werden) sind einfacher zu bedienen und knüpfen an bestimmte „Vorerfahrungen": ein Handy beispielsweise läßt sich grundsätzlich genauso bedienen wie ein herkömmliches Festnetztelefon. Die Grundfunktionen - anrufen bzw. Anrufannahme - sind identisch. Zusatzfunktionen - wie z. B. SMS, Einstellen von Klingeltönen, o. ä. sind optional und können vom Nut-

14. Natürlich ist klar, dass der Computer heute die Gesellschaft unmittelbar verändert; nicht nur privat, sondern insbesondere am Arbeitsplatz. Fehlende Computerkenntnisse können ein Grund für Arbeitslosigkeit sein; dies beeinflußt das gesamte Leben der arbeitslosen Person. Für die Arbeit ist allerdings eine nüchterne Auseinandersetzung mit dem Computer aus meiner Sicht dringend erforderlich.

zer erlernt werden oder auch nicht: die eigentliche Funktion des Telefonierens bleibt davon unberührt. Das Gleiche gilt für Stereoanlagen, Küchengeräte, Fernseher oder Videorecorder: die Zusatzfunktionen erleichtern oder erweitern die Bequemlichkeit im Umgang mit den Geräten (z. B. ShowView beim Videorecorder), verändern aber nicht ihre eigentliche Grundfunktionen.

Wer jedoch Texte mit einer Textverarbeitung erfassen will, muß dies erlernen. Hierbei gibt es verschiedene Wege („Versuch und Irrtum", Lernprogramm, Computerkurs, etc.). Die „Lernwege" sind von persönlichen Präferenzen, evtl. Vorerfahrungen und auch persönlichen Sichtweisen des einzelnen (kurz: Sozialisationserfahrungen) abhängig: wer wenig Scheu vor dem PC besitzt, wird eher dazu neigen, Dinge auszuprobieren.

Der Begriff der Sozialisation wird als

> „... ein Modell der wechselseitigen Beziehungen zwischen Subjekt und gesellschaftlich vermittelter Realität, eines interdependenten Zusammenhangs von individueller und sozialer Veränderung und Entwicklung ..." (HURRELMANN (1998: 64))

definiert. Die „Sozialisationserfahrungen" mit dem Computer beziehen sich hingegen auf persönliche Erfahrungen mit dieser Technik: sie können positiv oder negativ sein und bestimmen den zukünftigen Umgang. Wer eher ängstlich ist, dass er etwas „zerstören" könnte wird sich dem Computer anders nähern und Fehlermeldungen anders wahrnehmen als derjenige, der durch Ausprobieren lernt. Dies schlägt sich auf unterschiedliche Wahlmöglichkeiten (von reiner Anwendung (z. B. Textverarbeitung) bis hin zur Programmierung) nieder: eine Person, die dem Computer eher ängstlich oder kritisch gegenübersteht, wird kaum Erweiterungen (Speicher, Grafik-, Soundkarten, etc.) selbst einbauen. Sozialisation ist hier auf verschiedene Ziele gerichtet (Anwendung, Reparatur, Wartung, Programmierung, etc.), die sich ausschließen können, aber nicht müssen.

Umfassende Computerkenntnisse sind also grundsätzliche Voraussetzung für den Zugang zum Internet, gerade wenn man nicht über Online-Dienste wie AOL oder T-Online geht, sondern den Internetzugang mittels DFÜ-Netzwerk selbst konfiguriert:[15]

> „Internet-Anwender verfügen generell über eine größere Technikaffinität und damit über bessere technische Voraussetzungen als Computeranwender, die das Internet nicht nutzen. Dieser Trend wird zusätzlich durch die Dynamik des Mediums Internet verstärkt: Der Online-Bereich entwickelt sich im Verhältnis zum Offline-Bereich überdurchschnittlich schnell, so dass die technischen Unterschiede zukünftig stärker ausgeprägt sein werden." (ComCult Research (2001: 2))

Die Forschergruppe um RAMMERT et. al. (1991: 20f.) typisiert Computerkompetenzen in „Programmierer", „Anwender", „Hardware-Kenner" und „System-Kenner". Diese Kenntnisse werden zusammenfassend auch als „Computer literacy" bezeichnet:

> „Unter 'Computer Literacy' wird hier zunächst die Gesamtheit von *prozeduralen* und *deklarativen Wissensbeständen* verstanden, die dem Individuum einen *kompetenten Umgang mit dem Computer* und damit eine individuell wie sozial erfolgreiche Teilnahme an der computerorientierten Gesellschaft ermöglichen." (RICHTER, NAUMANN und GROEBEN (2001: 2))

Neben den Kompetenzen spielt bei den Autoren auch die persönlich wahrgenommene Sicherheit im Umgang mit Computern eine wichtige Rolle (vgl. RICHTER et. al. (2001: 2f.)).

Überraschenderweise hält sich die Zahl der Studien, die sich mit Computerstilen befassen, in Grenzen. Tabelle 2 auf Seite 37 gibt einen Überblick über die wichtigsten Ergebnisse. Hierbei wird deutlich, dass ebenso wie bei der Clusterung von Internetnutzern die Nutzungsarten die wichtigste Rolle spielen. Das ist nicht überraschend, bietet doch dieser Zugangsweg Informationen über das Verhalten

15. Das hat auch die AOL-Werbung mit Boris BECKER deutlich gemacht: „Bin ich schon drin?"

und damit die Präferenzen der User. Dieser Weg soll auch für die nachfolgende Sekundäruntersuchung als Ausgangspunkt dienen.

TABELLE 2 AUSGEWÄHLTE STUDIEN: TYPEN DER COMPUTERNUTZUNG (VGL. RAMMERT (1990), RAMMERT ET. AL. (1991), BÜHL (1999), MAAZ ET. AL. (2000), HOFFMANN (2001))

Studie / Jahr / Frage	Ergebnisse (Typen)
MAAZ et. al. (2000): Generation N (nur Jugendliche) „Wie setzen sich Kinder und Jugendliche mit den Neuen Medien auseinander?“	die Wenignutzer die Standardanwender (Spiele, Schule) die Spieler und Onlinenutzer die Vielnutzer (Grafik, Text, Lernen)
WETZSTEIN, et. al.: Datenreisende - die Kultur der Cybernetzwerke (1995) „Welche sozialen und kulturellen Veränderungen bringen nichtkommerzielle Computernetzwerke für die Gesellschaft?“	Freak (Hacker, Cracker, Programmierer, Spieler) Hobby - User Pragmatiker
BÜHL et. al.: Computerstile (1999) „Wie wird mit dem PC im Alltag umgegangen?“	Internet Text (2 Subgruppen) Spieler (2 Subgruppen) Anwender (2 Subgruppen) Experte
RAMMERT et. al.: Vom Umgang mit Computern im Alltag (1991) „Welche unterschiedlichen Aneignungsmuster des Computers gibt es?“	Der Computer als ... ... lebensstilbildendes Medium ... qualifikatorische Ressource (3 Subgruppen) ... Passion ... intellektuelle Herausforderung
PFLÜGER: Computer und Mythos - Methapern eines geregelten Alltags (1990) „Wodurch läßt sich das Neue im Verhältnis Mensch und Computer fassen und was ist unter Computernutzung zu verstehen?“	der passionierte Nutzer der hobbymäßige Nutzer der gezwungene Nutzer der beeinflußte Nutzer
ECKERT et. al.: Auf digitalen Pfaden (1991) „Was machen die Menschen mit den Medien?“	Hacker/Crasher Programmierer / Cracker Spieler (Sportler, Denker, Dramaturg

Die Studien untersuchen vor allem Nutzungsmuster (z. B. Hacker, Programmierer, Spieler) und Anwendungsbereiche (z. B. Hobby, gezwungen, passioniert, beeinflußt) - weniger die sozialstrukturellen Merkmale. Umfangreiche sozialstrukturelle Eigenschaften stehen

kaum im Mittelpunkt, sondern eher die Anwendung (außer das Geschlecht). Ähnliches gilt auch für die Internetnutzung (Tabelle 3 auf Seite 39). Auch hier dominiert, ähnlich wie bei den Computerstilen, der Nutzungsaspekt. Das ist nicht verwunderlich, ist doch eine der Hauptfragestellungen der Technikforschung, wie Menschen damit umgehen bzw. zurechtkommen. Aber: gibt es dominante Schichtungen, die die Techniknutzung erklären - z. B. in Form von Bildungszertifikaten, Alter, Geschlecht oder Kulturvorlieben:

TABELLE 3 AUSGEWÄHLTE STUDIEN: TYPEN DER INTERNETNUTZUNG (VGL. SPIEGEL-VERLAG (2000), SCHEID (1999), G+J ELECTRONIC MEDIA SERVICES (2000, 2001), VAN EIMEREN, ET. AL. (2000, 2001), GRÜNE UND URLINGS (1996)

Studie (Erscheinungsjahr)	Ergebnisse (Typen)
SPIEGEL ONLINE - OFFLINE (vgl. SPIEGEL-Verlag (1997: 9ff.))	• die Versierten • die Pragmatiker • die Begeisterten • die Ängstlichen • die Desinteressierten
Stern - Markenprofile (zit. nach SCHEID (1999: 25))	• Professionelle Nutzer • Service-Interessierte • Shopping-Interessierte • News-Interessierte • Desinteressierte
SCHEID: Chattende Spieler, surfende Infosucher uns shoppende Profis - Entwicklung einer Nutzertypologie für deutschsprachige Internetnutzer (Diplomarbeit, 1999)	• die Internet-Profis • die Newbies • die Alltagsnutzer • die qualifizierten Infosucher • die kommunikativen Spieler
ARD-/ZDF-Studien (siehe van Eimeren et. al.)	• Optimisten • Pragmatiker • Pessimisten
GRÜNE und URLINGS: Motive der Onlinenutzung. Ergebnisse der psychologischen Studie „Die Seele im Netz“ (1996)	• Pionier / Profi • Spät(er) Berufene • Neugierige • Aufgeschreckte • Abgeschreckte • Indifferente

3 Zur Theorie der sozialen Schichtung heute - Ableitung der Hauptfragestellung am Beispiel der PC-Nutzung

GEIGERs Analysen der Sozialstruktur eröffnen heute neue Blickwinkel - vor allem deshalb, weil seine Art der Schichtungsanalyse seit seinem Tod kaum breite Beachtung gefunden hat. Allerdings muß man sich von der Schichtanalyse der 50er und 60er Jahre eher lösen, die zumeist weit hinter GEIGER zurückbleibt. GEISSLER (1995) nennt hier einige Punkte wie z. B. „untheoretisches“ oder „unhistorisches Vorgehen“, „einseitige Orientierung auf das Sozialprestige“, „Ausblendung begrifflicher und methodologischer Reflexionen“, „deterministische und stereotype Deutung der Zusammenhänge“, usw. (vgl. GEISSLER (1995: 284f.)).

Die Analysen der Nachkriegszeit haben den Schichtbegriff aus heutiger Sicht eher diskreditiert als gestärkt. Häufig ist er als rein sozialstatistischer Begriff mißbraucht worden, wovor GEIGER immer gewarnt hat. Aus diesem Grund ist es sinnvoll, wieder zu einem Schichtbegriff zu kommen, der (sozialstatistische) Lagemerkmale mit Werthaltungen (Mentalitäten) verknüpft - oder durch Kultur- und Freizeitvariablen erweitert. Eine derartige Vorgehensweise findet sich in der neueren Literatur bei SCHROTH (1999).

Ein vertikales Modell - hierin sind sich wohl die meisten Sozialstrukturforscher einig - ist heute nicht mehr ausreichend, um eine Gesellschaft adäquat zu beschreiben. Es hat in den letzten Jahrzehnten diverse Umbrüche gegeben, die dagegensprechen: die Entwertung der Schulabschlüsse oder auch die Einschätzung von Firmen durch Arbeitnehmer hinsichtlich der Arbeitsplatzsicherheit.[16]

16. Anm.: Vor etwa zwanzig Jahren wurden große Firmen wie Banken oder Versicherungen als absolut sichere Arbeitsplätze angesehen. Selbst bis vor einigen Jahren galten Arbeitsplätze im EDV-Bereich als relativ zukunftssicher. Wie schnell sich dies verändern kann, haben die Kurse und Pleiten am sog. „Neuen Markt“ gezeigt

Die nachfolgende Untersuchung geht von folgenden Prämissen aus:

- PC-Nutzung ist in erster Linie durch vertikale Ungleichheiten definiert, wobei neben den „klassischen" Variablen Bildung, Beruf oder Einkommen auch Alter und Geschlecht ausdrücklich herangezogen werden
- Nicht alle vertikalen Variablen können zur Analyse herangezogen werden, sondern nur die dominantesten (wichtigsten): hierbei müssen auch hohe Korrelationen (die z. B. zwischen Beruf und Bildung existieren könnten) berücksichtigt werden. Allerdings wird es immer Zusammenhänge zwischen Variablen geben. Ziel ist es, Gruppen anhand sozialstruktureller Merkmale zu beschreiben
- Die gefundenen Segmente werden weiter hinsichtlich Kultur- und Freizeitvariablen untersucht. Somit wird ein vielfältiges Bild der PC-Nutzer entworfen. Die Gruppen werden sowohl empirisch als auch theoretisch gebildet
- Methodisch kommen neben der deskriptiven Statistik verschiedene multivariate Verfahren zum Einsatz. Da die meisten Variablen des Datensatzes nicht metrisch sind, fällt die Auswahl auf die logistische Regression und die Diskriminanzanalyse. Beide Verfahren geben - wie Entscheidungsbäume - Fehlklassifikationstabellen aus
- Das Hauptverfahren für diese Arbeit sind allerdings unterschiedliche Entscheidungsbaumalgorithmen, die in der Soziologie sehr selten bzw. nie eingesetzt werden. Sie bilden den Schwerpunkt der Analysen, da sie flexibel mit allen Skalenniveaus umgehen können. Die „klassischen" Verfahren dienen der Überprüfung der Ergebnisse. Dahinter steht die Frage, ob Entscheidungsbäume zu einem ähnlichen Ergebnis kommen oder nicht und somit das „klassischen" soziologischen Verfahren ebenbürtig sind
- Das empirische Vorgehen ist sowohl theoriegeleitet als auch explorativ: das theoretische Vorgehen bezieht sich auf die beschriebene sozialstrukturelle Analyse mit anschließender Untersuchung der Kultur- und Freizeitvariablen. Welche Variablen als „dominant" bzw. „subordiniert" anzusehen sind, wird allerdings explorativ über die Ergebnisse der Zusammenhänge ermittelt

Zusammenfassende Ziele sind also, Entscheidungsbaumalgorithmen in der Soziologie zu etablieren oder sie zu verwerfen und eine sozialstrukturelle Untersuchung der PC-Nutzung anhand unterschiedlicher Variablen durchzuführen, die theoretisch auf Theodor GEIGER beruhen.

KAPITEL III METHODISCHER HINTERGRUND

Bei einer empirischen Untersuchung stellt sich die Frage nach der Forschungsmethode (Primär- vs. Sekundärerhebung, Online- vs. Offlinebefragung, etc.), gefolgt von den eingesetzten statistischen Verfahren. Ein Schwerpunkt dieser Arbeit liegt zum einen auf Data Mining-Verfahren, die durch Entscheidungsbäume zum Einsatz kommen, zum anderen auf multidimensionaler Visualisierung in Form von Parallel-, Spine und Mosaic Plots. Daneben werden die multinominale logistische Regression und Diskriminanzanalyse herangezogen, deren Unterschiede und Vorgehen ausführlich erläutert werden. Vor und nach dem Einsatz der multivariaten Statistik stehen deskriptive Auswertungen und Interpretationen.

Die Analyse gliedert sich in folgende Schritte

- Deskription der sozialstrukturellen Variablen und Auswahl der wichtigsten („dominanten") Merkmale für die weitergehende Analyse
- Beschreibung der gefundenen Segmente mit den Mitteln der Entscheidungsbäume
- Vergleich der Ergebnisse mit den anderen Verfahren (Diskriminanzanalyse, multinominale logistische Regression)
- Deskriptive Beschreibung der Kultur- und Freizeitvariablen und Auswahl der wichtigsten („subordinierten") Merkmale für die weitere Beschreibung der Gruppen
- Bildung von Gruppen anhand der Entscheidungsbaumalgorithmen („dominante Schichtungen")
- Beschreibung der gefundenen Segmente

In diesem Kapitel werden jedoch zuerst die methodischen Fragen erläutert. Diese basieren auf der Begründung für den Einsatz der Methoden bis zur Beschreibung der verschiedenen multivariaten Verfahren und deren Klassifikationsergebnisse, die anhand eines einfachen Modells mit einer abhängigen und zwei unabhängigen Variablen dargestellt werden.

Auch wenn das zum Einsatz kommende Beispiel recht alltäglich scheint, besteht das Ziel dieses Kapitels nicht in einer Ergebnispräsen-

tation, sondern vor allem in der Charakterisierung der Entscheidungsbaumalgorithmen, da dieses Verfahren kaum jemals in der aktuellen Soziologie zum Einsatz kommt. Dieses Defizit will diese Arbeit ausgleichen.

Diese Untersuchung versucht, dominante Schichtungen für die PC-Nutzung anhand einer Offlinebefragung zu identifizieren. Der große Vorteil gegenüber der Onlinebefragung liegt in der Erreichbarkeit auch von Nichtnutzern. Die Nichtnutzer, deren Aussagen als „Kontrollgruppe“ interessante Ergebnisse erwarten lassen, könnten bei einer Onlinebefragung überhaupt nicht erreicht werden.

1 Möglichkeiten des methodischen Vorgehens

Zu Beginn eines Forschungsprojekts muss die Frage der angemessenen Methoden geklärt werden, insbesondere die Frage einer qualitativen oder quantitativen Vorgehensweise - jeder Weg bietet Vor- bzw. Nachteile - und es liegt in der Abwägung des Forschers/der Forschergruppe, adäquate Methoden im Hinblick auf eine spezifische Forschungsfrage auszuwählen. Nur dieses Kriterium sollte - unter Einbeziehung der bestehenden Forschung - den Ausschlag für qualitatives oder quantitatives Vorgehen bestimmen, nicht persönliche Vorlieben bzw. Abneigungen des Forschers, wie sie in vielen Wissenschaftsdisziplinen zu finden sind (vgl. u. a. HOFFMANN (2002a: 90ff.)).

1.1 Primär- vs. Sekundäranalysen

Die Entscheidung, einer Primär- oder einer Sekundäranalyse den Vorzug zu geben, hängt von einer Reihe von Variablen (z. B. finanzieller, zeitlicher, technischer Art sowie von der Verfügbarkeit von Daten, bezogen auf die Forschungsfrage) ab (vgl. u. a. FRIEDRICHS (1990: 353ff.)). Grundsätzlich müssen bei Forschungsfragen, zu denen keine empirischen Befunde vorliegen, Primäranalysen durchgeführt wer-

den. Die Problemstellungen werden operationalisiert und stringent im Forschungsdesign und der Methode (qualitativ, quantitativ, online, offline, etc.) umgesetzt.

Andererseits sind die heute verfügbaren Datensätze aus Primärerhebungen sehr zahlreich, häufig nicht vollständig ausgeschöpft und können durchaus in vielen Fällen für Sekundäranalysen herangezogen werden (vgl. DIEKMANN (1995: 173)). Hier bietet sich vor allem das Zentralarchiv für empirische Sozialforschung an der Universität zu Köln für eine Recherche an, das eine Vielzahl von Datensätzen aus unterschiedlichsten sozialwissenschaftlichen Erhebungen bereitstellt (z. B. EUROBAROMETER- oder ALLBUS-Befragungen).

Neben der Kostenersparnis (Porto-, Druckkosten) spielt auch die Qualität der erhobenen Daten eine nicht unerhebliche Rolle. Somit sind Sekundäranalysen - insbesondere bei geringem finanziellen Budget - eine interessante Alternative zu Primärerhebungen - wenn geeignete Datensätze für die jeweilige Fragestellung existieren.

Für die Forschungsfrage dieser Arbeit - der Analyse von dominanten Schichtungen von PC-Nutzern - trifft dies zu.[17] Im Rahmen des EUROBAROMETER 56.0 (Erhebungszeitpunkt: 2001) wurden eine Reihe von Techniknutzungs-, Freizeit-, Sozialstruktur- und Finanzvariablen erhoben. Verknüpft mit dem sehr offenen Ansatz von GEIGER wird untersucht, ob (und wenn ja: welche) Variablen die generelle PC-Nutzung beschreiben.

17. Im folgenden werden ausschliesslich die Begriffe des PC oder Personal Computers benutzt, nicht jedoch von „Computern" da sich Computer in einer Reihe von elektronischen Geräten (z. B. Handys, Autos, etc.) befinden.

1.2 Deduktiv-nomologisches vs. exploratives Vorgehen

In den Sozialwissenschaften gibt es unterschiedliche Standpunkte zum forschungslogischen Ablauf, speziell zum deduktiv-nomologischen (hypothesengeleiteten) und explorativen (entdeckenden) Vorgehen. Beim hypothesengeleiteten Vorgehen werden - im strengsten Falle - vor der Instrumentenerhebung Hypothesen aufgestellt, die Methode gewählt, Daten erhoben, hinsichtlich der Hypothesen ausgewertet und interpretiert. Weitere Zusammenhänge werden nicht untersucht.

Beim explorativen Vorgehen werden - im Extremfall - keine/kaum Hypothesen gebildet und die erhobenen Daten eher spielerisch nach ad-hoc-Hypothesen miteinander in Beziehung gesetzt.

Beide Vorgehensweisen bieten sowohl Vor- als auch Nachteile: ein zu „engstirniges“ Vorgehen bei der Auswertung des Datensatzes läßt leicht wichtige Zusammenhänge übersehen, theorieloses Vorgehen führt im schlimmsten Fall zur Scharlatanerie, wo möglicherweise Zusammenhänge postuliert werden, die real nicht existieren. So könnten sich - bei einer hohen Fallzahl - signifikante Effekte ergeben, die im Nachhinein ohne Hypothesen gedeutet werden und zu völlig falschen Schlüssen führen..

Mir scheint hier ein Vorgehen angebracht, dass durchaus Grundhypothesen bildet, jedoch Offenheit gegenüber den Daten aufbringt. Die Daten werden aufgrund möglicher sinnvoller Zusammenhänge selektiert (z. B. Alter, Geschlecht, Bildungsgrad, Beruf, Familienstand, Haushaltsnettoeinkommen, etc.). Wenig sinnvolle Variablen (z. B. Telefonbesitz, Interviewdauer) werden nicht untersucht. Dieses Vorgehen sollte nicht als „theorielos“ eingeschätzt werden. Dahinter steht die Annahme, dass sich dominante Variablen dem sozialen Wandel unterliegen und sich über die Zeit ändern. Auch gibt es nicht uner-

hebliche Korrelationen zwischen einzelnen Merkmalen (z. B. Bildung und Beruf oder Alter und Bildung), die in gewissen Sinne Berücksichtigung finden müssen.

Der Arbeit liegt weiter die Annahme zugrunde, dass dominante Schichtungen nur durch alte Ungleichheiten (incl. Geschlecht und Alter) repräsentiert werden können und subordinierte Segmentierungen nur durch Kultur- und Freizeitvariablen, die ebenfalls explorativ nach Wichtigkeit untersucht werden müssen.

Es ergeben sich somit dominante und subordinierte Variablen für die untersuchte Stichprobe. Damit läßt sich PC-Nutzung allgemein beschreiben (z. B. PC-Nutzung ist grundsätzlich abhängig von Alter und Beruf (dominant), PC-Nutzer besuchen häufiger als Nichtnutzer das Kino (subordiniert), hören deutlich weniger Volksmusik (subordiniert) und hören dafür mehr Rock-/Popmusik (subordiniert)) - aber auch für jede Gruppe nochmals spezifisch (z. B. könnten ältere PC-Nutzergruppen existieren, die Volksmusik (subordiniert) hören, dafür keine Rockmusik (subordiniert)).

Dominante und subordinierte Schichtungen lassen sich problemlos durch multivariate Verfahren finden - die einzelnen Gruppen lassen sich jedoch nur (vor allem, wenn sie geringe Fallzahlen aufweisen) deskriptiv beschreiben. Deshalb findet ein steter Wechsel zwischen multivariaten und deskriptiven Verfahren statt.

Das Vorgehen ist somit weder rein explorativ noch deduktiv - allerdings mit einem Schwerpunkt zu explorativem Vorgehen - was für die Auswertung eines quantitativen Datensatzes vielleicht eher etwas ungewöhnlich erscheint. Es lassen sich jedoch grundsätzlich alle multivariaten Verfahren auch dazu nutzen, Zusammenhänge zu entdekken - mit der Chance, neue Sachverhalte zu erkennen.

In dieser Arbeit wird unterstellt, dass „alte“ soziale Ungleichheiten sowie Alter und Geschlecht vertikale, Kultur- und Freizeitvariablen horizontale Ungleichheitsdimensionen abbilden. Da Kultur- und Freizeitaktivitäten ohne (Bildungs- oder monetöre) Ressourcen schwierig umzusetzen sind, kommt den vertikalen Variablen eine dominante, den Kultur- und Freizeitvariablen subordinierte Bedeutung zu.

Welche vertikalen bzw. horizontalen Variablen nun dominant wirken - das wird über die Zusammenhangsmaße mit der PC-Nutzung definiert: je höher das Zusammenhangsmaß, desto deutlicher leistet die Variable einen Beitrag zur Erklärung der PC-Nutzung und desto dominanter wirkt sie für diese Fragestellung.[18] Ist das Zusammenhangsmaß zwischen PC-Nutzung und den Variablen <0.2, liegt eine zu geringe Kohäsion vor, so dass die Variable aus der Analyse ausgeschlossen wird.

1.3 Operationalisierung des PC-Nutzers

Die Nutzung eines PCs läßt sich - auch im Datensatz - auf unterschiedliche Weise operationalisieren. Die Häufigkeit wurde ordinal erfaßt:

18. Es wird hierbei nicht intendiert, ein allgemeines Sozialstrukturmodell zu entwickeln - dominante und subordinierte Variablen beziehen sich in dieser Arbeit ausschließlich auf die PC-Nutzung.

ABBILDUNG 1 Häufigkeit der PC-Nutzung (N = 2047, %)

Q 39: Häufigkeit: PC-Nutzung

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	nie	1054	51,3	51,5	51,5
	seltener als einmal im Monat	40	2,0	2,0	53,5
	ein- bis dreimal im Monat	55	2,8	2,8	56,3
	einmal die Woche	108	5,3	5,3	61,6
	mehrmals die Woche	314	15,3	15,4	77,0
	täglich	468	22,9	23,0	100,0
	Gesamt	2038	99,6	100,0	
Fehlend	weiss nicht	9	,4		
Gesamt		2047	100,0		

Von 2047 befragten Personen nutzen 1054 keinen PC, 984 nutzen ihn. 9 Befragte haben sich zu dieser Frage nicht geäußert.

Die Variable besitzt ordinales Skalenniveau, die Ausprägungen lassen sich nach Höhe der PC-Nutzung ordnen, aber die Abstände sind nicht gleich groß. Somit läßt sich keine metrische Variable bilden.

Ist es sinnvoller, mit der ordinalen oder einer nominalen Variable zu arbeiten? - Das hängt vom weiteren Vorgehen ab: nicht immer erhöht eine ordinale Variable den Informationsgehalt - zum Beispiel in diesem Fall. Die häufigere Nutzung eines PC suggeriert auch die zunehmende Kompetenz im Umgang mit dieser Technik - in der Praxis ist dies jedoch sehr fraglich. Wer in einem Büro arbeitet, keinen Bezug zum PC besitzt, gezwungenermaßen mit dem Gerät arbeitet und vielleicht gerade einen einfachen Text mit Word schreiben und ausdrucken kann, würde in die Kategorie „mehrmals die Woche“ fallen, ohne auch nur in Ansätzen eine gewisse Kompetenz im Umgang zu besitzen. Steht bei dieser Person zuhause ein PC und werden am Wochenende emails abgerufen und beantwortet, so wird der PC sogar

täglich genutzt. Andererseits würde ein Programmierer, der fünf Tage die Woche programmiert, jedoch das Wochenende mit seiner Familie verbringt, nur unter „mehrmals die Woche“ aufgeführt, obwohl die Kompetenz im Umgang mit dem Rechner wesentlich größer ist.

Hier liegt - und das muss offen zugegeben werden - eine gewisse Ungenauigkeit in den Daten für diese Fragestellung. Auf der anderen Seite ist es sehr schwierig, Kompetenzen mit dem Computer zu erfassen - denkt man zum Beispiel am Umgang im Büro. Ein Befragter kann täglich mit dem PC umgehen und Stunden brauchen, einen Brief zu schreiben, ein anderer braucht wenige Minuten dazu. Beide würden sie angeben, mehrmals die Woche Bürotätigkeiten auszuführen - obwohl die Kompetenz sich sehr weit unterscheidet. Um den nominalen und den ordinalen Fall zu prüfen, werden die Fragestellungen anhand einer nominalen, einer ordinalen PC-Nutzungsvariablen untersucht - vor allem, um die Möglichkeiten und Grenzen der Entscheidungsbäume darzustellen.

Andererseits darf nicht unterschätzt werden, dass multivariate quantitative Verfahren viele Werte nivellieren, indem sie z. B. einen Gruppenmittelwert berechnen. Auch unter Hinzuziehung der Standardabweichung bleiben Differenzen - z. B. zwischen Personen eines Clusters - bestehen.

Ein weiterer Punkt soll ebenfalls nicht unerwähnt bleiben: Zielvariablen mit vielen Ausprägungen (wie z. B. die ordinal erfaßte PC-Nutzung) können im schlechtesten Fall zu überhaupt keinem multivariaten Ergebnis führen, wenn sich nicht bestimmte Merkmale auf Subgruppen zurückführen lassen. Zusätzlich ist es fraglich, ob der Erkenntnisgewinn sehr groß sein kann, wenn die Gruppen „seltener als einmal im Monat“ oder „ein bis dreimal im Monat“ charakterisiert werden.

Aus den o. g. Bedenken wurde hauptsächlich auf eine nominale Zielvariable (PC-Nutzung ja bzw. nein) zurückgegriffen - mag sie auch zu einer gewissen Ungenauigkeit beitragen. In diesem Sinne scheinen nominale Variablen realitätsgerechter zu sein als metrische oder ordinale. Angenommen, ein Befragter nutzt manchmal den PC nur einmal die Woche, manchmal mehrmals, ein anderer manchmal täglich, manchmal fünf Tage die Woche: weder qualitative noch quantitative Forschung könnten diese Antwort richtig verorten, da bei beiden Kategorien gebildet werden müssten, um die Antwort zuzuordnen. Schließlich wird man nach einer Regel vorgehen und diesen Befragten einer Kategorie zuordnen - mit der dann gerechnet wird bzw. die dann interpretiert wird. Beide Arten der Forschung produzieren hier Ungenauigkeiten, die in der Praxis natürlich nicht sehr stark ins Gewicht fallen.

Da es das erklärte Ziel der Arbeit ist, Entscheidungsbaumalgorithmen zu erläutern, werden neben dem Heranziehen der nominalen PC-Nutzer-Variable auch eine ordinale und eine metrische Variante vorgestellt, die zeigen, wie unproblematisch und flexibel Entscheidungsbäume im Gegensatz zu vielen anderen multivariaten Verfahren mit unterschiedlichen Skalenniveaus umgehen.

1.4 Eingesetzte Verfahren

In dieser Arbeit kommen sowohl Data-Mining in Form von Entscheidungsbäumen als auch „herkömmliche“ statistische Verfahren (Regression, Diskriminanzanalyse) zum Einsatz. Die Verfahren werden in diesem Kapitel anhand eines einfachen Beispiels erläutert.

Die vorliegenden Variablen des untersuchten EUROBAROMETER-Datensatzes sind meistens nominal oder ordinal skaliert, d. h., sie können für viele „klassische“ multivariate Verfahren nicht herangezogen werden (z. B. Faktorenanalyse).

Seit der Version 9 stellt SPSS jedoch auch einige statistische Analysen zur multivariaten Untersuchung nominaler und ordinaler Zusammenhänge zur Verfügung (z. B. multinominale logistische Regression), die visuelle und auch mathematische Anschaulichkeit ist allerdings in vielen Fällen etwas eingeschränkt, weshalb diese Verfahren weniger für das Verständnis von Zusammenhängen bei statistischen Laien (z. B. Kunden der Marktforschung) herangezogen werden.

Zusätzlich kann es bei einigen Verfahren zu Verzerrungen kommen, die im Nachhinein nicht oder nur sehr schwer überprüft werden können[19]:

> „Nominale Merkmale, die mehr als zwei mögliche Merkmalsausprägungen haben, werden in binäre (Hilfs-)Variablen zerlegt, und jeder Merkmalsausprägung (Kategorie) wird entweder der Wert 1 (Eigenschaft vorhanden) oder der Wert 0 (Eigenschaft nicht vorhanden) zugewiesen." (BACKHAUS et al. (2000: 332))

Dabei räumen die Autoren ein, dass „bei großer und unterschiedlich großer Anzahl von Kategorien solche Ähnlichkeitsmaße zu starken Verzerrungen führen können." (BACKHAUS et al. (2000: 332), vgl. auch BACKHAUS et al. (2004: 420))[20]. Da ein Großteil der Auswertungen für diese Arbeit auf nominalskalierten Daten beruhen, scheinen Entscheidungsbäume bessere Kriterien für Gruppenbildung zu liefern als o. g. Verfahren, da sie mit allen Skalenniveaus umgehen können. Sie sind voraussetzungsfrei hinsichtlich der Verteilung (z. B. keine Nor-

19. Bei einer dichotomen Variablen (z. B. männlich - weiblich) gibt es zwei Ausprägungen. Es kann weder eine Reihenfolge, noch können Abstände (z. B. „weiblich ist doppelt so gut wie männlich") gebildet werden. Dichotomisierung heisst in diesem Fall, dass die Variablenausprägung männlich den Wert 0, weiblich den Wert 1 erhält und ein kontinuierlicher Bereich zwischen 0 und 1 unterstellt wird (z. B. 0.1, 0.2, 0.3, ...), der natürlich nicht existiert.

20. Diesem Problem kann teilweise durch unterschiedliche Gewichtungen entgegengewirkt werden. Wenn z. B. bei der Sonntagsfrage fünf Parteien (CDU/CSU, SPD, GRÜNE, FDP, PDS) zur Auswahl stehen, ergeben sich bei Dichotomisierung fünf Variablen (CDU/CSU ja - nein, SPD ja - nein, usw.). Werden diese fünf Dummy-Variablen nicht mit 1, sondern nur mit jeweils 0.2 gewichtet, liegt die Gewichtung ebenfalls wieder bei (5 * 0.2 =) 1.

malverteilung der unabhängigen Variablen wie die Diskriminanzanalyse) - und auch die Art der Beziehungen (linear, loglinear, ...) spielen keine Rolle wie z. B. bei der Regressionsanalyse. Der Vergleich der unterschiedlichen Kennzahlen wird Differenzen offenlegen.

1.5 Deskriptive Verfahren: Verwendete statistische Masse

Die deskriptive (= beschreibende) Statistik unterscheidet zwischen uni- und bivariaten Häufigkeitsverteilungen. Im ersten Fall wird eine, im letzten werden zwei Variablen betrachtet.

ABBILDUNG 2 PC-Nutzung: Häufigkeitsverteilung (N = 2048)

pcuser Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,00 Non User	1054	51,5	51,7	51,7
	2,00 User	984	48,1	48,3	100,0
	Gesamt	2038	99,6	100,0	
Fehlend	7,00	9	,4		
Gesamt		2047	100,0		

Bei dieser univariaten Verteilung der Ausprägungen einer Variablen werden die einzelnen Kategorien dargestellt: von 2047 Befragten insgesamt sind 1054 PC-Nichtnutzer, 984 User. 9 Personen haben diese Frage nicht beantwortet. Die Spalte Prozent gibt die Anteile der jeweiligen Kategorien einschließlich der fehlenden Werte zur Basis 100 an, die Spalte „Gültige Prozent“ zeigt die jeweiligen Prozentwerte, bezogen auf die gültigen Antworten (N = 2038). Die letzte Spalte gibt die kumulierten („aufaddierten“) Prozentwerte an.

Häufig ist es jedoch interessant, Zusammenhänge zwischen zwei Variablen (= bivariate Betrachtung) herzustellen, um nähere Informationen über die PC-Nutzer zu erhalten: sind sie eher jünger oder älter, eher männlich oder weiblich, etc.?

Für diese Problemstellung gibt es sog. „Kreuztabellen“, die zwei Variablen horizontal und vertikal darstellen - hier am Beispiel von PC-Nutzung und Geschlecht.

ABBILDUNG 3 HÄUFIGKEIT DER PC-NUTZUNG NACH GESCHLECHT (N = 2038)

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) * D10 Geschlecht Kreuztabelle

Anzahl

		D10 Geschlecht		
		männlich	weiblich	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Non User	445	609	1054
	User	514	470	984
Gesamt		959	1079	2038

Im Vergleich zur Häufigkeitstabelle fällt auf, dass die fehlenden Werte nicht in der Kreuztabelle auftauchen. Die Basis ist 2038 Fälle (unten rechts). Es gibt 445 nichtnutzende Männer und 609 Frauen, die ebenfalls nicht mit dem PC umgehen - im Gegensatz zu ihren 514 männlichen und 470 weiblichen PC-nutzenden Pendants. Die reinen Werte sagen jedoch wenig über die jeweiligen Anteile aus, da die Gesamtzahl der Männer mit 959 leicht unter denen der Frauen mit 1079 liegt. Bezogen auf die Frage, ob der Anteil der Männer oder der Frauen an der PC-Nutzung höher ist, gibt es die Möglichkeit, die Absolutwerte zu prozentuieren:

ABBILDUNG 4 HÄUFIGKEIT DER PC-NUTZUNG NACH GESCHLECHT (N, SPALTENPROZENTE)

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) * D10 Geschlecht Kreuztabelle

			D10 Geschlecht		
			männlich	weiblich	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Non User	Anzahl	445	609	1054
		% von D10 Geschlecht	46,4%	56,4%	51,7%
	User	Anzahl	514	470	984
		% von D10 Geschlecht	53,6%	43,6%	48,3%
Gesamt		Anzahl	959	1079	2038
		% von D10 Geschlecht	100,0%	100,0%	100,0%

Während rund 47 % der Männer Nonuser sind, nutzen 56 % der Frauen keinen PC. Folglich liegt der Anteil in den Nutzerkategorien bei den Männern mit rund 54 % etwa 10 % höher als bei den Frauen mit etwa 44 %.

Diese Werte zeigen, dass Männer den PC häufiger nutzen als Frauen - ist ein Prozentsatz von rund 10 % gerechtfertigt, um zu sagen, Männer nutzen deutlich häufiger den PC als Frauen?

Für diese Fragestellung wurden Kennzahlen - sog. Zusammenhangsmaße - entwickelt, die anhand der Verteilungen der beiden Variablen eine Kennzahl errechnen, mit der die Frage beantwortet werden kann.

Während bei ordinalen und metrischen Variablen die Richtung eines Zusammenhangs berechnet werden kann (z. B. mit zunehmenden Lebensalter - gemessen in Jahren - steigt das verfügbare Haushaltsnettoeinkommen - gemessen in Euro), ist das bei nominalen Variablen, die keine Rangfolge besitzen, nicht möglich. Bei der Frage nach Einkommen und Alter kann der Zusammenhang positiv sein (je älter eine Person ist, desto höher ist auch das verfügbare Haushaltsnettoeinkommen). Es kann kein Zusammenhang existieren - wenn

zum Beispiel alle Personen ein gleiches Einkommen haben, unabhängig vom Lebensalter - oder es kann negativ sein, wenn vor allem jüngere Befragte höhere, Ältere geringere Einkünfte beziehen.

Nominale Zusammenhangsmaße können nur die Ähnlichkeit bzw. Unähnlichkeit zweier Variablen messen: sind sich PC-Nutzer hinsichtlich des Geschlechts ähnlich (Männer und Frauen nutzen den PC annähernd gleich) oder unähnlich (Männer (oder Frauen) nutzen den PC wesentlich häufiger)? Eine Relation wie beim Einkommen kann es nicht geben, da zwischen nominalen Variablen keine Rangfolge gebildet werden kann.

Als statistische Kennzahlen für nominale Zusammenhänge werden in dieser Arbeit Cramers v (bzw. Phi für Vier-Felder-Tafeln) als Stärke des Zusammenhangs[21] und der Unsicherheitskoeffizient (als Richtungsmaß) herangezogen. Diese beiden Kennzahlen beruhen auf unterschiedlichen Logiken: während Cramers v durch die Chi-Quadrat-Statistik gebildet wird, basiert der Unsicherheitskoeffizient - wie Lambda oder Goodman & Kruskals tau - auf dem Konzept der sog. „proportionalen Fehlerreduktion“ (vgl. BÜHL & ZÖFEL (2002a: 244ff.), JANN (2002: 75ff)).[22]

1.5.1 Chi-Quadrat-basierte Maße

Cramers v bzw. Phi können einen Wert zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang) annehmen, wobei es bei Cramers v durch die fehlende Rangfolge keine nähere Beschreibung

21. Für Vierfeldertafeln wurde eine spezielle Kennzahl Phi entwickelt, die ebenfalls zum Einsatz kommt. Vier-Felder-Tafeln geben zwei Ausprägungen der abhängigen und zwei der unabhängigen Variablen wieder, so dass sich vier Felder ergeben. Ein Beispiel hierfür ware PC-Nutzung (ja - nein) und Geschlecht (männlich - weiblich).

22. Die Aussagekraft der Chi-Quadrat-basierten Maße ist durch den Informationsgehalt von Nominaldaten geringer als bei metrischen Skalen. Aufgrund der Unzufriedenheit mit der Interpretierbarkeit schlugen GUTTMAN, GOODMAN und KRUSKAL in den 40er und 50er Jahren des letzten Jahrhunderts die PRE-Maße vor, die klare Schlüsse der unabhängigen auf die abhängige Variable zuläßt (vgl. BENNINGHAUS (1979: 84ff.)).

der Art der Beziehung gibt.[23] Ein Zusammenhang ab 0.2 gilt als geringer Zusammenhang (vgl. BÜHL und ZÖFEL (2002a: 243)), der in den Sozialwissenschaften als bedeutsam gilt. Allerdings sollten gerade Nominalmaße mit Vorsicht interpretiert werden, da sich die Stärke des Zusammenhangs eher über den Vergleich verschiedener Tabellen als über die absolute Stärke der Korrelation ergibt (vgl. NORUSIS (1998: 354)). Für dichotome (0, 1) Variablen kann durch den Phi-Koeffizienten die Richtung bestimmt werden, wie im nachfolgenden Beispiel beschrieben. Im Gegensatz zum Kontingenzkoeffizienten ist Cramers v nicht von der Größe der Tabelle abhängig.

ABBILDUNG 5 Formel für Cramers v (vgl. BENNINGHAUS 1979: 109 bzw. 100)

$$v = \sqrt{\frac{x^2}{N \min(r-1, c-1)}}$$

Der Chi Quadrat-Wert wird ausführlich auf Seite 63 beschrieben. BENNINGHAUS erläutert den Nenner:

> „... wobei r die Anzahl der Zeilen und c die Anzahl der Spalten bezeichnet. Der Ausdruck 'min' steht für Minimum und besagt, daß zunächst zu prüfen ist, ob die Anzahl der Zeilen oder der Spalten geringer ist; der kleinere Wert geht in die Berechnung des Koeffizienten ein." (BENNINGHAUS (1979: 109))

Im Fall ordinaler bzw. metrischer Variablen läßt sich eine Richtung des Zusammenhangs durch den PEARSONschen Korrelationskoeffizienten r bzw. KENDALLs tau b bzw. SPEARMANs Rho ermitteln, der zwischen -1 (negativer Zusammenhang), 0 (kein Zusammenhang) und 1 (perfekter positiver Zusammenhang) liegt.

Für die Frage nach PC-Nutzung und Geschlecht ergibt sich:

23. In neueren SPSS-Versionen kann Phi zwischen -1 (negativer Zusammenhang) und +1 (perfekter positiver Zusammenhang) liegen. Weitere Erläuterungen folgen im Text.

ABBILDUNG 6 NOMINALE ZUSAMMENHANGSMASSE ZWISCHEN PC-NUTZUNG UND GESCHLECHT (PHI, CRAMERS V)

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	-,100	,000
	Cramer-V	,100	,000
Anzahl der gültigen Fälle		2038	

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

Da es sich bei der Auswertung um eine sog. 4-Felder-Tafel handelt (PC-Nutzung: ja - nein, Geschlecht: männlich - weiblich), wird Phi anstatt Cramers v interpretiert. Der Wert ist mit 0.1 identisch, allerdings hat Phi ein negatives Vorzeichen.

Phi errechnet sich nach folgender Formel:

ABBILDUNG 7 Formel für Phi (vgl. BENNINGHAUS (1979: 100)

$$\Phi = \frac{N(ad-bc)^2}{(a+b)(c+d)(b+c)(b+d)}$$

Die Besonderheit der Richtungsangabe des Zusammenhangs ist nur bei dichotomen Variablen möglich. Die Richtung wird durch die Kodierung der Variablen erzeugt. Im Beispiel ist die Nichtnutzung mit 0, die Nutzung mit 1 definiert (männlich = 0, weiblich = 1). Eine positive Korrelation würde vorliegen, wenn die Werte bei den mit 0 kodierten Variablen kleiner sind als die mit 1 ausgegebenen - also wenn Männer häufiger Nichtnutzer wären als Frauen. Die Tabelle macht aber deutlich, dass Männer (mit 0 kodiert) häufiger den PC nutzen (mit 1 kodiert) als Frauen.

Die Stärke des Zusammenhangs, repräsentiert durch Phi, ist mit -0.100 sehr gering: es gibt also keinen starken Zusammenhang zwischen Geschlecht und PC-Nutzung.

Daraus wird deutlich, dass Phi kein „echtes" Richtungsmaß ist, da es nur auf zwei dichotomen Ausprägungen von Variablen beruht. Es gibt nur an, ob die Reihenfolge der Kodierung gleichlaufend ist (positives Phi) oder nicht (negatives Phi).

Der Chi-Quadrat-Wert ist ein Ähnlichkeits-Unähnlichkeits-Maß und vergleicht die Randsummen von Merkmalen. „Ähnlichkeit/Unähnlichkeit" bezieht sich hier auf die (prozentuale) Ausprägungen der Merkmale, die als Kriterium herangezogen werden. Da bei Nominalskalen keine Rangfolge gebildet werden kann und die Abstände unbekannt sind (z. B. beim Familienstand: ledig - verheiratet - geschieden - verwitwet oder der PC-Nutzung: Nutzer - Nichtnutzer), wird die Ähnlichkeit - Unähnlichkeit an der Besetzung der einzelnen Zellen untersucht.

Die in der Formel für Phi angegebenen Platzhalter a, b, c und d werden auf die Vierfeldertabelle folgendermaßen verteilt: a = männlich/nonuser, b = weiblich/nonuser, c = männlich/user und d = männlich/nonuser:

ABBILDUNG 8 ERWARTETE UND BEOBACHTETE HÄUFIGKEITEN ZWISCHEN GESCHLECHT UND HÄUFIGKEIT DER PC-NUTZUNG

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) * D10 Geschlecht Kreuztabelle

			D10 Geschlecht		
			männlich	weiblich	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Non User	Anzahl	445	609	1054
		Erwartete Anzahl	496,0	558,0	1054,0
	User	Anzahl	514	470	984
		Erwartete Anzahl	463,0	521,0	984,0
Gesamt		Anzahl	959	1079	2038
		Erwartete Anzahl	959,0	1079,0	2038,0

Bei völligem „Unwissen“ über die Verteilung einer Variablen würde man vereinfachend eine Gleichverteilung unterstellen und ganz bestimmte Werte **erwarten**, die in der Tabelle als Dezimalzahlen im unteren Teil jeder Zelle ausgegeben ist. Beim Beispiel Geschlecht (männlich - weiblich) und PC-Nutzung (ja - nein), würde man, ausgehend von den Randsummen der Tabelle **erwarten,** dass z. B. aus 1054 Nichtnutzern und 959 Männern 496 männliche Nonuser resultieren. Diese Zahl errechnet sich proportional aus den Randsummen:

```
1054 (Nonuser ges.) * 959 (Männer ges.) / 2038 (Stichprobe ges.).
```

Diese Berechnung wird nun für alle Zellen durchgeführt. Empirisch **(= beobachtet)** stellt sich nun heraus, dass **tatsächlich in der Stichprobe** 445 Männer den PC nicht verwenden. Durch die Berechnung der Randsummen würden aber hier 496 männliche User **erwartet.** Die Differenz bezeichnet man als Residuen (= -51). Anders ausgedrückt: ohne jegliches Vorwissen würde man für diese Zelle 51 Männer mehr erwarten, die den PC nutzen.

ABBILDUNG 9 UNSTANDARDISIERTE ERWARTETE UND BEOBACHTETE HÄUFIGKEITEN ZWISCHEN GESCHLECHT UND HÄUFIGKEIT DER PC-NUTZUNG

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) * D10 Geschlecht Kreuztabelle

			D10 Geschlecht		
			männlich	weiblich	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Non User	Anzahl	445	609	1054
		Erwartete Anzahl	496,0	558,0	1054,0
		Residuen	-51,0	51,0	
	User	Anzahl	514	470	984
		Erwartete Anzahl	463,0	521,0	984,0
		Residuen	51,0	-51,0	
Gesamt		Anzahl	959	1079	2038
		Erwartete Anzahl	959,0	1079,0	2038,0

Ein Blick auf die Tabelle macht deutlich, dass sowohl dieser Wert als auch alle anderen Werte für alle Zellen um 51 von den erwarteten Werten differieren, die Abweichung ist aber nicht sehr stark. Somit läßt sich nicht eindeutig feststellen, ob die erwarteten von den beobachteten Häufigkeiten stark abweichen, da sich für jede Tabelle mit unterschiedlicher Fallzahl unterschiedliche Differenzen zwischen erwarteten und beobachteten Werten ergeben. Deshalb wurde eine Standardisierung entwickelt, die auf der Normalverteilung beruht.

ABBILDUNG 10 STANDARDISIERTE ERWARTETE UND BEOBACHTETE HÄUFIGKEITEN ZWISCHEN GESCHLECHT UND HÄUFIGKEIT DER PC-NUTZUNG

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) * D10 Geschlecht Kreuztabelle

			D10 Geschlecht		
			männlich	weiblich	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Non User	Anzahl	445	609	1054
		Erwartete Anzahl	496,0	558,0	1054,0
		Standardisierte Residuen	-2,3	2,2	
	User	Anzahl	514	470	984
		Erwartete Anzahl	463,0	521,0	984,0
		Standardisierte Residuen	2,4	-2,2	
Gesamt		Anzahl	959	1079	2038
		Erwartete Anzahl	959,0	1079,0	2038,0

Diejenigen Zellen, die einen Absolutwert > 2 (positiv wie auch negativ) aufweisen, sind für die Stärke des Zusammenhangs verantwortlich. Die Standardisierung bedeutet in diesem Zusammenhang eine sog. z-Transformation. Vereinfacht ausgedrückt: die Werte werden in eine Standardnormalverteilung („Glockenkurve“) überführt. Die Standardisierung entsteht durch die Symmetrie der Glockenkurve: der Mittelwert ist immer 0, die Standardabweichung = 1. Weichen nun Werte deutlich von diesem Mittel ab ($\geq$ 1.96), so sind sie für die Stärke des Zusammenhangs verantwortlich. Anschaulich ausgedrückt unterstellt die Normalverteilung, dass die mittleren Kategorien deutlich höher besetzt sind als die Randkategorien. Dies läßt sich auch empirisch finden, wenn man z. B. die Körpergröße von Erwachsenen betrachtet, die durchschnittlich bei Frauen etwas geringer ist als bei Männern. So kann man allgemein sehen, dass es mehr Männer mit einer Größe von z. B. 175 cm gibt als mit 155 cm oder 225 cm. Die „Durchschnittskategorien“ sind also deutlich höher besetzt als die

Randkategorien. Dieser Sachverhalt wird durch die Glockenkurve ausgedrückt.

Der Chi-Quadrat-Test „vergleicht“ nun die erwarteten und die beobachteten Häufigkeiten („Wie weit weichen 496 von 445, 609 von 558, ... ab?“ „Wie ähnlich bzw. unähnlich sind sich die erwarteten bzw. beobachteten Häufigkeiten?“). Wären die Werte identisch, würden also z. B. 496 männliche Nichtnutzer erwartet und könnte dies auch durch die empirische Untersuchung bestätigt werden, gäbe es keine Unterschiede zwischen den Ausprägungen, so ist der Chi-Quadrat-Wert 0. Je mehr jedoch diese Kategorien voneinander abweichen (wären also - bei gleichen Randsummen - von den 959 Männern z. B. 900 Nichtnutzer, würde sich der Chi-Quadrat-Wert immer weiter von 0 entfernen. Damit wären die Ergebnisse nicht mehr ähnlich, sondern unähnlich, was durch den Koeffizienten Cramers v (bzw. Phi für Vierfeldertabellen) ausgedrückt wird, der zwischen 0 und 1 liegen kann. Anders ausgedrückt: wenn sich die empirisch beobachteten Häufigkeiten proportional zu den Zeilen- bzw. Randsummen verhalten, besteht kein Zusammenhang.

ABBILDUNG 11 CHI-QUADRAT BASIERTE STATISTISCHE ZUSAMMENHANGSWERTE ZWISCHEN GESCHLECHT UND HÄUFIGKEIT DER PC-NUTZUNG

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	20,491[b]	1	,000		
Kontinuitätskorrektur[a]	20,091	1	,000		
Likelihood-Quotient	20,520	1	,000		
Exakter Test nach Fisher				,000	,000
Zusammenhang linear-mit-linear	20,481	1	,000		
Anzahl der gültigen Fälle	2038				

a. Wird nur für eine 2x2-Tabelle berechnet

b. 0 Zellen (,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 463,03.

Da die erwarteten Werte nicht sehr stark von den beobachteten abweichen, liegt der Chi-Quadrat-Wert mit rund 20 in der Nähe von 0. Das Ergebnis ist durch die hohe Fallzahl signifikant. Dies bedeutet, dass sich die PC-Nutzung - hinsichtlich Männer und Frauen - nicht deutlich unterscheidet - weder in der Stichprobe noch in der Grundgesamtheit.[24]

Berechnet wird der Wert folgendermaßen:

$$2038\ (445 * 470 - 609 * 514)^2 / 959 * 979 * 1054 * 984 = 20.491.$$

Die gegensätzlichen Ausprägungen der Variablen (nichtnutzende Männer - nutzende Frauen und umgekehrt) werden multipliziert, um negative Werte zu vermeiden quadriert und mit dem Umfang der Stichprobe multipliziert.[25] Dieses Ergebnis wird durch die multiplizierten Randsummen geteilt. Der Chi-Quadrat-Wert von 20 weicht nicht

24. FISHERs Exact Test wird nur bei kleinen Stichproben herangezogen. Der Zusammenhang linear mit linear gilt nur für Ordinaldaten, da hier untersucht wird, ob eine monotone Steigung vorliegt. SPSS gibt die Werte jedoch immer standardmäßig aus.

sehr weit von 0 ab - und auch nicht von den erwarteten Häufigkeiten. Er ist dafür hochsignifikant - das Ergebnis kommt also nicht zufällig zustande[26]. Daraus folgt, dass die (generelle) PC-Nutzung keine „Männerdomäne" ist.

Angenommen, die PC-Nutzung wäre - wie in nachfolgender Tabelle unterstellt - eher eine Männerdomäne, so würden sich die erwarteten und beobachteten Häufigkeiten deutlich unterscheiden, da die erwarteten Häufigkeiten identisch mit obiger Tabelle sind. Zum Beispiel würden 496 nichtnutzende Männer erwartet, es sind jedoch nur 54. Somit weichen die erwarteten von den beobachteten Häufigkeiten sehr stark ab.[27]

ABBILDUNG 12 HOHE ABWEICHUNGEN: PC-NUTZUNG NACH GESCHLECHT (N = 2038)

	Männer	Frauen	gesamt
PC-Nichtnutzer	54	1000	1054
PC-Nutzer	905	79	984
	959	1079	2038

Der Chi-Quadrat-Wert verändert sich drastisch:

$$2038\ (54 * 79 - 1000 * 905)^2 / 1054 * 984 * 959 * 1079 = 1540.71.$$

Der Chi-Quadrat-Wert, der im Datensatz bei rund 20 lag, steigt nun mit der künstlichen Veränderung auf etwa 1549 - ein sehr großen Un-

25. Eine Schwäche des Chi-Quadrat-Wertes liegt darin, dass durch die Einbeziehung des Stichprobenumfangs N der Wert beeinflusst wird. Aus diesem Grund läßt sich nicht mehr so leicht von einem „hohen" oder „niedrigen" Chi Quadrat-Wert sprechen, sondern er muss immer in Abhängigkeit des Stichprobenumfangs und den Ergebnissen interpretiert werden.

26. Da die Berechnung des Signifikanzwertes von der Stichprobengröße abhängt, werden häufig bei großen Stichproben die meisten Ergebnisse signifikant. Aus diesem Grund wird im folgenden nur noch darauf verwiesen, wenn Werte nicht signifikant sind.

27. Um in der Terminologie des Chi-Quadrat-Werts zu bleiben, wurden die Begriffe der „erwarteten" und „beobachteten Häufigkeiten" sprachlich nicht variiert.

terschied zwischen PC-nutzenden Männern und nichtnutzenden Frauen läge vor.

In den Sozialwissenschaften wird ein Wert von > .20 für metrisch skalierte Variablen als bedeutsamer Zusammenhangswert angesehen. Für Nominaldaten ist dies aber leider nicht so ohne weiteres übertragbar (vgl. NORUSIS (1998: 354), BROSIUS (1998: 412)):

> „Chi-square-based measures are difficult to interpret. Although they can be used to compare the strength of association in different tables, the strength of association being compared isn´t easily related to an intuitive concept of association." (NORUSIS (1998: 354))

Für diese Arbeit bedeutet das, dass alle sinnvollen unabhängigen Variablen mit der PC-Nutzung korreliert werden und eine Rangfolge der Zusammenhänge gebildet wird. Je höher der Zusammenhang, desto bedeutsamer ist die abhängige Variable für die Analyse. Allerdings wird unterstellt, dass alle Kultur- und Freizeitvariablen keinen dominanten Charakter haben können, da sie von sog. „alten", vertikalen Ungleichheiten wie Geschlecht, Bildung oder beruflicher Stellung abhängig sind. PRE-basierte Maße

1.5.2 PRE-Maße

Im Gegensatz zu den Chi-Quadrat-basierten Maßen (u. a. Phi und Cramers v) verwenden PRE-Maße (Unsicherheitskoeffizient, Lambda) einen anderen Zugang zur Assoziation zweier Variablen:

> „**Proportional reduction in Error** (PRE) measures, unlike chi-square-based measures, have a clear interpretation. They look at how much better you can predict the values of a dependent variable when you know the values of an independent variable. PRE measures compare the errors in two different situations: one where you don't use the independet variable for prediction and one where you do." (NORUSIS (1998: 354))

Während Chi-Quadrat-basierte Maße die Stärke eines Zusammenhangs angeben, bilden PRE-Maße die Richtung eines Zusammenhangs ab. Die dahinterstehende Frage lautet: „Um wieviel (Prozent)

kann ich die Ausprägungen einer abhängigen Variable vorhersagen, wenn ich die Ausprägungen der unabhängigen Variable kenne“?

ABBILDUNG 13 HÄUFIGKEIT DER PC-NUTZUNG NACH GESCHLECHT (N, SPALTENPROZENTE)

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) * D10 Geschlecht Kreuztabelle

			D10 Geschlecht		
			männlich	weiblich	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Non User	Anzahl	445	609	1054
		% von D10 Geschlecht	46,4%	56,4%	51,7%
	User	Anzahl	514	470	984
		% von D10 Geschlecht	53,6%	43,6%	48,3%
Gesamt		Anzahl	959	1079	2038
		% von D10 Geschlecht	100,0%	100,0%	100,0%

Die abhängige Variable ist die PC-Nutzung, die unabhängige Variable das Geschlecht. Ohne Einbeziehung des Geschlechts sind 51.7 % der Befragten PC-Nutzer, 48.3 % Nonuser (Gesamt-Spalte).

Könnte man aufgrund dieser Information - ohne Einbeziehung des Geschlechts - die PC-Nutzung perfekt voraussagen, z. B. dass alle Männer PC-Nutzer sind und alle Frauen nicht, würde man einen Zusammenhang von 1 erhalten - die abhängige Variable ließe sich perfekt durch die unabhängige vorhersagen.

Der Unsicherheitskoeffizient (vgl. JANN (2002: 50ff. und 78ff.), BAUR (2003: 20ff.), LIENERT (1973: 580ff.)), der auch unter dem Begriff der “Transinformation“ (vgl. LIENERT (1973: 581), JANN (2002: 78)) bekannt ist (und in dieser Arbeit zum Einsatz kommt), berechnet die statistische (Un-)Abhängigkeit einer abhängigen bezogen auf eine unabhängige Variable. Bei einem Wert von 1 läßt sich die abhängige Variable perfekt durch die unabhängige Variable erklären, bei 0 wird

durch die unabhängige Variable nichts erklärt (vgl BAUR (2003: 26)). [28]

Die Formel lautet:

ABBILDUNG 14 Formel für PRE-Maße (vgl. BAUR (2003: 26))

$$PRE = \frac{Va - Vb}{Va}$$

Va bezeichnet den Vorhersagefehler ohne Kenntnis (der PC-Nutzung), Vb den Vorhersagefehler bei Kenntnis der PC-Nutzung, dividiert durch den Vorhersagefehler ohne Kenntnis der PC-Nutzung. Wenn man also das Geschlecht kennt - mit wieviel Prozent Wahrscheinlichkeit mache ich weniger Fehler bei der Vorhersage der PC-Nutzung?

ABBILDUNG 15 PRE BASIERTE STATISTISCHE ZUSAMMENHANGSWERTE ZWISCHEN GESCHLECHT UND HÄUFIGKEIT DER PC-NUTZUNG

Richtungsmaße

			Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Unsicherheitskoeffizient	Symmetrisch	,007	,003	2,270	,000[c]
		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) abhängig	,007	,003	2,270	,000[c]
		D10 Geschlecht abhängig	,007	,003	2,270	,000[c]

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Chi-Quadrat-Wahrscheinlichkeit für Likelihood-Quotienten.

Im o. g. Beispiel läßt sich die Variable PC-Nutzung nicht sehr gut durch das Geschlecht (7 %) vorhersagen. Der Unsicherheitskoeffizi-

28. Auf die Darstellung der Formeln zur Berechnung des Unsicherheitskoeffizienten wird verzichtet, da sie sehr komplex sind und wenig zur Anschaulichkeit des Ergebnisses beitragen.

ent, der keine Verteilungsannahmen macht, ist ein Vorhersagemaß, mit wieviel Prozent Wahrscheinlichkeit sich die Ausprägung einer Variable durch eine andere voraussagen läßt. Hierbei sind zwei Maße zu unterscheiden: der symmetrische Unsicherheitskoeffizient, der zur Anwendung kommt, wenn nicht geklärt werden kann, welche Variable abhängig und welche unabhängig ist. Die asymmetrischen Kennzahlen werden in den Zeilen PC-Nutzung bzw. Geschlecht abhängig, angegeben. In diesem Fall kann entschieden werden, ob die PC-Nutzung oder das Geschlecht abhängig ist (hier: PC-Nutzung).

Die Logik hinter Chi-Quadrat und PRE-Maßen ist somit ähnlich: bei ersteren Maßzahlen werden (durch die Spalten- und Zeilensummen) erwartete und beobachtete (= empirisch erhobene) Häufigkeiten verglichen. Je geringer die Unterschiede, desto mehr tendiert Chi Quadrat gegen 0. Bei den PRE-Maßen werden die Zeilen-/Spaltensummen einer Variablen verglichen mit den Spalten- und Zeilensummen einer zusätzlichen Variablen. Sind die Unterschiede hoch, dann erhöht sich der Prozentsatz, mit der die Ausprägungen der eine Variable durch die andere vorhergesagt werden kann.

Im Gegensatz zu den PRE-Maßen GOODMAN & KRUSKAL s TAU liefert der Unsicherheitskoeffizient genauere Ergebnisse, da er nicht nur die Zellen mit der größten Häufigkeit bzw. die Randsummen untersucht, sondern eine Analyse jeder Zelle vornimmt. Dadurch ist der Informationsgehalt wesentlich höher, allerdings die Berechnung auch komplexer - was allerdings heute durch die Möglichkeiten der EDV keinen Nachteil darstellt.

1.6 Multivariate Verfahren

Multivariate untersuchen - im Gegensatz zu bivariaten Verfahren - Zusammenhänge zwischen mehr als zwei Variablen. Hierbei muß zwischen Methoden (z. B. Faktoren-, Clusteranalyse), die alle in die

Analyse eingehenden Variablen miteinander in Beziehung setzt und Verfahren unterschieden werden, die untersuchen, wie mehr als zwei unabhängige Variablen mit einer abhängigen Variable kausal zusammenhängen. War die Fragestellung bei der bivariaten Analyse „Wie stark hängt die PC-Nutzung vom Geschlecht ab?“, lautet hier z. B. die Frage: „Welche der unabhängigen Variablen Alter, Haushaltsnettoeinkommen, berufliche Stellung, Geschlecht und Schulbildung beeinflussen wie stark und ggfs. in welche Richtung (unter Konstanthaltung der anderen Variablen) die PC-Nutzung“? - Rechnernutzung wird also nicht anhand einer, sondern mehrerer Variablen erklärt.

Wie hoch die Erklärungskraft von dominanten Variablen, bezogen auf die PC-Nutzung ist, soll kurz am Beispiel der linearen Regression aufgezeigt werden. Hierbei wird unerlaubterweise auf die ordinal skalierte PC-Nutzungsvariable zurückgegriffen - lineare Regression setzt eigentlich metrisches Skalenniveau bei der abhängigen Variablen voraus.

ABBILDUNG 16 Lineare Regression: Erklärungskraft sozialstruktureller Variablen (Alter, Beruf, Haushaltsnettoeinkommen, Bildung und Lebensgemeinschaft

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,666[a]	,444	,442	1,64405

a. Einflußvariablen : (Konstante), D8: Schulbildung der Befragten, D7: Lebensgemeinschaft (recodiert), D29: Haushaltsnettoeinkommen, D15AR: Berufliche Stellung, D 11 Alter

Aus der Abbildung geht hervor, dass die PC-Nutzung durch Alter Beruf, Haushaltsnettoeinkommen, Schulbildung und die Art der Lebensgemeinschaft mit R-Quadrat = .444 erklärt werden kann. Das bedeutet: von allen Abweichungen vom Mittelwert der Variable PC-Nutzung können 44.4 % durch die unabhängigen Variablen in diesem Modell erklärt werden.

An dieser Stelle ist deutlich darauf hinzuweisen, dass diese Prozentangaben sich grundsätzlich auf das Modell, nicht auf die Realität beziehen. PC-Nutzung hat viele Einflußfaktoren ob persönliche oder berufliche sei dahingestellt. Sie können durch keinen Datensatz vollständig repliziert werden. Dieses Ergebnis bedeutet also, dass die unabhängigen Variablen, die in die Analyse eingehen, rund 44 % des Modells erklären - zu 56 % müssen andere Variablen herangezogen werden.

Wenn rund 40 % des Modells durch nur vier sozialstrukturelle Variablen erklärt werden deutet dies auf eine deutliche Vormachtstellung „alter Ungleichheiten“ hin. Wie eingangs erläutert, kosten Freizeitaktivitäten Geld - und sind damit in gewisser Weise vom Einkommen abhängig. Ein Teil der Freizeitaktivitäten (z. B. Kinobesuch) könnte deutlich altersspezifisch sein. Als weiteres vertikales Merkmal wird aus diesem Grund ausdrücklich das Alter einbezogen, das heute häufig (vor allem im Bezug auf Beruf, Techniknutzung und Aneignung neuen Wissens) sozialstrukturell wirkt - auch wenn SCHULZE Alter noch als horizontales Merkmal definiert hat. Heute wirkt insbesondere das Alter im Berufsleben eher negativ, da höheres Alter nicht mehr mit „Erfahrung“, sondern mit „Nicht-mehr-Schritt-halten-können“ gleichgesetzt wird. So zeigen Untersuchungen des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) in Nürnberg, dass die Zahl der älteren Arbeitslosen (55 Jahre und älter) in den 90er Jahren deutlich anstiegen (vgl. KOLLER et al. (2003: 19ff.)). Dies ist nicht nur auf den Umgang mit dem PC,

sondern auch auf zunehmend verschärfte Globalisierungsbedingungen zurückzuführen, die gute Fremdsprachenkenntnisse, Bereitschaft zu (weltweiter) Mobilität, Gesundheit, Überstunden, Anpassung an fremde Kulturen und eine Leistungsethik voraussetzen, die weniger sozialstaatlich bzw. familiär als vielmehr verantwortlich für sich persönlich und sowohl von materiellen als auch von vermeintlichen Selbstverwirklichungszielen gekennzeichnet ist. Diese Haltung traut man deutlich häufiger eher jüngeren als älteren Menschen zu.

Neben den sog. „Kausalmodellen" in der multivariaten Statistik gibt es Verfahren (z. B. Faktoren-, Clusteranalyse), die nicht von einer abhängigen und verschiedenen unabhängigen Variablen ausgehen, sondern die alle Variablen miteinander in Beziehung setzen.

Ein Beispiel für nichtkausale multivariate Fragestellungen ist z. B. die Untersuchung von Itembatterien über politische Einstellungen mit Hilfe der Faktorenanalyse. Hier lautet das Forschungsziel: welche Items werden ähnlich beantwortet? - So ist bei den Aussagen „Bei der Arbeitsplatzvergabe sollten Deutsche vor Ausländern berücksichtigt werden" und „Ausländer sollten keine staatliche Unterstützung erhalten" zu erwarten, dass Menschen mit ähnlicher politischer Einstellung diese Fragen eher bejahen oder verneinen. Eine abhängige Zielvariable gibt es hier nicht.

Ein in der (quantitativen) Soziologie sehr verbreitetes Verfahren für den Fall der unabhängigen Variablen bezogen auf eine abhängige Variable ist die Regression. Allerdings geht diese Methode einen Schritt weiter und untersucht, wie jede einzelne unabhängige Variable für sich genommen auf die abhängige Variable wirkt, wenn man die anderen unabhängigen Variablen konstant hält. Es ist durchaus denkbar, dass die unabhängigen Variablen untereinander Zusammenhänge aufweisen. Z. B. könnte bei der Fragestellung, wie Alter,

Schulbildung und Geschlecht auf die Rechnernutzung wirken, die Schulbildung auch abhängig vom Alter sein (durch die Bildungsexpansion ist Jüngeren der Weg in weiterführende Schulen ermöglicht worden). Dies kann mit Hilfe der linearen („Wenn die abhängige Variable(n) um eine Einheit steigt, sinkt (bzw. steigt) die unabhängige(n) Variable(n) um eine Einheit"). multiplen Regression - allerdings nur für metrische Daten in SPSS - untersucht werden.[29] Durch das fehlende metrische Skalenniveau bei dem zu untersuchenden Datensatz wird auf die logistische Regression zurückgegriffen, die für Nominaldaten geeignet ist.

Im Mittelpunkt der Methoden dieser Arbeit steht ein, der Regression ähnliches Verfahren, die Entscheidungsbäume, auf das im weiteren detailliert eingegangen wird. Der Vorteil dieser Methode liegt im flexiblen Umgang mit allen Skalenniveaus. Ist die abhängige Variable metrisch, können sog. „Regressionsbäume" generiert werden, bei Ordinal- oder Nominaldaten spricht man von „Klassifikationsbäumen". Ein weiterer Vorteil liegt in der grafischen Aufbereitung der Ergebnisse in Form einer Baumstruktur.

Da die Entscheidungsbäume in der Soziologie bislang kaum Beachtung fanden, soll in dieser Arbeit ein entscheidender Schritt unternommen werden, diese Methoden zu etablieren. Um die Ergebnisse zu kontrollieren bzw. um zu überprüfen, ob Entscheidungsbäume mehr oder weniger Informationen als andere multivariate Methoden liefern, wird auf die multinominale logistische Regression und die Diskriminanzanalyse zurückgegriffen. Diese Methoden können u. a. (dichotome) nominalskalierte Daten untersuchen und vergleichbare

29. Es existiert ein DOS-Programm (LEM), das von Jeroen K VERMUNT entwickelt wurde und Nominaldaten für diese Fragestellung verarbeitet (weitere Informationen siehe unter: http//spitswww.uvt.nl/~vermunt. Auf den Einsatz in dieser Arbeit wurde aber aufgrund der nicht unproblematischen Konvertierung und dem Problem des unter Windows XP sehr restriktiv gehandhabten DOS-Modus, verzichtet.

Kennzahlen liefern. Um festzustellen, ob Entscheidungsbäume statistisch und auch inhaltlich gleiche oder andere (bessere oder schlechtere) Ergebnisse liefern, wird in dieser Arbeit anhand des EUROBAROMETER 56.0-Datensatzes ein Vergleich der Methoden durchgeführt.

TABELLE 4 GRUNDFRAGEN AUSGEWÄHLTER MULTIVARIATER VERFAHREN (VGL. BÜHL UND ZÖFEL (2002A: 329, 431, 487), SPSS (2001B: 5))

Multivariates Verfahren	Erläuterung
Entscheidungsbäume	Wie lassen sich konkrete Gruppen aufgrund unterschiedlicher unabhängiger Variablen bilden? (z. B. 25-35jährige Akademiker mit einem Haushaltsnettoeinkommen zwischen 2300 und 2900 DM)?
Diskriminanzanalyse	Wie stark und in welche Richtung hängen die Funktionen mit den Variablen zusammen? Wo liegen in diesem, durch die Funktionen aufgespannten, Raum die User und die Non-User?
(Multinominale logistische) Regression	Wie stark wirken unahängige auf eine abhängige Variable bei Konstanthaltung der anderen unabhängigen Variablen?

Die Tabelle zeigt, dass alle Verfahren ähnliche Probleme beantworten. Die Fragestellungen sollen an einem Beispiel illustriert werden: gegeben ist die abhängige Variable „PC-Nutzung?“ mit den Ausprägungen Ja und Nein. Diese Variable soll anhand des Bildungsgrades (ordinal) und dem Alter (metrisch) erklärt werden. Dahinter steckt die Frage: sind PC-Nutzer eher älter (oder jünger), sind sie eher besser (oder eher schlechter) gebildet? Läßt sich überhaupt ein Zusammenhang zwischen den unabhängigen Variablen, bezogen auf die abhängige PC-Nutzung finden?[30]

Das Alter wurde metrisch erfaßt. Beim Bildungsabschluss wurde - um den europäischen Vergleich des eingesetzten Datensatzes zu er-

30. Dieses recht einfache Beispiel soll die Methoden für den Leser anschaulich machen.

möglichen - das Alter erfragt, in dem der höchste Schulabschluss erworben wurde. Diese metrische Kategorie beinhaltet einige Ungenauigkeiten, da sie schon innerhalb des deutschen Schulsystems mit unterschiedlichen Schultypen schwer zu übertragen ist. Aus dem bildungsbezogenen Alter wurde deshalb eine ordinale Variable mit den Ausprägungen „Volks-, Hauptschule", „erweiterter Hauptschulabschluss/Mittlere Reife", „(fachgebundene) Hochschulreife", „Student (zum Zeitpunkt der Befragung)" und „Hochschulabsolvent" gebildet.

ABBILDUNG 17 Recodierte Schulbildung der Befragten (N = 2047)

D8: Schulbildung der Befragten

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	bis 15 Jahre (Volks-, Hauptschule)	506	24,7	24,7	24,7
	16 - 17 Jahre (Mittl Reife/erw. HS-Abschluss)	635	31,0	31,0	55,7
	18 - 20 Jahre (Fachgeb.) Hochschulreife	473	23,1	23,1	78,8
	Studium	128	6,3	6,3	85,1
	21 Jahre + abgeschlossenes Hochschulstudium	305	14,9	14,9	100,0
	Gesamt	2047	100,0	100,0	

Die bis 15jährigen wurden als Abgänger der Volks- bzw. Hauptschule angesehen (9jährige Schulausbildung). Nach 10 bzw. 11jähriger Ausbildung wird in Deutschland entweder der Realschul- oder der erweiterte Hauptschulabschluß erreicht (dabei ist ein Jahr berücksichtigt, das durch Wiederholung einer Klasse entstehen kann). Mit 17 Jahren kann aber ansonsten keine Hochschulzugangsberechtigung erworben werden, dies ist in der Regel erst nach 12 bzw. 13 Jahren möglich (Kategorie: 18 - 20 Jahre (fachgeb.) Hochschulreife). Die Studierenden konnten direkt über die Berufsangabe erfaßt werden, so dass alle Befragten über 21 Jahre als Akademiker recodiert wurden. Deutlich wird, dass diese Zuordnung einige Ungenauigkeiten enthält, die sich jedoch auf die Gesamtergebnisse nicht dramatisch auswirken dürften.

ABBILDUNG 18 KORRELATION ZWISCHEN ALTER (METRISCH) UND ALTER, IN DEM DER HÖCHSTE SCHULABSCHLUSS ERWORBEN WURDE (METRISCH)

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Intervall- bzgl. Intervallmaß	Pearson-R	,230	,022	10,689	,000[c]
Ordinal- bzgl. Ordinalmaß	Korrelation nach Spearman	-,020	,025	-,895	,371[c]
Anzahl der gültigen Fälle		2047			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf normaler Näherung

Da es sich beim Vergleich um metrische Daten handelt, muß der PEARSONsche Korrelationskoeffizient r = .230 zur Interpretation herangezogen werden. Allerdings sind sehr viele der Zellen mit 0 besetzt. Die Interpretation wird dadurch erschwert. Dies spricht ebenfalls für ein ordinales Zusammenhangsmaß. Auch wurden die Befragten, die im Augenblick noch studieren, auf 0 gesetzt, da sie noch keinen höchsten Abschluß erreicht haben. Somit kommt es zu der wenig hilfreichen Annahme, dass die Höhe des Schulabschlusses mit dem Alter steigt. Unterstellt man in den Sozialwissenschaften bedeutsame Korrelationen > .20, würde man hier ein Artefakt produzieren.

In der Regel darf der PEARSONsche Korrelationskoeffizient nicht sehr weit von den Ergebnissen von SPEARMANs Rangkorrelation liegen - das ist hier jedoch der Fall. Eine weitere, negative Auswirkung entsteht durch die Bildungsvariable, die den Austritt aus dem Bildungssystem als Lebensalter wiedergibt und die im Gegensatz zum tatsächlichen Lebensalter eine andere Dimension aufweist.

Aus diesem Grund wurde die metrische Variable in eine ordinale Variable recodiert - mit zugegebenermaßen einigen Ungenauigkeiten. Die Gruppe der Studierenden wurde als eigene Kategorie zwischen den Absolventen der (fachgebundenen) Hochschulreife und den Hochschulabsolventen recodiert. Die Korrelation verändert sich drastisch:

ABBILDUNG 19 KORRELATION ZWISCHEN ALTER (METRISCH) RECODIERTER SCHULBILDUNG (ORDINAL)

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-b	-,217	,017	-13,150	,000
	Kendall-Tau-c	-,235	,018	-13,150	,000
	Korrelation nach Spearman	-,287	,022	-13,531	,000[c]
Intervall- bzgl.	Pearson-R	-,243	,021	-11,306	,000[c]
Anzahl der gültigen Fälle		2047			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Basierend auf normaler Näherung

Neben den bereits herausgearbeiteten Maßen SPEARMAN und PEARSON-r kommen nun zwei Kennzahlen von KENDALL dazu.

Auf den ersten Blick ist deutlich, dass alle vier Kennwerte negativ sind - je älter also ein Befragter ist, desto geringer ist die Schulbildung. Die Maße sind mit -.217 bzw. -.235 deutlich niedriger als der Koeffizient von SPEARMAN. Dies ist darauf zurückzuführen, dass SPEARMANs und PEARSONs Koeffizienten Monotonie unterstellen. Während PEARSONs r eine Normalverteilung voraussetzt, geht SPEARMANs Koeffizient davon aus, dass die Abstände zwischen den Kategorien gleich sind (z. B. dass der Abstand zwischen Volks- und Hauptschule einerseits und

Mittlerer Reife/erw. HS-Abschluß ebenso groß ist wie zwischen Studierenden und Akademikern).

Die Maße von KENDALL machen genau diese Annahme der gleichen Abstände nicht - ebenso wie Voraussetzungen über die Normalverteilung. Sie sind in der Regel somit immer niedriger als der SPEARMANsche Koeffizient.

Die Korrelation nach SPEARMAN (für ordinale Daten) ergibt -.287[31]. Der Korrelationswert kann zwischen - 1 (perfekter negativer Zusammenhang) und + 1 (perfekter positiver Zusammenhang) liegen. Ein negativer Wert in diesem Fall deutet an, dass mit zunehmenden Alter die Bildungsabschlüsse abnehmen - was durch die Bildungsexpansion plausibel ist. Somit ist Alter und Schulbildung miteinander korreliert, was bei vielen multivariaten Betrachtungen ein Problem sein kann. Allerdings würden sich - ohne jegliche Korrelationen - keine Aussagen über eine Stichprobe treffen lassen. Beispielsweise ergibt sich:[32]

- PC-Nutzung und Schulbildung v = .411, u = .062
- PC-Nutzung und Alter e = .542 (e^2 = .294), u = .248
- Alter und Schulbildung v = .429, u = .179

Wenn es z. B. keinen Unterschied hinsichtlich des Alters und Bildungsgrads (bezogen auf PC-Nutzung) gibt, können auch keine sinnvollen Gruppen gefunden werden.

Die Chi-Quadrat basierten Zusammenhangsmaße liegen alle deutlich über 0.2, liefern also einen hohen Beitrag zur Erklärung von PC-Nutzung.

31. PEARSONs r und SPEARMAN unterscheiden sich kaum - so dass der Einsatz der ordinalen Variable realitätsgerechter ist.

32. Zukünftig wird der Wert Cramers v mit v, Eta^2 mit e^2 und der Unsicherheitskoeffizient mit u abgekürzt. Phi wird - um Verwechslungen mit der Wahrscheinlichkeit p auszuschließen - ausgeschrieben.

Eta ist - wie der Unsicherheitskoeffizient - ein asymmetrisches (PRE-)Maß, das versucht, die Abweichung einer Variablen (hier: PC-Nutzung) vom Mittelwert unter Hinzunahme einer weiteren Variablen (hier: Alter) zu spezifizieren. Voraussetzung ist eine metrische, abhängige Variable und eine weitere Variable, deren Skalenniveau nominal, ordinal oder metrisch sein kann. Obwohl SPSS bei Kreuztabellen nur Eta ausgibt, ist immer die Quadrierung erforderlich - genauso wie beim multiplen Korrelationskoeffizient, der u. a. bei der linearen Regression zum Einsatz kommt.

ABBILDUNG 20 Lage- und Streuungsparameter für Alter (N = 2047)

Statistiken

D 11 Alter

N	Gültig	2047
	Fehlend	0
Mittelwert		46,41
Standardabweichung		17,862
Minimum		15
Maximum		94

Über die gesamte Stichprobe liegt das Durchschnittsalter (Mittelwert) bei rund 46 Jahren - die jüngsten Befragten sind 15, die ältesten 94 Jahre alt. Diese Lagemaßzahl gibt jedoch keinen Aufschluss darüber, wie sie entsteht: es wäre bei zwei Beobachtungen möglich, dass ein Befragter 45, der zweite 47 Jahre alt ist, es ist aber genauso denkbar, dass ein Befragter 26 Jahre und einer 66 Jahre alt ist. Von dieser Information ist abhängig, wie gut der Mittelwert eine Verteilung beschreibt. In ersterem Beispiel wäre dies sehr gut gelungen, im zweiten Fall nicht.

Die Standardabweichung, die immer zusammen mit dem Mittelwert interpretiert werden muss, definiert einen (in diesem Fall: (Alters)Be-

reich, in denen 2/3 der Befragten liegen (hier = 17.862). Dies bedeutet, dass 2/3 aller Befragten zwischen (rund) 46 +/- 18 Jahren, also zwischen 28 und 64 Jahre alt sind. 1/3 aller Befragten sind unter 28 bzw. über 64 Jahre alt.

Eta^2 greift auf den Mittelwert und die Abweichungen zurück und vergleicht jeden Wert mit dem Mittelwert. Damit sich kleinere und größere Differenzen vom Mittelwert (in diesem Fall: jüngere und ältere Befragte) nicht gegenseitig „aufheben", werden die Werte quadriert:

Genau wie bei anderen PRE-Maßen wird nun eine zweite Variable (die PC-Nutzung) herangezogen, um zu prüfen, ob sich unter Hinzunahme von weiteren Informationen die Vorhersagekraft verbessern läßt. Im Vergleich dieser beiden Variablen ergibt sich:

ABBILDUNG 21 Eta für PC-Nutzung und Geschlecht (N = 2038)

Richtungsmaße

			Wert
Nominal- bzgl. Intervallmaß	Eta	D 11 Alter abhängig	,497
		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) abhängig	,542

Eta wird in SPSS als Richtungsmaß ausgegeben: der Nutzer muss entscheiden, ob Alter oder die Häufigkeit der PC-Nutzung die abhängige Variable ist. Im Gegensatz zum PEARSONschen r kann Eta^2 höher sein, da dieses Maß nichtlineare Zusammenhänge erfassen kann.

Der Wert 0.542 (Eta^2 = 0.294) wird in diesem Fall verwendet, da die PC-Nutzung vom Alter abhängt. Der Eta-Wert kann - wie die meisten

Nominalmaße - zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang) liegen. Ein Wert von 0.294 zeigt hier eine deutliche Abhängigkeit der PC-Nutzung vom Alter.

Eta macht - auch durch das metrische Skalenniveau einer Variablen - mehr Annahmen als der Unsicherheitskoeffizient und kann dadurch höher sein. So geht Eta davon aus, dass die Standardabweichungen für alle Merkmalsausprägungen gleich ist, was eine Normalverteilung unterstellt.

ABBILDUNG 22 Vergleich Eta und Unsicherheitskoeffizient für Alter und PC-Nutzung (N = 2038)

Richtungsmaße

			Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Unsicherheitskoeffizient	Symmetrisch	,070	,004	16,716	,000[c]
		D 11 Alter abhängig	,041	,002	16,716	,000[c]
		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) abhängig	,248	,015	16,716	,000[c]
Nominal- bzgl. Intervallmaß	Eta	D 11 Alter abhängig	,497			
		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) abhängig	,542			

a. Die Null-Hyphothese wird nicht angenommen.

b. Unter Annahme der Null-Hyphothese wird der asymptotische Standardfehler verwendet.

c. Chi-Quadrat-Wahrscheinlichkeit für Likelihood-Quotienten.

Der Wert des Unsicherheitskoeffizienten (rund 0.25) ist häufig niedriger als von Eta^2 (ca. 0.29). Dies liegt am unterschiedlichen Skalenniveau und den sich daraus ergebenden Verteilungen für Intervallskalen (z. B. Mittelwerte, Standardabweichungen), soll aber in dieser Arbeit nicht näher untersucht werden.

Für nominale multivariate Analysen greifen bestimmte Entscheidungsbaumalgorithmen die Idee des Chi-Quadrat-Werts auf. Die Vorgehensweise ist - vereinfacht ausgedrückt - folgende: es wird nach derjenigen unabhängigen Variablen gesucht, die z. B. den

größten Chi-Quadrat-Wert (also in irgendeiner Form einen Zusammenhang mit der Zielvariablen) aufweist. Der Datensatz wird in zwei oder mehr Gruppen aufgeteilt. Im nächsten Schritt wird untersucht, welche unabhängige Variable nun den höchsten Chi-Quadrat-Wert für jede Subpopulation liefert, usw.

Unterstellt, die wichtigste unabhängige Variable, bezogen auf die PC-Nutzung ist das Alter: Jüngere nutzen eher den PC, Ältere weniger. Angenommen, es gäbe keinen Unterschied zwischen den Bildungsabschlüssen der Jüngeren und den Älteren, so gäbe es auch keine sinnvolle Trennung zwischen den eher jüngeren und älteren Befragten hinsichtlich der Bildung - die Lösung wäre damit nicht optimal.

Dies würde bedeuten, dass in einem ersten Schritt eine Trennung aufgrund des Alters erfolgt - der Bildungsabschluss aber keine weitere Erklärungskraft in diesem Modell besitzt. Durch die Bildungsexpansion ist jedoch gerade dies gegeben: so ergeben sich neben Jüngeren und Älteren vier Gruppen: besser bzw. schlechter ausgebildete jüngere Befragte und die älteren Pendants.

Das heißt: je weniger die Variablen untereinander korreliert sind, desto weniger gut können multivariate Verfahren Gruppen segmentieren. Das ist bei Entscheidungsbäumen etwas anders, da in jedem Schritt der Analyse (Trennungen) alle Variablen auf ihren Beitrag (z. B. Chi-Quadrat) geprüft werden. Korrelationen unter den unabhängigen Variablen spielen keine so große Rolle - auch wenn sie als gewisse Störfaktoren gesehen werden können. Ohne Korrelationen gibt es aber auch keine sinnvollen Gruppen. Dies ist ein weiterer Vorteil der Entscheidungsbäume: Variablen, die ein deutlich höheres Zusammenhangsmaß liefern als andere unabhängige Variablen, können auf den ersten Blick identifiziert werden, da für jede Stufe des Ent-

scheidungsbaumprozesses immer die wichtigste (die Variable mit dem jeweils höchsten Zusammenhangswert) ausgewählt wird.[33] Liegt keine Korrelation vor, wird auch kein Entscheidungsbaum generiert.

Sowohl Entscheidungsbäume als auch die Diskriminanzanalyse und die Regressionsanalyse geben sog. „Fehlklassifikationsschemata" aus, d. h. eine tabellarische Übersicht, wieviele Personen falsch bzw. richtig zugeordnet wurden. Für das Beispiel der PC-Nutzung bedeutet das: wieviele PC-Nutzer bzw. Non-User sind tatsächlich auch durch das jeweilige Verfahren erkannt worden, welche nicht? - Die Anzahl der falsch klassifizierten Personen wird prozentual zur Gesamtzahl der untersuchten Stichprobe ausgegeben. Wurden also von 1000 Befragten 50 Nichtnutzer irrtümlich als Nutzer, 50 Nutzer irrtümlich als Nichtnutzer klassifiziert, sind (50 + 50) / 1000 = 0.10 bzw. 10 % durch das Verfahren fehlklassifiziert. Durch den Vergleich kann überprüft werden, ob ein Verfahren bessere Gruppen findet als ein anderes. Da Entscheidungsbäume in der Soziologie kaum angewandt werden, soll diese Arbeit dazu beitragen, diese Methode entweder neben „bewährten" Verfahren wie der Regression zu etablieren - oder es als für die Soziologie nicht geeignetes Verfahren verwerfen.

1.7 Ableitungen für diese Arbeit

Für diese Arbeit wurde - um Aussagen auch verallgemeinern zu können - ein quantitatives Vorgehen gewählt. Die Rahmenbedingungen der Fragestellungen - PC-Nutzung, dominante Schichtungen - sind, wie der Theorieteil gezeigt hat, gut erforscht, so dass eine grundlegende qualitative Arbeit aufgrund der geringen Fallzahl für diese Arbeit eher ein Nachteil wäre.[34]

33. Hierbei muss angemerkt werden, dass natürlich das Gesamtergebnis des Entscheidungsbaums ebenfalls eine Rolle spielt. Somit ist das Ergebnis ebenfalls nicht völlig verzerrungsfrei.

Ziel ist es, PC-Nutzung aufgrund vielfältiger Variablen im Sinne der dominanten Schichtungen nach GEIGER zu untersuchen. Zum Einsatz kommen Entscheidungsbaum-Verfahren, als Kontrollmethoden werden aufgrund der Vergleichbarkeit die logistische Regression und die Diskriminanzanalyse herangezogen. Um auch dem statistisch weniger versierten Leser die komplexen Verfahren zu verdeutlichen, werden die Ergebnisse aller multivariater Verfahren an einem einfachen, eher alltäglichen Beispiel eingeführt, dessen Ziel es ist, dem Leser die Methoden näherzubringen.[35] Als abhängige Variable wurde die dichotomisierte PC-Nutzung aus Frage 39 des EUROBAROMETER-Datensatzes (PC-Nutzung: ja - nein), als unabhängige Variablen das offen erfragte Lebensalter (Frage D 11) und die Schulbildung (Frage D 8, recodiert) herangezogen.[36]

Deskriptive statistische Ergebnisse in Form von (überwiegend) nominalen Zusammenhangsmaßen (Cramers v, Unsicherheitskoeffizient) liefern die Grundlage für dominante Schichtungen. Unterstellt wird, dass alte Ungleichheiten, erweitert durch die Variablen Geschlecht und Alter die Basis der Schichtstruktur wiederspiegelt. Subordinierte Schichtungen - Kultur- und Freizeitaktivitäten mit hohen Zusammenhang mit der PC-Nutzung - sind daraus abgeleitet.

Dies bedeutet, dass multivariat ein Modell mit der abhängigen Variable PC-Nutzung und den unabhängigen dominanten sozialstrukturellen Variablen, die einen hohen Zusammenhangswert mit der PC-

34. vgl. für die PC-Nutzung die Literatur von RAMMERT (1990), RAMMERT et al. (1991) und BÜHL (1999), für die neuere Sozialstruktur- bzw. Schichtungsanalyse vgl. u. a. FLAIG et al. (1997), GEISSLER (1996).

35. Zumindest für zwei Variablen sind ganz einfache grafische Verfahren, wie sie auch in dieser Arbeit vorgestellt werden, ohne großen Aufwand möglich.

36. Die Schulbildung wurde - wohl um der europäischen Vergleichbarkeit der vielfältige Bildungssysteme Rechnung zu tragen - als Alter, in dem der höchste Bildungsabschluss erreicht wurde, erfasst. Die methodische Zusammenfassung wird im Ergebniskapitel detailliert erläutert und soll an dieser Stelle nicht verwirren.

Nutzung aufweisen, gebildet werden. Die gefundenen Gruppen lassen sich gegebenenfalls durch die subordinierten Schichtungen weiter beschreiben. Grafische Verfahren sollen die gefunden Zusammenhänge zusätzlich anschaulich erläutern. Es ist zu erwarten, dass das Ergebnis alltäglich sein wird, dies ermöglicht es aber dem Leser, die Methoden besser zu verstehen, da er sich bei wenigen, verständlichen, überschaubaren Variablen eher auf das methodische Vorgehen konzentrieren kann. Trotzdem sind die Ergebnisse - vor allem auch in ihrer Unterschiedlichkeit, wie sie von den einzelnen Algorithmen ermittelt werden, sehr aufschlussreich.

Die Anzahl der Variablen ist durch den Datensatz und die inhaltliche Fragestellung begrenzt. Von den multivariaten Verfahren eignen sich aufgrund ihrer Vielfältigkeit im Umgang mit unterschiedlichen Skalenniveaus insbesondere Entscheidungsbäume, die einen Schwerpunkt dieser Arbeit ausmachen. Anhand nominaler und ordinaler Variablen werden sog. Klassifikationsbäume gebildet. Regressionsbäume entstehen durch den Einsatz metrischer Daten. Die Fragestellung dieser Arbeit nach PC-Nutzung (ja - nein) ist somit ein Beispiel für ein typisches Klassifikationsproblem.

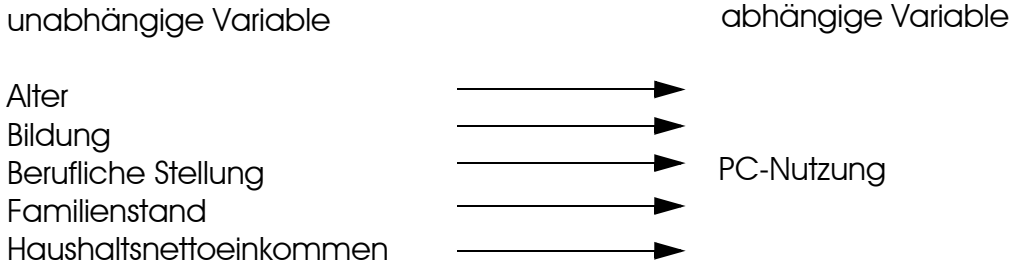
Auch weniger an mathematischen Formeln Interessierten soll der Zugang zu diesen Verfahren mit dieser Arbeit eröffnet werden. Die Erläuterungen zu den Algorithmen sind deshalb weniger mathematisch als verbal geprägt: das Verständnis für die Möglichkeiten und Grenzen der Methoden und Kennzahlen steht hierbei im Vordergrund. Durch die sprachliche Umsetzung werden die mathematischen Formeln an dieser Stelle nicht berücksichtigt, um dem Leser einen anderen Zugang zur Thematik zu eröffnen, die mehr darauf abzielt, die inhaltlichen Zusammenhänge zu verstehen. Da die Literatur zu diesem Verfahren äußerst umfangreich ist (alleine das Handbuch zu Answertree hat rund 350 Seiten, daneben gibt es unzählige

Veröffentlichungen mit unterschiedlichsten Schwerpunkten) kann es vorkommen, dass die Algorithmen nicht bis ins letzte Detail erläutert werden - was auch eine Arbeit innerhalb der angewandten Wissenschaften vor Probleme stellen würde. Für weitergehende, vor allem mathematische Erläuterungen, sei auf die von SPSS herausgegebenen Handbücher (SPSS (2001a, 2001b)) und die Veröffentlichungen von BALTES-GÖTZ (2001, 2004a) hingewiesen.[37] Um den Umfang der Arbeit nicht allzusehr zu strapazieren, werden die anderen Verfahren (logistische Regression, Diskriminanzanalyse), die seit langer Zeit in der Soziologie verwendet werden, in aller Kürze behandelt und auf die umfangreiche Literatur (z. B. BÜHL und ZÖFEL (2002a: 359ff., 431ff., 487ff.), URBAN (1993), BACKHAUS (2004: 417ff.)) verwiesen.

2 Einführung in die kausalen multivariaten Verfahren

Häufig ist es interessant, neben der reinen Häufigkeitsverteilung bzw. der bivariaten Analyse auch Zusammenhänge zwischen verschiedenen Variablen herzustellen.

ABBILDUNG 23 Abhängige und unabhängige Variable(n) bei multivariaten Fragestellungen



37. Im Internet finden sich unzählige Anwendungsbeispiele aus verschiedensten Disziplinen. Aufgrund der vielfältigen Forschungen auf diesem Gebiet und des schnellen Wandels des Internets wird hier auf Suchmaschinen verwiesen.

War die Frage bei den bivariaten Verfahren: „Wie beeinflusst das Alter (unabhängig) die PC-Nutzung (abhängig)?“ werden hier mehrere Variablen zur Untersuchung herangezogen.

Anhand eines einfachen Modells werden Entscheidungsbäume, die logistische Regression und die Diskriminanzanalyse vorgestellt und erste Vergleiche gezogen. Eingeleitet wird dieses Unterkapitel durch ein rein grafisches Verfahren, die Parallelplots. Ziel ist es, auch denen, die nicht mit multivariaten Methoden vertraut sind, anhand eines einfachen Regressionsmodells (eine abhängige, zwei unabhängige Variablen) diese zu erläutern. Erst im nächsten Hauptkapitel werden die Verfahren dann - allerdings ohne nochmals ausführlich auf die Definition der Kennzahlen einzugehen - zu einem vollständigen Modell ausgebaut.

2.1 Grafische Verfahren

Stellen Sie sich vor, sie sind in einer fremden Stadt und müssen vom Hauptbahnhof zu ihrem Ziel mit öffentlichen Verkehrsmitteln gelangen. Sie haben die Möglichkeit, jemanden zu fragen oder sich einen Stadtplan mit dem öffentlichen Verkehrsliniennetz zu besorgen.

In einer uns unbekannten Stadt ist es sicherlich einfacher, sich mit einem grafischen Plan der öffentlichen Verkehrsmittel zurechtzufinden als mit einer verbalen Erläuterung. Auf dem Plan ist ersichtlich (meist noch anhand unterschiedlicher Farben für die jeweiligen S-, U-Bahnen, Busse und Straßenbahnen) welche Linien uns zum Ziel bringen. Eine Erklärung wie: „Sie fahren jetzt zwei Stationen mit der Linie 2 Richtung x, dann steigen Sie in die Linie 1 Richtung y, fahren fünf Stationen und warten dann auf den Bus Nr. 38 in Richtung z“ ist sicherlich verwirrender und schwerer zu merken als grafisch aufbereitete Informationen - ohne Informationsverlust. Visualisierung hat auch in diesem Sinne nichts mit einer Negierung der Sprache zu tun - auch

visualisierte Informationen müssen, vielleicht noch viel aufwendiger als Kennzahlen, sprachlich umgesetzt werden.

Dies ist schon ein komplexes Beispiel: tagtäglich nehmen wir ganz nebenbei grafische, feststehende Informationen wahr und reagieren darauf (Verkehrszeichen, Hinweisschilder) und Symbole (z. B. @, www, +, ...). Wir würden es wohl als völlig unsinnig ansehen, an einer Kreuzung auf das Schild mit dem geschriebenen Text „Vorfahrtsstraße", „Überholverbot", o. ä. zu stoßen. Die einzige Ausnahme ist eine (Kilo-)Meterangabe (z. B. Rastplatz 2 Kilometer) bzw. Erläuterungen zu den Schildern.

Der Vorteil von Visualisierungen - z. B. bei Verkehrsschildern - liegt in der Kodierung von Informationen in einem Symbol. Bei einer Unterführung, die z. B. nur 2.50 m hoch ist und die nur jeweils von einer Seite befahren werden kann, reichen zwei Symbole aus, um diesen Sachverhalt zu beschreiben. Verbal müßte ein Schild etwa so lauten: „In 200 Metern gelangen Sie zu einer Unterführung. Die Unterführung ist nur 2.50 m hoch und nur jeweils in eine Richtung befahrbar. Diese Seite hat Vorfahrt." Wenn Sie an einem Schild mit diesen verbalen Informationen mit einer Geschwindigkeit von 60 km vorbeifahren, haben Sie keine Chance, das Schild zu lesen - vor allem, wenn Sie sich auf die Straße konzentrieren müssen.

Ein frühes Beispiel für Visualisierung ist John SNOWs Karte der Cholera-Epidemie in London von 1854 (vgl. SPENCE (2001: 7), TUFTE (1997: 27ff.)):

> „From the General Register Office, Snow obtained a list of 83 deaths from cholera. When plotted on a map, these data showed a close link between cholera and the Broad Street pump. Persistent house-by-house, case-by-case detective work had yielded quite evidence about a possible cause-effect relationship ..." (TUFTE 1997: 28))

Dies ist ein Beispiel für deduktives Vorgehen: ausgehend von der Hypothese, dass sich Cholera nicht über die Luft (wie ursprünglich angenommen), sondern über das Wasser verbreitet, untersuchte Snow die Todesfälle anhand eines grafischen Plots, der bei 83 Fällen das Problem schneller verdeutlicht als eine Kreuztabelle:

> „Cholera broke out in the Broad Street area of central London on the evening of August 31, 1854. John Snow, who had investigated earlier epidemics, suspected that the water from a community pump-well at Broad and Cambridge Streets was contaminated." (TUFTE (1997: 28))

Grafische Darstellungen sind somit teilweise statistischen Verfahren ebenbürtig, von der Verständlichkeit auch überlegen. In der deskriptiven Statistik haben sie allerdings häufig nur die Rolle eines illustrierenden Charakters: „hard facts" (errechnete statistische Kennzahlen) sollen illustriert werden.

Ein weiteres Beispiel für Visualisierung ist die Karte der Londoner Tube, die Harry BECK 1932 entwickelte (vgl. SPENCE (7ff.)). Es zeigt das Liniennetz der Tube und damit London „von unten". Dieses Verfahren hat heute jedes Öffentliche Verkehrsunternehmen übernommen, um die Strecken der U-Bahnen, Straßenbahnen, S-Bahnen und Busse darzustellen.

Einer der Pioniere der neueren Visualisierung ist Edward TUFTE (vgl. TUFTE (1983, 1990, 1997)):

> „Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color." (TUFTE 1983: 9))

Hierbei folgt für TUFTE, dass nur aufgrund einer guten Visualisierung auch gute Forschungsergebnisse erzielt werden können („Clear and precise seeing becomes as one with clear and precise thinking." (TUFTE (1997: 53)).

Das Problem der Visualisierung liegt nicht in der Möglichkeit, sondern in der Komplexität der Darstellung. Durch große Datenmengen mit vielen Variablen und Datensätzen steigen die Anforderungen an die Mehrdimensionalität. Wie lassen sich multidimensionale Daten in einem vorgegebenen Raum (z. B. A4, Monitor) realisieren? - Auch hier gibt es einige Möglichkeiten, z. B. die Navigation durch Scrollen bei Websites, die Aufteilung von Informationen auf unterschiedliche Seiten und die Verbindung mit Links, etc. Neuere Ansätze finden sich u. a. bei SPENCE (2001: 111ff.).

Statistik- und Tabellenkalkulationsprogramme bieten umfangreiche Grafikmöglichkeiten (Diagramme) zur Reduktion (z. B. Mittelwert-Balkendiagramm), Zusammenfassung der Dateninformationen (z. B. prozentuales Balkendiagramm) an. Hier können jedoch sinnvoll nur drei Dimensionen abgebildet werden (z. B. Geschlecht, Einkommen und Bildungsabschluss). Für eine weitere Variable (z. B. Altersgruppen) muss eine zusätzliche Grafik erstellt werden (z. B. Männer bzw. Frauen hinsichtlich Alter und Geschlecht) - oder eine gestapelte Grafik (z. B. gestapelte Balken) gewählt werden.

Die Techniken der Informationsvisualisierung sind vielfältig und eine Beschreibung würde den Umfang der Arbeit strapazieren. Weiterführende Informationen finden sich bei WEGMAN (2003b, 2003c), KEIM (1997), oder THEARLING (2001).

Die grafische Darstellung von Daten ist eine der herausragendsten Stärken von Data Mining. Ähnlich wie bei der Korrespondenzanalyse steht die Veranschaulichung der Daten im Vordergrund.

> „Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations." (TUFTE (1983: 13))

Die Visualisierungsmöglichkeiten, die vor allem in der Informatik erforscht werden, ermöglichen einer Reihe von wissenschaftlichen Dis-

ziplinen die komplexe Informationsvisualisierung - zum Beispiel bei der Darstellung geografischer Daten (Landschaften, Kartographie, zukünftige Bauvorhaben, etc.), die mit Zahlen nicht dargestellt werden könnten. TUFTE (1983: 16ff.) belegt dies mit sechs Karten der USA, die - nach Counties geordnet - unterschiedlichfarbige Informationen anhand von fünf Signifikanzlevels zu den an Krebs gestorbenen Personen (für weisse Männer und Frauen, allgemein und spezifiziert nach Krebsarten 1950 - 1969) liefern. Er bemerkt: „Only a picture can carry such a volume of data in such a small space.“[38].

Ziel der Visualisierung ist es, neue Zusammenhänge zu entdecken, diese Informationen zusammenzufassen und präsentieren. Nachfolgend werden einige Visualisierungsverfahren, die bis jetzt kaum in der Soziologie angewandt worden sind, vorgestellt - obwohl sie überwiegend von SCHNELL (1999) beschrieben wurden.

2.1.1 Parallele Koordinaten

Die Methode der Parallelen Koordinaten oder Parallelplots ist geeignet, hochdimensionale Daten grafisch darzustellen und somit einen Überblick über die Daten zu erhalten (vgl. WEGMAN (2003b: 7)). Er bemerkt:

> „Visualization for Data Mining can realistically hope to deal with somewhere on the order of 10^6 to 10^7 observations.“ (WEGMAN (2003b: 6))

Bei Parallelplots werden die ausgewählten Variablen als parallele Achsen dargestellt und die Werte innerhalb der Achsen abgetragen. Jeder Datensatz wird als horizontale Linie dargestellt, der die Achsen bei dem jeweiligen Wert für den Datensatz schneidet. Ziel ist es, einen Überblick über eine Vielzahl von Variablen zu gewinnen.

38. TUFTE geht von ca. 21000 Informationen pro Landkarte aus.

Das Verfahren wurde Ende der 1970er/Anfang der 1980er Jahre von Alfred INSELBERG (vgl. z. B. INSELBERG (2000), INSELBERG und DIMSDALE (1990), FUA (1999), WARD (1994)) entwickelt:

> „Our goal is the visualization of complex problems with many parameters - *Multivariate Visualization* or equivalently *Multidimensional Visualization* and we shall concentrate on *Information Visualization*." (INSELBERG (2000: 8))

> „Believe it or not, the fascination with Dimensionality may predate Aristotle and Ptolemy who argued that space had only three dimensions. By the nineteenth century, mathematicians like Riemann, Lobachevsky and Gauss unshackled our imagination and higher-dimensional and non-Euclidean geometries came into their own." (INSELBERG (2000: 8))

Bedauerlicherweise gibt es im Augenblick keine sehr anwendungsorientierten Programme zur Darstellung von Parallelplots, die kostengünstig sind. Es existieren einige Java-Applets, die aber keine SPSS (oder Excel-)Dateien einlesen. Abhilfe schaffen hier Excel-Implementierungen für Statistik, z. B. XLSTAT (http://www.xlstat.com). Leider unterliegt dieses Programm den Excel-Konventionen - so dass nur 256 Dimensionen pro Grafik dargestellt werden können, was in der Matrix durch Fälle * Variablen ausgedrückt wird. XLSTAT behilft sich bei zu großen Datensätzen mit einer Stichprobenziehung - was keine befriedigende Lösung darstellt.[39]

Allerdings gibt es zwei freie Lösungen, die auch Excel- (xdmv) bzw. ascii-Daten (aus SPSS) (Mondrian) einlesen können - xdmv teilweise nur über einige Umwege.[40], [41]

Alter, Bildung und PC-Nutzung wurden als Variablen ausgewählt, die die Achsen bestimmen. Die Achsen werden vertikal und parallel nebeneinander gestellt und die Skalierung den jeweiligen Werteberei-

39. Trotzdem ist XLSTAT für kleinere Datensätze nicht die schlechteste Wahl, da sowohl gerichtete als auch ungerichtete Plots erzeugt werden können. Die Erstellung der Grafik unterscheidet sich nicht von anderen EXCEL-Grafiken und kann ebenso einfach bearbeitet wie in andere Programme (z. B. Textverarbeitung) eingefügt werden.

chen dimensioniert. Der „untere“ Wert (1) repräsentiert hier die Nichtnutzer, der obere (Wert 2) die Nutzer. Der Wertebereich des Alters, (15 - 94 Jahre) und der Bildung (1 - 5) werden analog als weitere Achsen abgetragen. Die Achsenwerte sollen allerdings nicht verwirren: Je höher die Werte auf den Achsen, desto älter sind die Befragten, haben eine höhere Bildung und nutzen den PC (und umgekehrt).

Jeder Datensatz wird als „Zickzack“-Linie abgetragen, Ein Nichtnutzer, der 55 Jahre alt ist und einen Hauptschulabschluss als höchsten Bildungsgrad angegeben hat, wird auf der ersten Achse (PC-Nutzer) mit 1 (Nichtnutzer), auf der zweiten Achse mit seinem tatsächlichen Alter (55 Jahre) und auf der dritten Achse mit dem Wert 1 abgetragen. Die Linie beginnt also auf der ersten Achse unten, steigt auf 55 Jahre bei der zweiten Achse Alter und sinkt wieder auf der dritten Achse auf den Wert 1 (Hauptschule). Das Gesamtbild ergibt sich aus den 2038 Fällen des Datensatzes.

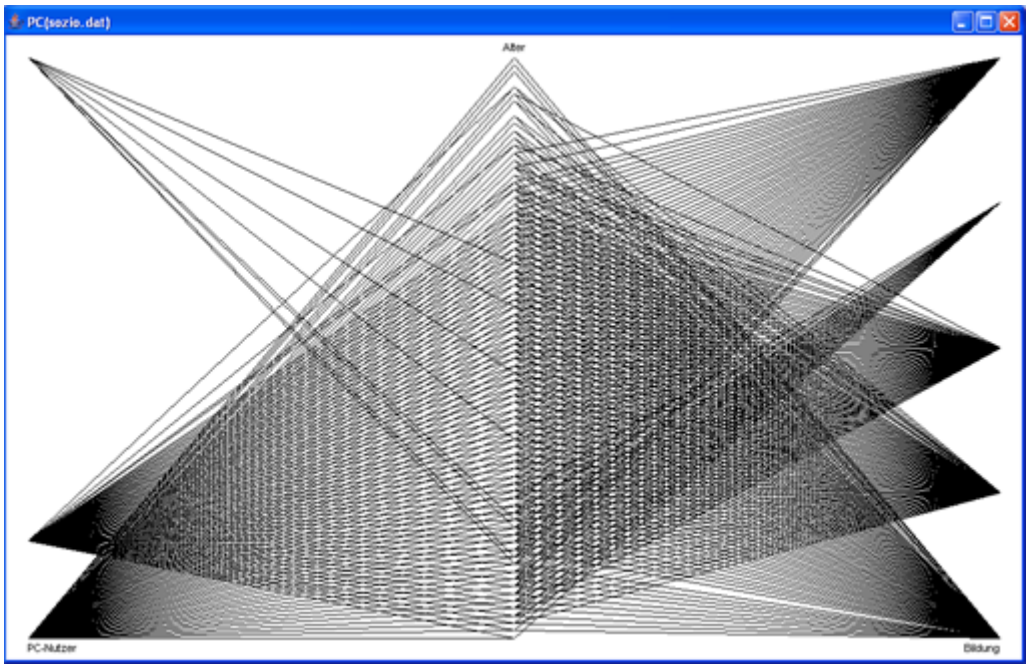
Je mehr Linien sich „überlappen“, desto größer ist der Zusammenhang zwischen zwei Variablen. Die deutlichsten Konzentrationen finden sich zwischen PC-Nutzern mit eher geringerem Alter und höherer Bildung - und umgekehrt.

Mit Mondrian sieht der Plot folgendermaßen aus:

40. Xdmv verarbeitet csv-konvertierte Excel-Dateien. Die SPSS-Daten können in SPSS als Excel-Datei gespeichert werden und anschließend mit Excel ins csv-Format konvertiert werden (ein Format, dass die Variablenwerte mit Kommata trennt). Leider verwendet die deutsche Excel-Version als Trennzeichen nicht das Komma, sondern das Semikolon, so dass mit einem beliebigen Editor die Semikolons durch Kommata ersetzt werden müssen. Die gewonnene Datei wird nochmals in Excel eingelesen, als csv-Datei abgespeichert und liegt nun in einem Format vor, in dem sie in das von Xmdv verwendete *.okc-Format umgewandelt werden kann. Hierfür gibt es ein im DOS-Fenster ausführbares Programm: `excel2xmdv`. Mit der Anweisung `excel2xmdv dateiname.csv` wird die Datei umgewandelt und kann in Xmdv geöffnet werden.

41. xdmv ist erhältlich unter http://davis.wpi.edu/ ~xdmv, mondrian unter http://stats.math.uni-augsburg.de/mondrian/mondrian.html.

ABBILDUNG 24 Parallele Koordinaten: PC-Nutzung, Alter und Bildung mit Mondrian



Aus der Grafik läßt sich ersehen, dass die Variable PC-User drei Kategorien hat: an der untersten (Nichtnutzer) lassen sich viele „Linien“ in Richtung eines hohen Alters erkennen (Alter, mittlere Kategorie), die dann auf der dritten (Bildungs-)Achse stark absinken (Volks- und Hauptschulabschluss).

Die zweite PC-User-Kategorie (Nutzer) haben eher einen Schwerpunkt im unteren Alterssegment, dafür ist der Bildungsgrad tendenziell höher.

Die dritte Kategorie der PC-User - diejenigen, die keine Angaben machten, sind in mittleren Altersjahrgängen vertreten, wobei die Jüngeren auch hier einen höheren Bildungsabschluss erreicht haben.

Mondrian verzichtet in dieser Abbildung auf Achsen, sondern trägt an der entsprechenden Stelle nur die Variablennamen ab. Dadurch wird das Ganze etwas übersichtlicher. Allerdings scheint es auch hier noch keine sinnvolle Exportfunktion für Grafiken zu geben. Zwar gibt

es neben der „Bildschirmfoto“-Funktion von Windows die konventionelle Methode, Grafiken über die Zwischenablage zu kopieren - das Ergebnis ist aber qualitätsmäßig nicht überragend. Auch laufen beide Programme nicht sehr stabil. Trotzdem scheint es sinnvoll, diese Möglichkeiten für zukünftige Arbeiten aufzuzeigen - nicht nur, weil Visualisierung heute und in der Zukunft ein zentrales Thema ist bzw. sein wird, sondern weil damit auch konkrete (soziologische) multivariate Ergebnisse transparenter gemacht werden können.

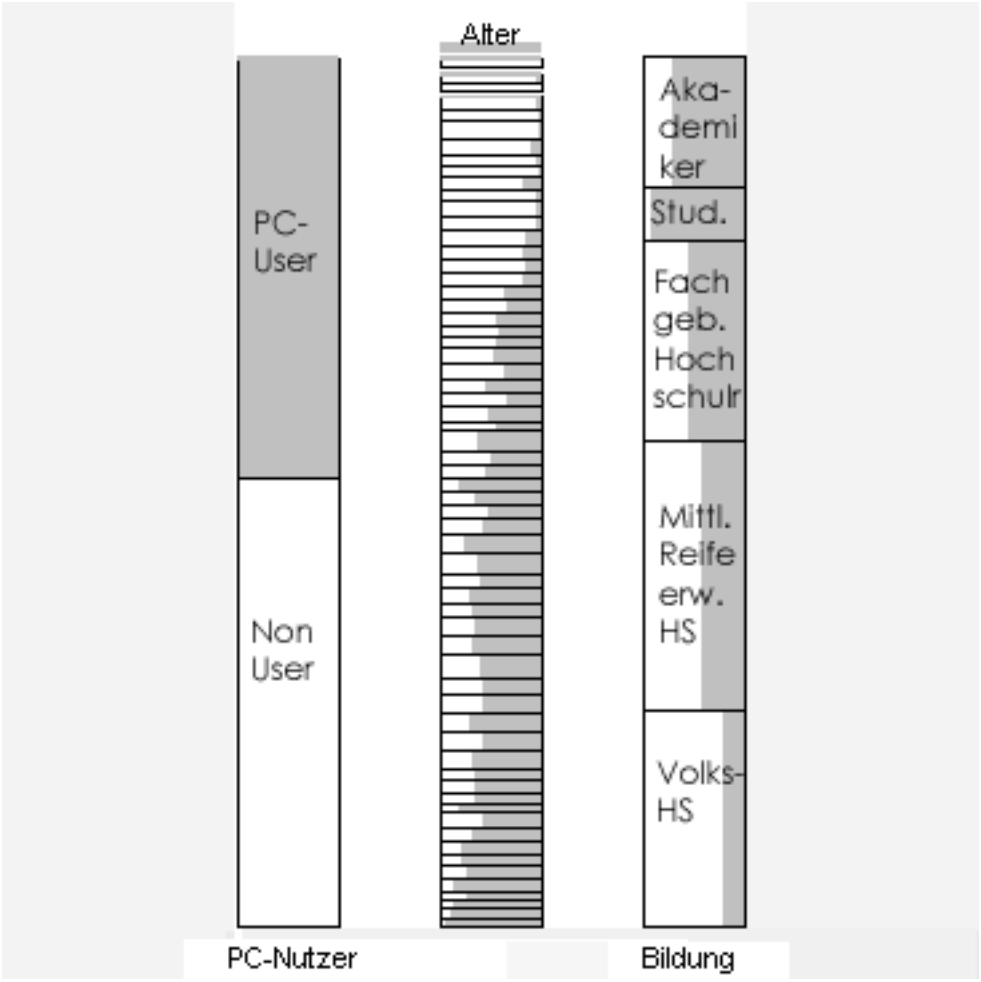
An dieser Stelle wird auf eine Darstellung von xmdv aus folgenden Gründen verzichtet:

- der Datenexport aus SPSS ist sehr kompliziert
- die Dimensionen der Grafik sind nicht identisch mit denen des Datensatzes (z. B. wird bei den PC-Nutzern, die mit 1 und 2 codiert werden ein Wertebereich von 0.95 bis 2.05 angegeben, was nur zu Verwirrungen führt)
- die Grafik, die nur über die Zwischenablage kopiert werden kann, läßt sich zwar auf dem Bildschirm gerade noch erkennen, der Output ist jedoch mangelhaft

2.1.2 Spine Plots

Neben den Parallelen Koordinaten bietet Mondrian parallele (gestapelte) Balkendiagramme, sogenannte Spine Plots an, was einen wesentlich besseren Überblick über die Datenstruktur verschafft.

ABBILDUNG 25 Spine Plots mit Mondrian (Alter, Bildung, PC-Nutzung



Die Grafik erinnert an gestapelte Balkendiagramme - und nichts anderes sind Spine Plots - allerdings bieten sie einige zusätzliche Informationen.

Alle Balken sind gleich hoch und repräsentieren 100 % einer Variablen (z. B. pcuser). Etwa die Hälfte (weiß) sind Nichtnutzer (codiert mit 1), die andere Hälfte (graue Fläche) Nutzer (codiert mit 2).

Diese codierte Information (weiss bzw. grau) wird nun mit den jeweiligen Anteilen auf die „Nachbarbalken" Bildung und Alter übertragen. Somit läßt sich an jedem Balken der jeweilige Anteil der PC-Nutzer

grafisch ablesen.[42] Zusätzlich verdeutlicht die unterschiedliche Breite der Balken die Anteile.

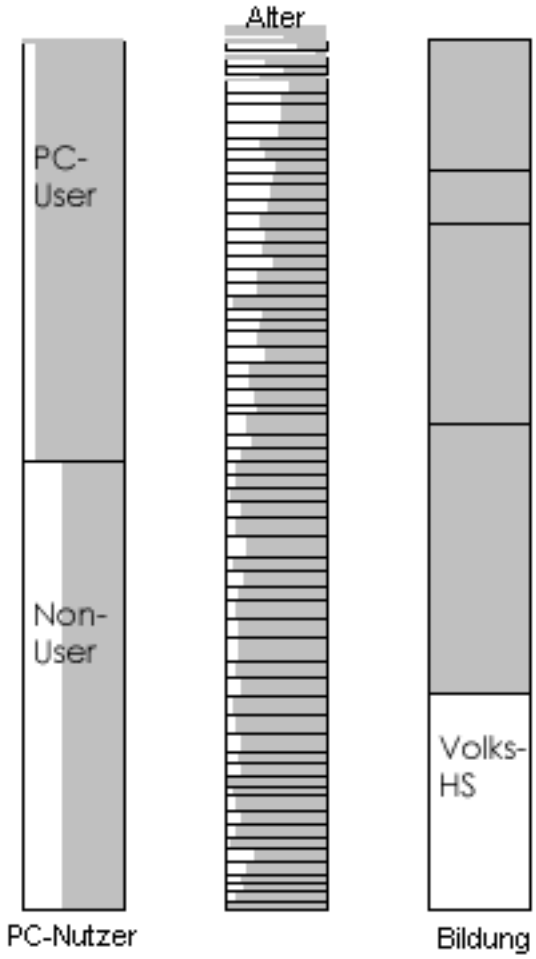
Die grau schraffierten Flächen charakterisieren die Nutzer, die weißen die Nichtnutzer - definiert durch den ersten Balken („pcuser“).[43] Deutlich wird dies am Beispiel der Bildung (oberstes Segment: Hochschulabsolventen, zweitoberstes Segment: Studierende, drittes Segment: (Fach)Hochschulreife-Abgänger, vorletztes Segment: Mittlere Reife, unterstes Segment: Hauptschulabgänger). Die Nichtnutzer, (weiß) haben in den unteren Segmenten höhere Anteile als in den oberen. Dasselbe gilt auch für die oberen Altersjahrgänge.

Somit lassen sich auf einen Blick Nutzer- bzw. Nichtnutzeranteile erkennen. Die Nichtnutzer sind eher älter und haben niedrigere Bildungsabschlüsse. Natürlich ist es auch möglich, jedes andere Segment der Grafik anzuklikken und sich z. B. anzusehen, wie sich die Hochschulabsolventen durch die Variablen PC-Nutzung und Alter definieren. Durch die Interaktivität und die hohe Übersichtlichkeit eignen sich diese Boxplots besser als Parallelplots zur Darstellung von Daten.

42. Ein wünschenswertes Feature für zukünftige Versionen wäre die Einblendung der Werte (Anteile) jeder Variablen. Nur mit einem Klick mit der rechten Maustaste erhält man Kontextinformationen über die Anzahl und die Zusammensetzung des jeweiligen Balkenabschnitts. Ein weiteres Problem (was in diesem Fall nicht ins Gewicht fällt, aber später noch diskutiert wird) ist das Problem, dass Mondrian noch nicht in der Lage ist, fehlende Werte aus einer Grafik auszuschliessen. Werden z. B. fehlende Werte in SPSS als sysmis gesetzt, erscheinen sie also als Komma, weigert sich Mondrian, die Grafik darzustellen. Es ist nur möglich, den fehlenden Werten einen Zahlenwert (z. B. 99) zuzuordnen und in die Grafik aufzunehmen - was zu Verwirrung führen kann.

43. Die Farben lassen sich individuell über Options/Preferences in Mondrian anpassen.

ABBILDUNG 26 Spine Plots mit Mondrian (Alter, Bildung, PC-Nutzung): Verteilung der Hauptschulabsolventen auf PC-Nutzung und Alter



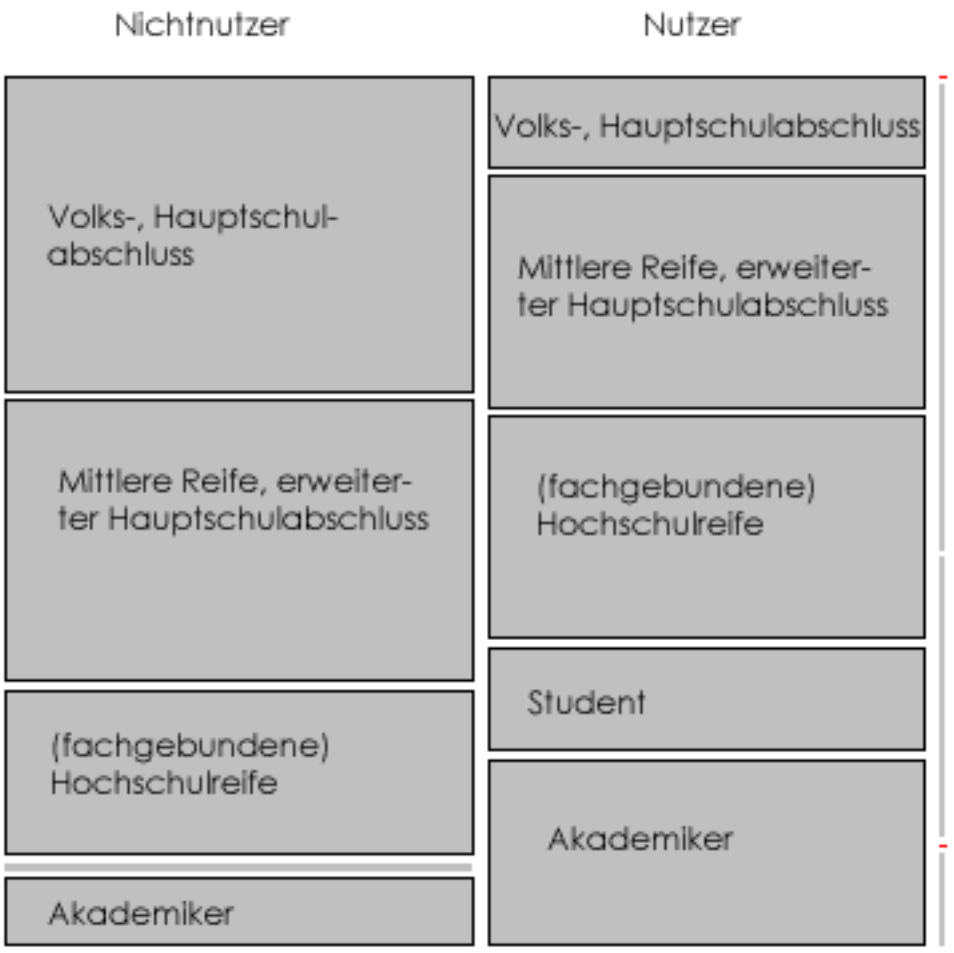
Durch einfaches Anklicken des rechten Balkens Bildung (Hauptschule) werden die Verteilungen hinsichtlich der andern Variablen (als weiße Flächen) deutlich: es sind mehr PC-Nichtnutzer als Nutzer mit Hauptschulabschluss und eher Ältere. Somit läßt sich jedes Segment untersuchen - am Bildschirm durch die hohe Interaktivität besser als auf dem Papier.

2.1.3 Mosaic Plots und multiple Balkendiagramme

Mosaic Plots stellen eine Erweiterung von Spine Plots dar. Auf der x-Achse wird die erste Variable (z. B. PC-Nutzung) abgetragen, die y-

Achse nach der zweiten Variablen. Für das Beispiel PC-Nutzung und Bildung sieht das folgendermaßen aus:

ABBILDUNG 27 Mosaic-Plot mit Mondrian (PC-Nutzung und Bildung)

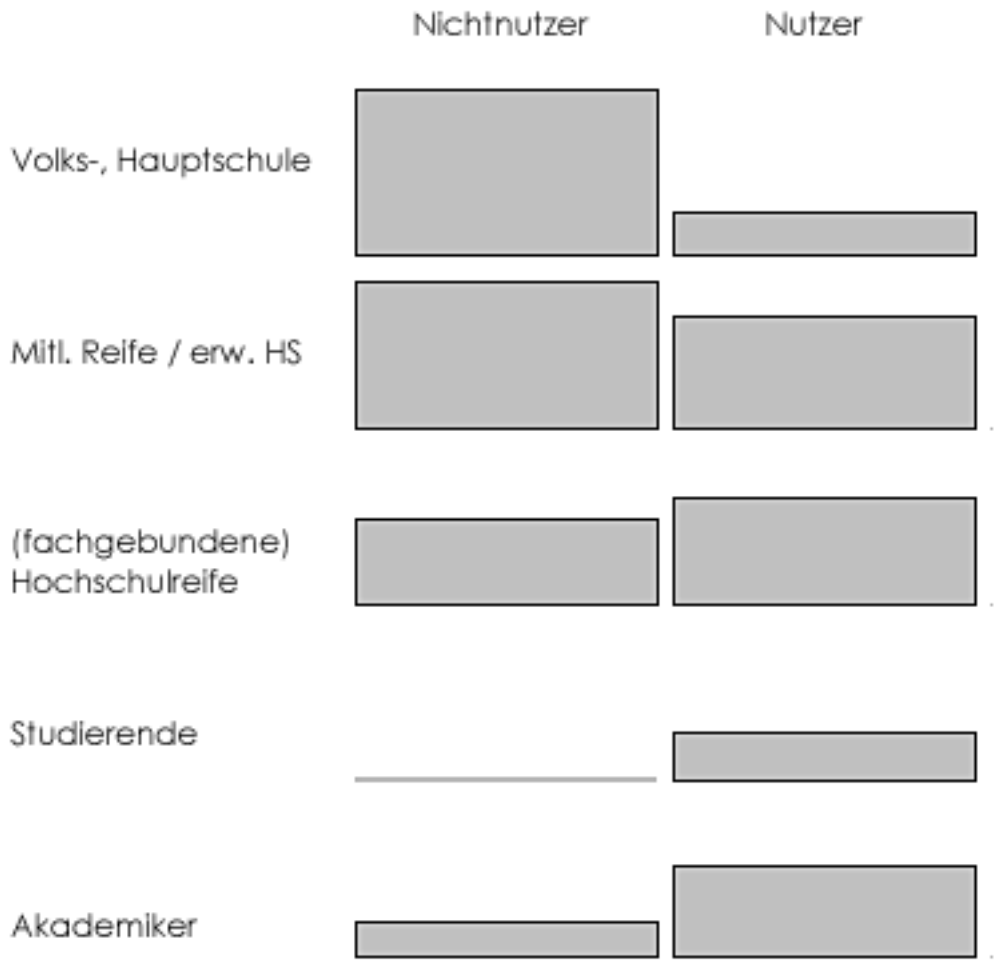


Die linken Balken repräsentieren die Nichtnutzer, die rechten die Nutzer. Die Bildungsabschlüsse sind von oben (Hauptschulabschluss) nach unten (Hochschulabschluss) angeordnet. Die Größe der Rechtecke entspricht der tatsächlichen Fallzahl. So ist auf der linken Seite der Anteil der Studierenden kaum zu erkennen (vorletzter Balken): durch die geringe Fallzahl ergibt sich nur ein ganz dünner grauer Strich.

Deutlich läßt sich z. B. ablesen, dass PC-Nutzung und Bildung schon bei den unteren Schulabschlüssen deutliche Unterschiede zeigt: etwa eine gleichhohe Nutzerzahl mit Haupt- und Realschule steht Nichtnutzern mit Hauptschulabschluss gegenüber.

Etwas übersichtlicher ist evtl. das multiple Balkendiagramm:

ABBILDUNG 28 Multiples Balkendiagramm mit Mondrian (PC-Nutzung und Bildung



Hier werden die Anteile deutlicher: von oben nach unten werden die Bildungsabschlüsse, von links nach rechts die PC-Nutzung definiert. Auf einen Blick ist ersichtlich, dass die Anteile bis zur (fachgebunde-

nen Hochschulreife) bei den Nichtnutzern höher sind, darüber bei den Nutzern.

Es ist auch möglich, die Verfahren der Parallelen Koordinaten und Boxplots in einer Grafik darzustellen - leider läuft an dieser Stelle das Programm noch nicht ganz stabil. Allerdings bringt dieses Feature keinen zusätzlichen Informationsgewinn.

Die Vorteile multivariater Visualisierungstechniken liegen in der hohen Anschaulichkeit auch für komplexe Fragestellungen, der hohen Zugänglichkeit und gestatten somit einen guten Überblick über die rechnerisch zu erwartenden Ergebnisse. Sie können sowohl explorativ als auch hypothesengeleitet eingesetzt werden. So lassen sich auch Cluster durchaus mit diesem Verfahren erkennen. Eine weitere Stärke dieser Verfahren, was „auf dem Papier“ nicht so deutlich wird, ist die hohe Interaktivität am Rechner. Hier lassen sich mit einem Mausklick sofort die Anteile der anderen, zu untersuchenden, Variablen anzeigen.

Nachteile dieses Verfahrens ist - zumindest für den Nichtinformatik-Bereich - die im Augenblick fehlende Implementierung in Standardsoftware ohne Limitationen (z. B. Excel) bzw. die für viele Geisteswissenschaftler sicherlich komplexe Konvertierung von Dateien in für diese Zwecke entwickelten Java-Programme und deren Formate. Auch der Grafikexport aus xmdv ist im Augenblick nur über die Funktion der Bildschirmkopie (ALT-rechte SHIFT-Taste-DRUCK) möglich. Die Qualität der ausgegebenen Grafik könnte qualitativ verbessert werden. Einen Fortschritt bietet Mondrian mit dem Export über die Windows-Zwischenablage. Allerdings wären hier auch Verbesserungen der Grafikausgabe wünschenswert.

Ein schwerwiegenderes Problem ist der Umgang mit fehlenden Werten, die Mondrian im Augenblick nicht bzw. nur als Kategorie darstel-

len kann, was zu erschwerten Interpretationen führt. Ein weiteres Problem ist, dass keine Prozentanteile in der Grafik dargestellt werden können - die Interpretation würde sehr erleichtern werden.

Auch wenn die mangelnde Funktionalität es im Augenblick in dieser Arbeit an der einen oder anderen Stelle so aussehen ließe, als wären die Grafiken dilettantisch - die Möglichkeiten der Visualisierung für zukünftige Arbeiten, auch wenn sie im Augenblick noch nicht perfekt funktionieren, kann nicht hoch genug bewertet werden. Dies gilt auch für die Möglichkeiten, die sich den Geisteswissenschaften bieten - für den qualitativen wie für den quantitativen Fall. Man kann davon ausgehen, dass innerhalb einiger Jahre diese Verfahren wie selbstverständlich in Standardsoftware implementiert sein werden.

2.2 Entscheidungsbäume: eine Einordnung

2.2.1 Einführung

In der Regel stehen statistische Verfahren am Ende jeder quantitativen Untersuchung. Eine Vielzahl von Daten werden also in statistischen Kennwerten zusammengefaßt (z. B. Mittelwert, Cluster, Faktoren, Standardabweichung, etc.). Zur Darstellung der Ergebnisse werden in der deskriptiven Statistik zweidimensionale Diagramme (Balken, Säulen, Kreis, Linien, Lorenzkurve, ...) herangezogen. Eine Dreidimensionalität (z. B. dreidimensionale Säulen) dient zumeist eher der ästhetischen Grafikdarstellung denn der Erhöhung des Informationsgrades - obwohl drei- oder vieldimensionale Darstellungen immer wichtiger werden. Im Bereich der multivariaten Statistik gibt es wenige Verfahren, die ihre Ergebnisse anschaulich in Grafiken umsetzen (z. B. Korrespondenzanalyse, MDS).

Seit Beginn der 1990er Jahre hat sich eine unüberschaubare Variablen- und Datenmenge angesammelt, die mit herkömmlichen statistischen Verfahren nicht mehr ausgewertet werden kann:

> „As the amount of information available for access has grown rapidly in the last few decades, researchers on finding new tools and techniques for navigating and analysing information have become more and more important. One of the effective ways to amplify cognition is the use of computer-supported, interactive visual representations of abstract, non-physically based data that are called information visualization." (NGUYEN und HUANG (2003: 3))

WEGMAN (2003a: 6) zeigt dieses Problem anhand der Huber Taxonomy of Data Set Sizes auf:

TABELLE 5 THE HUBER TAXONOMY OF DATA SET SIZES (ZIT. NACH WEGMAN (2003: 6))

Beschreibung	Data Set Size in Bytes	Storage mode
Tiny	10^2	Piece of Paper
Small	10^4	A few Pieces of Paper
Medium	10^6	A Floppy Disk
Large	10^8	Hard Disk
Huge	10^{10}	Multiple Hard Disks, e. g. Raid Storage
Massive	10^{12}	Robotic Magnetic Tape, Storage Silos

Die vorhandenen technischen Voraussetzungen und die Digitalisierung von Daten bieten die Möglichkeiten, Informationen zu sammeln (z. B. Kundendaten über das Internet) und diese kostengünstig - für spätere Auswertungen - zu speichern. WEGMAN (2003: 24) zeigt dies am Beispiel für die Luftüberwachung auf: bei 6 - 12 Radarstationen, mehreren hundert Flugzeugen und 64 Byte Daten pro Radarstation,

Flugzeug und Antennendrehung entstehen 1 MB Daten pro Minute und rund 1,4 Gigabyte pro Tag. Nach der Huber Taxonomy of Data Set Sizes gilt die Menge der erhobenen Daten pro Tag schon als groß und benötigt eine Festplatte.

Hier setzt die Idee des Data Mining an: wenn Daten grundsätzlich entstehen (z. B. bei der Bestellung von Büchern über das Internet: email-Adresse, Produkte, die sich der Kunde angesehen hat, Lieferadresse, Bankverbindung, Geburtsdatum, bestellte Waren, etc.) - warum sollen die Informationen (z. B. Lesegeschmack, Alter, Geschlecht) nicht gesammelt und ausgewertet werden - in Form von individuellen Buchvorschlägen oder in einem besonderen Zuschnitt des Angebots?

Ziel des Data Mining ist es also, große Datenmengen hinsichtlich ausgewählter Variablen bzw. konkreter Fragestellungen zu strukturieren. Dieser Prozeß sollte insoweit theoriegeleitet sein, dass Vermutungen zugrundeliegen, denn Data Mining ist wie jedes andere statistische Verfahren auf Grundannahmen angewiesen, um sinnvolle Ergebnisse zu liefern.

Data Mining Verfahren, die sich zum Teil auf Statistik, zum Teil auf Visualisierung stützen, finden in verschiedenen Bereichen (Künstliche Intelligenz, Marketing, Medizin, Pharmazie, etc.) weite Verbreitung.

Die Definitionen des Data Mining sind - nach Standpunkt und Wissenschaft - vielfältig:

> „Data Mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.“ (TWO CROWS CORPORATION (1999: 1))

In dieser (sehr knappen) Definition findet sich einer der Grundgedanken des Data Mining (DM) wieder, die Vorhersage von gültigen Aussagen. Allerdings geht DM wesentlich weiter:

> „Data Mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules." (BERRY and LINOFF (2000: 7))

In dieser eher problemorientierten Definition werden zwei neue Aspekte des DM angesprochen: die große Datenmenge und die (halb-)automatische Verarbeitung. Die Automatisierung deutet keinesfalls auf Theorielosigkeit (häufig ein großes Vorurteil gegen diese Verfahren) von DM hin:

> „Data Mining is a tool, not a magic wand. It won´t sit in your database watching what happens and send you e-mail to get your attention when it sees an interesting pattern. It doesn´t eliminate the need to know your business, to understand your data, or to understand analytical methods. Data Mining assists business analysts with finding patterns and relationships in the data - it does not tell you the value of the patterns to the organization. Furthermore, the patterns uncovered by data mining must be verified in the real world." (TWO CROWS CORPORATION (1999: 1))

Die Gefahr, auf die hier hingewiesen wird, bezieht sich grundsätzlich auf die gesamte Statistik. Kennwerte müssen immer in Beziehung zur Realität interpretiert werden. Ein gutes Beispiel hierfür ist der Wert der Signifikanz: ist z. B. ein Zusammenhang sinnvoll, der Wert jedoch nicht signifikant, wird das Ergebnis in einem Forschungsbericht nicht verallgemeinert (auf die Grundgesamtheit übertragen) werden. Umgekehrt wird aber ein Ergebnis, das zwar hochsignifikant ist, aber keinen inhaltlich erklärbaren Zusammenhang liefert, eher nicht in einem Forschungsbericht auftauchen. Weder Statistikpakete wie SPSS oder Data Mining-Programme wie Clementine kennen die Realität und können demnach ihre Ergebnisse selbst „überprüfen". Dies bleibt in jedem Fall eine Aufgabe des Forschers.

2.2.2 Überblick über ausgewählte Data Mining Techniken

Die wichtigsten Data Mining-Techniken sind die Clusterung, Entscheidungsbäume und neuronale Netze (vgl. BERRY und LINOFF (2000: 102ff.)), die in nachfolgender Tabelle kurz dargestellt werden. Wäh-

rend die ersten beiden Verfahren auf statistischen Grundlagen beruhen, versuchen neuronale Netze als „selbstlernende“ Systeme Muster innerhalb von Samples zu erkennen, indem begonnen wird, Fälle zu untersuchen, Vorhersagen zu generieren - und diese Vorhersagen zu korrigieren, wenn neue Daten in die Analyse eingehen. Das macht es sehr schwierig, die gefundene Lösung nachzuvollziehen.

TABELLE 6 AUSGEWÄHLTE DATA-MINING-VERFAHREN

Verfahren	Erläuterung
automatische Clusterung (K-means-Clusterung)	Die Zahl der Cluster wird (theoriegeleitet oder explorativ) vorgegeben. Durch einen iterativen Prozeß wird versucht, möglichst gute Clusterzentren anzustreben. Ziel ist es, möglichst homogene Gruppen zu finden, die sich maximal kontrastieren. Dieses Verfahren ist für Fragestellungen geeignet, die eher nach Strukturen in den Daten suchen. Möglicherweise kann eine Interpretation dadurch erschwert werden (vgl. BERRY und LINOFF (2000: 103ff.)), wenn vorher keine Theorie- oder Hypothesenbildung erfolgt. Eine 3-Cluster-Lösung bringt in diesem Fall sicherlich ein anderes Ergebnis als eine 5-Cluster-Lösung. Die Auswahl und Interpretation der „richtigen“ bzw. optimalen Clusterlösung stellt hohe Anforderungen an den Forscher.[a]
Entscheidungsbäume	bieten die Möglichkeit, anhand einer abhängigen Ziel- und mehrerer unabhängiger Variablen Cluster zu bilden, die sich stufenweise - je nach Stärke des Zusammenhangs - aufbauen. Ziel ist es, Vorhersagen zu treffen oder Regressionen durchzuführen. Ein Kriterium für die Güte eines Baumes ist es, Fehlklassifikationen zu minimieren - bei möglichst homogenen Untergruppensegmentierungen (vgl. SUTTON (2003: 21)).
Neuronale Netze	Dieses Verfahren basiert nicht auf statistischen Formeln, sondern „lernt“ anhand der Daten, indem diese überprüft werden und Korrekturen vorgenommen werden, die dann wiederum in die Lösung eingehen. Ein einfaches Beispiel aus dem Alltag ist das Erkennen bekannter Menschen in einer Menge: anhand z. B. eines Kleidungsstücks wird jemand als bekannt eingeschätzt, es werden aber weitere Kriterien (z. B. Haarfarbe, Gesicht, Größe, ...) zur Überprüfung herangezogen. Sehr komplexes Verfahren mit schwer zu überprüfender Lösung

a. Beim explorativen Vorgehen kann durch Vergleich der Varianzen im Gesamtmodell und der einzelnen Clusterlösungen die Homogenität der Cluster überprüft werden. Der Sinngehalt muss jedoch hier durch den Forscher erfolgen.

3 Entscheidungsbäume

Entscheidungsbäume segmentieren („splitten“) Datensätze anhand einer abhängigen und verschiedener unabhängiger Variablen in Subgruppen. Das Ergebnis wird mathematisch und grafisch als Baumstruktur ausgegeben. Da die Methode der Entscheidungsbäume das hauptsächlich angewandte Verfahren in dieser Arbeit ist, soll nachfolgend auf die Logik des Verfahrens eingegangen werden.

Ausgangspunkt eines Entscheidungsbaums ist eine einfache Häufigkeitstabelle:

ABBILDUNG 29 Häufigkeitsverteilung: PC-Nutzer (N = 2047)

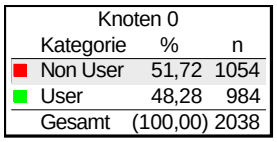
pcuser Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,00 Non User	1054	51,5	51,7	51,7
	2,00 User	984	48,1	48,3	100,0
	Gesamt	2038	99,6	100,0	
Fehlend	7,00	9	,4		
Gesamt		2047	100,0		

Dieser SPSS-Output zeigt, dass 1054 (51.7 gültige Prozent) PC-Nichtnutzer 984 (48.3 gültige Prozent) PC-Nutzern gegenüberstehen. Diese beiden Kategorien sind zur Basis 100 % in Beziehung gesetzt. Neun Befragte haben keine Angaben gemacht (.4 %). Sie tauchen in der Spalte „Prozent“ auf, nicht jedoch in der Spalte „Gültige Prozent“, da eine spezielle Auswertungskategorie von 9 Fällen nicht interessant ist.

Diese Häufigkeitsverteilung bildet den Ausgangspunkt für einen Entscheidungsbaum. Da es wenig Sinn macht, die fehlenden Werte zu segmentieren, werden diese Fälle - wie bei SPSS - aus der Analyse ausgeschlossen und auch nicht angezeigt. Bei diesem Beispiel handelt es sich um ein typisches Klassifikationsbeispiel:[44]

ABBILDUNG 30 PC-Nutzer:Wurzelknoten.



Der Knoten 0 - auch Stamm-, Haupt- oder Wurzelknoten genannt - enthält die gleichen Informationen wie eine Häufigkeitsverteilung in SPSS: die Anzahl und die Anteile der jeweiligen Variablenausprägungen.[45] Ziel ist es, anhand der gewählten unabhängigen Variablen Alter und Bildungsgrad möglichst homogene Untergruppen von PC-Nutzern und Nichtnutzern („Unterknoten“) zu finden. Die Homogenität wird aufgrund verschiedener statistischer Maße definiert, die von Algorithmus zu Algorithmus differieren (z. B. Chi-Quadrat, F-Test, Gini) und selbstverständlich auch vom Skalenniveau der eingesetzten Variablen abhängen.[46] Dahinter stehen verschiedene Konzepte (Statistik, Entropie, Maschinelles Lernen, ...).

Angenommen, der Chi-Quadrat-Test wird für die Untersuchung (nominaler) Daten herangezogen. Das Ergebnis ist eine Maßzahl für die Ähnlichkeit bzw. Unähnlichkeit der erwarteten und die beobachteten Häufigkeiten. Je mehr die erwarteten von den beobachteten Werten differieren, desto größer wird der Chi-Quadrat-Wert (siehe auch Abbildung 10 auf Seite 61).

Bei einer Befragung von 50 Männern und 50 Frauen zur PC-Nutzung (ja - nein) würde man ohne Vorwissen eine Gleichverteilung unterstellen: 25 Männer sind PC-Nutzer, 25 Nichtnutzer, je 25 Frauen nutzen

44. Es gibt zwei Arten von Entscheidungsbäumen: Klassifikationsbäume, die hauptsächlich in dieser Arbeit vorkommen und Regressionsbäume. Klassifikationsbäume basieren auf nominalen bzw. ordinalen, Regressionsbäume auf metrischen Variablen.

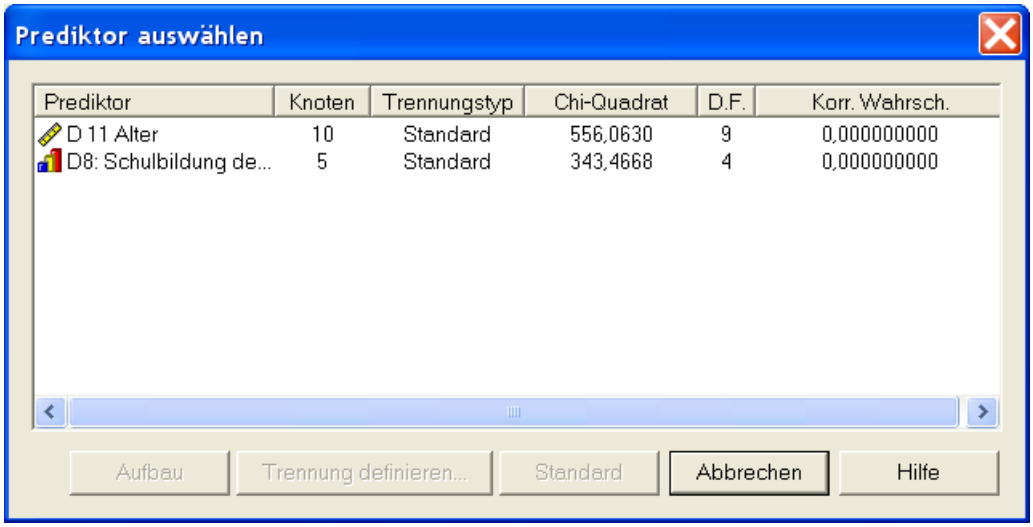
45. Es ist auch möglich, sich die Anteile (zusätzlich) auch als Balkendiagramme anzeigen zu lassen.

46. ANSWERTREE von SPSS unterscheidet vier Algorithmen zur Errechnung von Entscheidungsbäumen: CHAID, EXHAUSTIVE CHAID, CART (= C&RT) und QUEST.

den PC bzw. nicht. Wenn bei einer empirischen Untersuchung herauskommt, dass dieses Ergebnis tatsächlich auch in der Realität zu finden ist, ist der Chi-Quadrat-Wert gleich 0 - die erwarteten und die beobachteten Häufigkeiten sind identisch. Somit gibt es keine Unterschiede zwischen den untersuchten Gruppen. Die PC-Nutzung wäre also unabhängig vom Geschlecht. Je mehr allerdings die empirisch erhobenen Werte von den unterstellten Werten abweichen, desto höher wird der Chi-Quadrat-Wert. Je höher dieser Wert, desto deutlicher unterscheiden sich die Gruppen anhand der unabhängigen Variablen.

Das Ziel eines Chi-Quadrat-Tests im Sinne von Entscheidungsbäumen ist es, in einem ersten Schritt zu prüfen, welche unabhängige Variable den größten Chi Quadrat-Wert bei (vorbestimmter) statistischer Signifikanz besitzt.

ABBILDUNG 31 EXHAUSTIVE CHAID-Algorithmus: Statistische Werte der Prädiktoren (abhängige Variable: PC-Nutzung, unabhängige Variablen: Alter, Schulbildung)[47]



In der Spalte „Prediktor“ werden die unabhängigen, in die Analyse eingehenden Variablen, geordnet nach den Einflüssen auf die ab-

hängige Variable, angegeben. Im Beispiel weist das Alter mit 556 einen höheren Chi-Quadrat-Wert als die Schulbildung (343) auf. Das Alter hat somit eine höhere Bedeutung, bezogen auf die PC-Nutzung, als der Bildungsgrad. Beide Variablen sind höchst signifikant mit .000 (korrigierte Wahrscheinlichkeit). Die Trennung erfolgt aufgrund des standardmäßig ausgewählten Algorithmus, der Chi-Quadrat-Werte zur Analyse heranzieht. Die Spalte D, F bezieht sich auf die sog. „Freiheitsgrade". Diese sind bei der Chi-Quadrat-Statistik abhängig von der Anzahl der Spalten und Zeilen einer Tabelle.

Der Wert, der in der Spalte „Knoten" angegeben ist, entspricht der Anzahl der Merkmalsausprägungen (fünf Bildungsabschlüsse, zehn gefundene Alterskategorien). Da das Alter der Befragten zwischen 15 und 94 Jahren liegt, wurden vom Algorithmus 10 Altersgruppen, die sich hinsichtlich der PC-Nutzung unähnlich sind (statistisch: hohe Chi-Quadrat-Werte aufweisen) gebildet.

Bei der Chi-Quadrat-Statistik ist es wichtig - um es nochmals zu wiederholen - ob sich die erwarteten und beobachteten Häufigkeiten unterscheiden. Als Kriterium werden die Randsummen herangezogen. Der Chi-Quadrat-Wert ist dann hoch, wenn die erwarteten und beobachteten Häufigkeiten in den einzelnen Zellen sich deutlich unterscheiden. Anhand der (standardisierten) Residuen mit einen Absolutwert von 2 lassen sich Abweichungen in der Tabelle deutlich erkennen.

Die Zahl der Freiheitsgrade folgt einer ähnlichen Logik und betrachtet die Anzahl der Spalten und Zeilen einer Kreuztabelle.

47. Die Symbole vor den Variablennamen bezeichnen das jeweilige Skalenniveau: nominale Variablen werden als drei sich überlappende Kreise (siehe unten) dargestellt. Ordinale Variablen (z. B. Schulbildung) werden als drei unterschiedlich hohe Balken und metrische Daten (z. B. Alter) als eine Art „Lineal" für das Symbol für fortlaufende Wertebereiche stilisiert.

Die Berechnung der Freiheitsgrade ist denkbar einfach und lautet:

D, F = (Anzahl der Zeilen - 1) (Anzahl der Spalten - 1).

Die Freiheitsgrade geben an, wieviele Zellen man frei besetzen kann, um anschließend durch die Differenz der Randsummen auf den Rest der Tabelle zu schließen. Bei einer 2 x 2-Tabelle bedeutet das:

D, F (2, 2) = (2 - 1) (2 - 1) = 1.

ABBILDUNG 32 PC-Nutzung nach Geschlecht (N = 2038)

	Männer	Frauen	gesamt
PC-Nichtnutzer		1000	1054
PC-Nutzer			984
	959	1079	2038

Wenn also die Randsummen und die Anzahl der (in diesem Beispiel) weiblichen Non-PC-User bekannt ist, kann durch einfache Differenzbildung zuerst auf die nichtnutzenden Männer, anschließend auf den Rest der Tabelle geschlossen werden.

Für den Entscheidungsbaum bedeutet das, dass 9 Zellen bei den Altersgruppen, vier bei den Bildungsgruppen frei besetzt werden können.

D, F (Alter) = (10 Altersgruppen-1) (2 PC-Nutzergruppen-1) = 9.

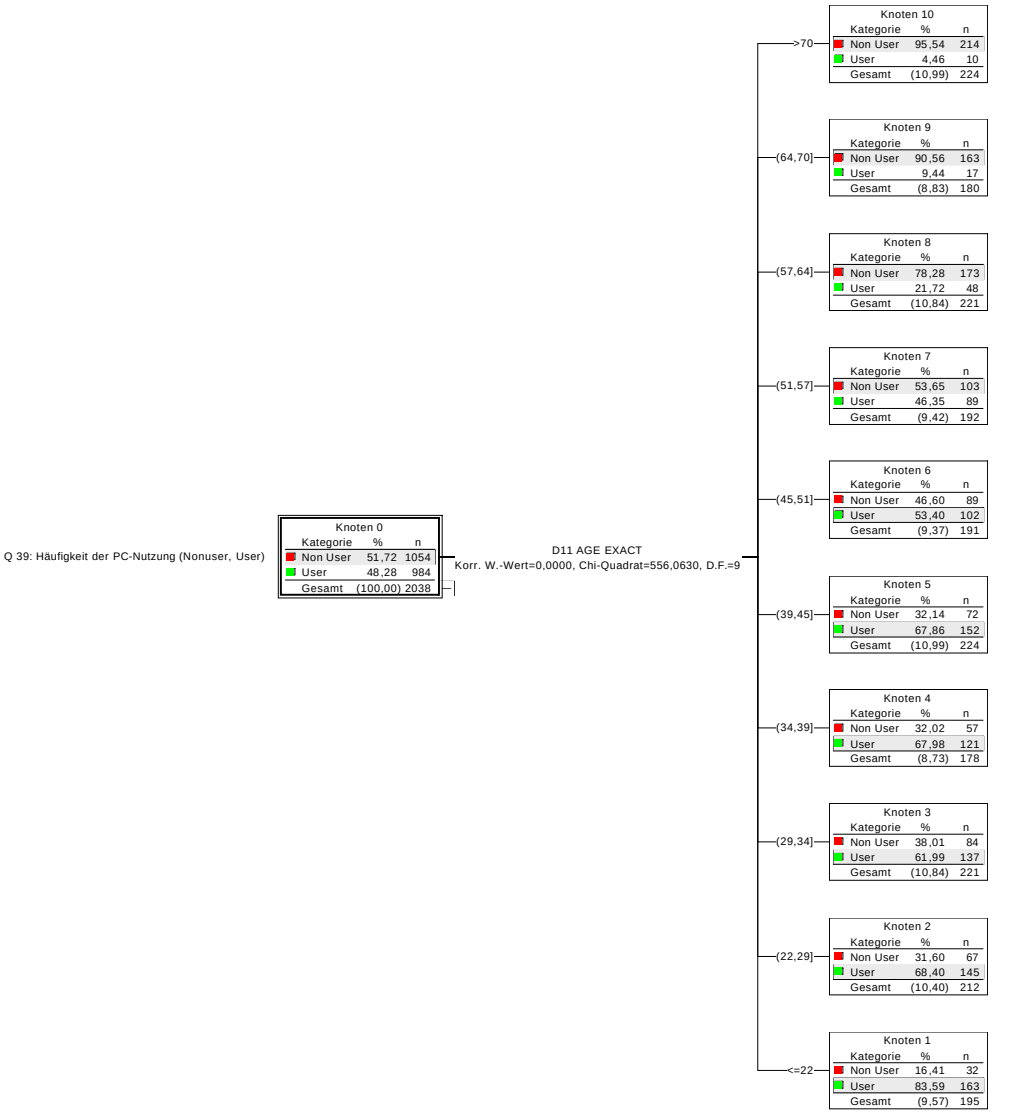
D, F (Bildung) = (5 Bildungsgruppen-1) (2 PC-Nutzergruppen-1) = 4.

Die Logik hinter den Freiheitsgraden ist eine zufallstheoretische Überlegung: so ist z. B. bei zehnmaligem Münzwurf zu erwarten, dass sich 5 x Wappen und 5 x Zahl ergibt. Es ist jedoch durchaus denkbar, dass bei unendlich vielen Würfen es auch einmal vorkommt, dass 10 x Wappen erscheint. Übertragen auf das Beispiel der PC-Nutzung und dem Alter bedeutet das: je geringer die Freiheitsgrade (bzw. die Anzahl der Zellen einer Tabelle) sind, desto höher ist auch die Gefahr eines sich zufällig ergebenden Wertes. Je größer also eine Tabelle ist,

desto unwahrscheinlicher ist es, dass die Ergebnisse zufällig zustandekommen (vgl. NORUSIS (1998: 313f.)). Folglich wäre die Wahrscheinlichkeit, dass sich die Kreuztabelle aus Bildungsgrad und PC-Nutzung mit 4 Freiheitsgraden eher zufällig ergibt als die der Altersklassen mit 9 Freiheitsgraden.

Bei der abhängigen Variablen PC-Nutzung Ja bzw. Nein und den beiden unabhängigen Variablen Alter und Schulbildung ergibt sich also beim sog. EXHAUSTIVE CHAID-Algorithmus (vgl. Abbildung 31 auf Seite 108) folgendes Bild: Das Alter weicht mit 556.0630 deutlicher von 0 ab als die Schulbildung mit 343.4668. Würde sich ein Chi-Quadrat-Wert von 0 zwischen Alter und PC-Nutzung ergeben, würde dies bedeuten, dass der PC in allen Altersstufen gleich häufig genutzt wird. Der gleiche Schluß gilt auch für die Schulbildung, wenn auch nicht ganz so deutlich. Da das Alter einen größeren Chi-Quadrat-Wert besitzt, wird für die erste Segmentierung des Baums das Alter herangezogen.

ABBILDUNG 33 EXHAUSTIVE CHAID-Entscheidungsbaum: PC-Nutzung und Alter



Die Darstellungsmöglichkeiten des eingesetzten Programms Answertree 3.1 sind - vor allem für breite Bäume - nicht sehr effizient gelöst: normalerweise werden die Bäume „von oben nach unten“ dargestellt: bei den vielen Unterknoten bietet sich jedoch in diesem Fall eine eher „vertikale“ Ansicht an.

Unterhalb des Hauptknotens entstehen zwei (oder mehr) sog. „Unterknoten“. Für jeden der Unterknoten wird geprüft, ob sich die Gruppe der gefundenen Elemente weiter anhand der unabhängigen Varia-

blen, des Chi-Quadrat-Werts, unter Berücksichtigung der Signifikanz, aufteilen läßt. Ist dies der Fall, entstehen unterhalb der soeben berechneten Unterknoten weitere Unterknoten. Dieser Prozess wird solange fortgesetzt, bis ein Stopkriterium erfüllt ist. Dies kann die Gruppengröße sein - oder auch nichtsignifikante oder unzureichend hohe Chi-Quadrat-Werte. Bei den zehn Segmentierungen wird auch sehr schnell das Ende des Baums erreicht sein.

Jeder Unterknoten enthält absolute und prozentuale Informationen über die Zielkategorie PC-Nutzer bzw. Nichtnutzer. Bei obiger Grafik wird deutlich, dass die PC-Nutzer vor allem in den jüngeren Altersgruppen zu finden sind.

Die unterste - und jüngste - Gruppe ist jünger als 23 Jahre - oder 22 Jahre und jünger, wie an dem „<="-Zeichen ersichtlich, was zu Beginn etwas ungewohnt ist. Altersgruppen werden in Klammern angegeben - und diese können sich auch - wie z. B. bei den beiden jüngsten gefundenen Segmenten durchaus „überlappen". Dies ist möglich, da der Algorithmus nach dem höchsten Chi-Quadrat-Wert segmentiert und 22jährige Nichtnutzer eher dem zweiten Alterssegment zurechnet, wo der Anteil der Nutzer, ausgegeben als N und in Prozent, geringer ist als in der ersten Gruppe. Der höchste Anteil in den jeweiligen Knoten wird schraffiert ausgegeben, so dass - hinsichtlich der Zielvariablen - schnell ein Überblick über die Fragestellung entstehen kann.

Für Entscheidungsbäume sind eine Reihe von Algorithmen verfügbar (vgl. u.a. BREIMAN et al. (1984), SPSS (2001a), LOH und SHIH (1997), BERRY und LINOFF (2000), WITTEN und FRANK (2001, insbes. S. 95ff.)).

Die ersten Entscheidungsbaumalgorithmen wurden in den 60er Jahren entwickelt. Bis heute wurden die Verfahren immer weiter verfeinert und verändert (vgl. u. a. SPSS (2001a: 185ff., BREIMAN et al.

(1984), WILKINSON (1992), NEVILLE (1999: 8ff.)). Statistisches Ziel eines Entscheidungsbaumprozesses ist es, anhand einer abhängigen und mehrerer unabhängiger Variablen möglichst homogene Gruppen zu bilden, die untereinander inhomogen sind - und dies mit möglichst geringen Fehlklassifikationen.[48] Herbei geht es entweder darum, möglichst gut trennende Variablen zu testen oder die unsichtbare Struktur verschiedener Variablen zu untersuchen:

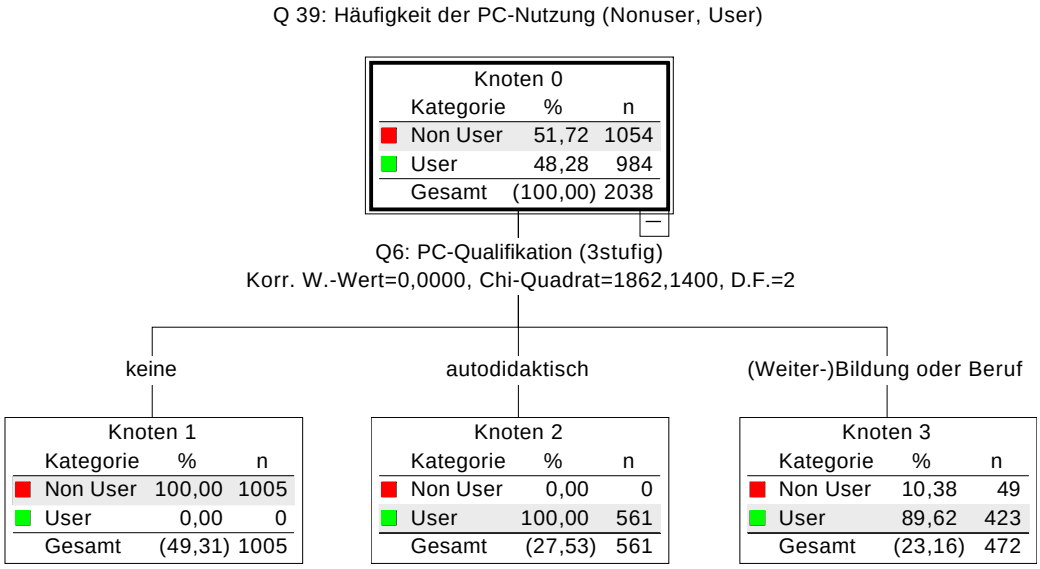
> „Depending on the problem, the basic purpose of a classification study can be either to produce an accurate classifier or to uncover the predictive structure of the problem." (BREIMAN et al. (1984: 6))

Die Ergebnisse dieser Analyse werden als grafisches Baumdiagramm ausgegeben. Die fehlklassifizierten Datensätze werden als „Kosten" errechnet: für nominale/ordinale Daten als Prozentsatz, für metrische Daten als (nicht erklärte) Varianz.[49] Hierbei kommt es nicht ausschließlich um eine möglichst geringe Fehlklassifikation, sondern auch auf eine möglichst gute Interpretierbarkeit der Daten an. Angenommen wir haben eine Zielvariable „PC-Nutzer ja/nein" und eine zweite Variable „PC-Nutzung: beruflich, privat oder sonst", so ist zu erwarten, dass es überhaupt keine Fehlklassifikationen gibt, da der PC nur innerhalb dieser Kategorien benutzt werden kann.

48. Fehlklassifikationskosten entstehen, indem Kategorien der abhängigen Variable irrtümlich vertauscht wurden (z. B. PC-Nutzer irrtümlich als Nichtnutzer erkannt wurden oder umgekehrt).

49. Answertree gibt den Anteil der erklärten Varianz nicht selbständig an, sondern er muss aus der Standardabweichung des Wurzelknotens und der Gesamtvarianz des Modells „per Hand" errechnet werden. Das Handbuch (SPSS (2001b: 199f.)) merkt hierzu an: „Die Risikoschätzung (bei metrischen Variablen, Anm. S. L.) ist nichts anderes als die knoteninterne Varianz ... Die knoteninterne Varianz ist hier 12,5322, während die Gesamtvarianz (die Risikoschätzung für den Baum mit nur einem Knoten) 84,4196 beträgt. Der Anteil der Varianz aufgrund von Fehlern beträgt 12,5322/ 84,4196 = 0,1485. Damit ist der Anteil der durch das Modell erklärten Varianz 100 % - 14,85 % = 85,15 %." Allerdings wird die Gesamtvarianz nicht direkt, sondern nur als Standardabweichung ausgegeben.

ABBILDUNG 34 EXHAUSTIVE CHAID-Entscheidungsbaum: PC-Nutzung und PC-Qualifikation



Der Informationsgewinn geht gegen Null. Alle Nichtnutzer ohne PC-Qualifikation werden richtig (ohne Qualifikation) segmentiert, ebenso diejenigen, die ihre PC-Kenntnisse autodidaktisch erwarben. Nur in der letzten Kategorie gibt es 49 Personen, die den PC aktuell nicht nutzen, irgendwann aber einmal PC-Kenntnisse durch Weiterbildung oder Beruf erwarben. Dieser Zusammenhang kann mit einer einfachen Kreuztabelle leichter und übersichtlicher erzeugt werden.

ABBILDUNG 35 HÄUFIGKEIT DER PC-NUTZUNG NACH PC-QUALIFIKATION (N, SPALTEN-%)

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) * Q6: PC-Qualifikation (4stufig) Kreuztabelle

			Q6: PC-Qualifikation (4stufig)				
			keine	(Weiter-) Bildung	Beruf	autodidaktisch	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Non User	Anzahl	1005	22	27	0	1054
		% von Q6: PC-Qualifikation (4stufig)	100,0%	16,5%	8,0%	,0%	51,7%
	User	Anzahl	0	111	312	561	984
		% von Q6: PC-Qualifikation (4stufig)	,0%	83,5%	92,0%	100,0%	48,3%
Gesamt		Anzahl	1005	133	339	561	2038
		% von Q6: PC-Qualifikation (4stufig)	100,0%	100,0%	100,0%	100,0%	100,0%

In den Zeilen ist die (Nicht-)Nutzung, in den Spalten die Art der PC-Qualifikation abgetragen. Ohne die Tabelle inhaltlich vertiefen zu wollen, wird deutlich, dass die Nichtnutzer in der Regel auch keine PC-Qualifikation besitzen, umgekehrt die Nutzer jedoch in irgendeiner Form die Kenntnisse über den PC (beruflich, durch (Weiter-)Bildung, autodidaktisch) erwarben.

Allerdings gibt es 49 (aktuelle) Nichtnutzer, die schon einmal Kenntnisse erwarben: addiert man in der Zeile der Nichtnutzer die Spalten (Weiter-)Bildung und Beruf, erhält man die 49 Befragten. Inhaltlich bringt dieser Hinweis jedoch keine Erkenntnis, da sich Zeilen und Spalten logischerweise bedingen.

Es ist in diesem Zusammenhang sinnvoller, einige Prozent der Fehlklassifikation hinzunehmen und dafür die Ergebnisse plausibel erklären zu können. Für dieses Beispiel ergibt sich eine Fehlklassifikation von 2.4 Prozent, die aus den 49 Nichtnutzern resultieren - ein erfreulich niedriger Wert, allerdings ohne jegliche Aussagekraft.

Entscheidungsbäume werden empfohlen, wenn Daten (nach bestimmten Regeln) geclustert bzw. Voraussagen getroffen werden sollen:

> „... the data mining task is classification of records or prediction of outcomes. Use decision trees when your goal is to assign each record to one of a few broad categories. Decision trees are also a natural choice when your goal is to generate rules that can be easily understood, explained, and translated into SQL or a natural language." (BERRY und LINOFF (2000: 121))

Auf den ersten Blick - was Klassifikation betrifft - scheint es kaum Unterschiede zur Clusteranalyse zu geben. Die Logiken hinter den beiden Verfahren sind jedoch grundverschieden: Entscheidungsbäume gewinnen ihre Ergebnisse aus der Untersuchung unabhängiger Variablen, bezogen auf eine abhängige, während bei der Clusteranalyse alle zu untersuchenden Variablen „gleichberechtigt" in die Analyse eingehen.

3.1 Interpretationshilfen bei Entscheidungsbäumen

In diesem Unterkapitel werden allgemeine Interpretationshilfen (Fehlklassifikationsmatrix, Gewinnübersicht) für alle verwendeten Entscheidungsbaum-Algorithmen dargestellt. Besonderheiten, die nur einzelne Algorithmen betreffen, werden im jeweiligen Unterkapitel zu finden sein.

3.1.1 Fehlklassifikationsmatrix, Regeln und Übersicht

Die Fehlklassifikationsmatrix ist ein Hauptbestandteil, um das Ergebnis eines Entscheidungsbaums zu bewerten. Es handelt sich hierbei um eine einfache Kreuztabelle, die den Anteil der richtig bzw. falsch klassifizierten Fälle ausgibt.

Fehlklassifikationen entstehen grundsätzlich bei der Aufteilung eines Samples - es sei denn, Zellen wären unbesetzt:

ABBILDUNG 36 PC-NUTZUNG NACH SCHULBILDUNG (N, ZEILEN-%)

D8: Schulbildung der Befragten * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

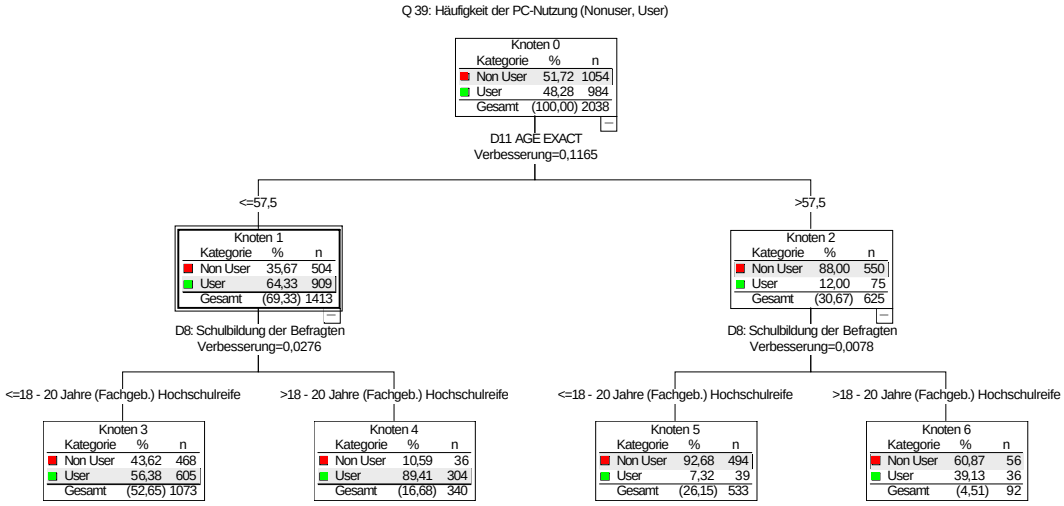
			Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)		
			Non User	User	Gesamt
D8: Schulbildung der Befragten	bis 15 Jahre (Volks-, Hauptschule)	Anzahl	399	107	506
		% von D8: Schulbildung der Befragten	78,9%	21,1%	100,0%
	16 - 17 Jahre (Mittl Reife/erw. HS-Abschluss)	Anzahl	355	275	630
		% von D8: Schulbildung der Befragten	56,3%	43,7%	100,0%
	18 - 20 Jahre (Fachgeb.) Hochschulreife	Anzahl	208	262	470
		% von D8: Schulbildung der Befragten	44,3%	55,7%	100,0%
	Studium	Anzahl	7	121	128
		% von D8: Schulbildung der Befragten	5,5%	94,5%	100,0%
	21 Jahre + abgeschlossenes Hochschulstudium	Anzahl	85	219	304
		% von D8: Schulbildung der Befragten	28,0%	72,0%	100,0%
Gesamt		Anzahl	1054	984	2038
		% von D8: Schulbildung der Befragten	51,7%	48,3%	100,0%

Beispielsweise sind rund 21 % der PC-Nutzer Volks-, bzw. Hauptschulabgänger, rund 79 % Non-User. Dadurch kommt es zwangsläufig - vor allem, wenn noch mehrere andere Variablen berücksichtigt werden, dazu, dass Nutzer als Nichtnutzer klassifiziert werden und umgekehrt. Dies ließe sich nur dann umgehen, wenn alle Akademiker PC-Nutzer und alle Volks-, Haupt- und Realschüler keine Nutzer wären.

Um die Beispiele in diesem Kapitel überschaubarer zu machen, wird zumeist auf einen binären Entscheidungsbaumalgorithmus (CART, QUEST) zurückgegriffen. Im Gegensatz zu dem weiter oben dargestellten EXHAUSTIVE-CHAID-Algorithmus fasst er alle Ausprägungen einer Variable immer zu zwei Unterknoten zusammen. Im Laufe des Kapitels werden Gemeinsamkeiten und Unterschiede herausgearbeitet.

Der CART-Algorithmus liefert für die abhängige Variable PC-Nutzung und die unabhängigen Variablen Alter und Bildungsgrad nachfolgendes Ergebnis, das weiter unten ausführlich erläutert wird.[50]

ABBILDUNG 37 CART-ENTSCHEIDUNGSBAUM: PC-NUTZUNG, ALTER UND BILDUNG



Im Gegensatz zu den zehn Segmenten von EXHAUSTIVE CHAID werden hier immer nur zwei Trennungen durchgeführt. Die wichtigste Variable zur Unterscheidung der Gruppen ist das Alter, gefolgt von der Schulbildung: die PC-Nutzer sind eher jünger und eher besser gebildet.

Der Splitpunkt beim Alter ist 57.5 Jahre. Im linken Teil des Baums sind die bis 57jährigen („<=“), rechts die Älteren (58 Jahre +) enthalten. Die Schulbildung, 5stufig erfaßt, trennt im Knoten 3 diejenigen mit Schulabschlüssen bis einschließlich (!) (fachgebundene) Hochschulreife, im Knoten 4 diejenigen mit Studium, bei denen der Anteil der Nutzer mit knapp 90 % deutlich höher liegt als im Knoten 3 mit rund 56 %. Bei den Älteren, wo die Nichtnutzung mit 88 % (Knoten 2) recht hoch liegt, ergibt sich ein ähnlich deutliches Bild - allerdings prozentu-

50. Um die Ergebnisse überschaubar zu halten, wurden nachfolgend zweistufige Bäume generiert.

al auf einem anderen Niveau. Trotzdem ist der Anteil der PC-Nutzer mit höheren Bildungsabschlüssen über fünfmal so hoch (vgl. Knoten 5 : Knoten 6).

Wenn verschiedene Variablen zur Gruppenbildung herangezogen werden, kann es zu sog. „Fehlklassifikationen" kommen: tatsächliche PC-Nutzer können als Nichtnutzer und umgekehrt klassifiziert werden. Für dieses Beispiel bedeutet das:

ABBILDUNG 38 CART-ENTSCHEIDUNGSBAUM: FEHLKLASSIFIKATIONSMATRIX FÜR PC-NUTZUNG, ALTER UND BILDUNG

Fehlklassifikationsmatrix		Tatsächliche Kategorie		
		Non User	User	Gesamt
Vorhergesagte Kategorie	Non User	550	75	625
	User	504	909	1413
	Gesamt	1054	984	2038

```
Risikostatistiken     Risikoschätzung              0,284102
                      Std.f. der Risikoschätzung   0.00998989
```

Die Anzahl der falsch eingeordneten Personen werden in ein Verhältnis zur Gesamtzahl der Befragten gesetzt:

504 falsch klass. Nichtnutzer + 75 falsch klass. Nutzer = 579.

579 / 2038 = 28.4 %.

„Std.f. der Risikoschätzung" zeigt den Standardfehler der Risikoschätzung. Je kleiner der Wert ist, desto vertrauenswürdiger ist die Risikoschätzung (Fehlklassifikation) - denn desto weniger weicht der Wert in der Grundgesamtheit von dem der Stichprobe ab.[51]

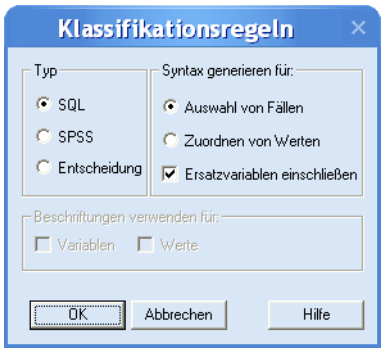
Die Bewertung des Ergebnisses ist nicht pauschal zu treffen, sondern hängt von der Anzahl und den Kategorien der unabhängigen Varia-

51. Der Wert bezieht sich auf die Konfidenzintervalle (= Vertrauensintervalle) einer Schätzung. Näheres findet sich u. a. bei BÜHL und ZÖFEL (2002: 118)).

blen ab. Bei jeder multivariaten Analyse führen eine Vielzahl von unabhängigen Variablen tendenziell zu besseren Ergebnissen als wenige - sie können sich jedoch auch störend gegenseitig aufheben. Gehen Mischtypen in die Analyse ein (z. B. ordinalskalierte Variablen mit den Ausprägungen ja - teils/teils - nein) oder auch das Lebensalter, wird es möglicherweise für den Algorithmus schwieriger, sinnvolle Trennungen zu finden, da bestimmte Eigenschaften nicht mit einem bestimmten Alter verknüpft sind. Tendenziell nutzen wahrscheinlich eher jüngere Personen den PC, aber es gibt natürlich auch Ältere, die dieses Medium in Anspruch nehmen - ein exaktes Alter für die Trennung ist hier wahrscheinlich schwer zu finden.

Regeln ermöglichen den Export der gefundenen Strukturen in andere Anwendungen:

ABBILDUNG 39 EINSTELLUNGSMÖGLICHKEITEN FÜR KLASSIFIKATIONSREGELN BEI ENTSCHEIDUNGSBÄUMEN MIT ANSWERTREE



Hierbei lassen sich verschiedene Einstellungen vornehmen. Ebenso ist auch eine Darstellung der SPSS-Syntax möglich, die sich problemlos kopieren und weiterverarbeiten läßt.

Die Übersicht gibt ein Protokoll des Gesamtprojekts aus - bestehend aus Dateinamen, Anzahl der Fälle, Name der Variablen, Einstellungen des Algorithmus, etc. und dient eher der Kontrolle der über die Menüs erstellten Entscheidungsbäume.

3.1.2 Gewinnübersicht

Neben der Risikostatistik wird eine Gewinnübersicht (vgl. BÜHL und ZÖFEL (2002b: 42ff.)) ausgegeben. Hierbei handelt es sich um eine Tabelle, die angibt, in welchen Knoten die meisten „Gewinne" (im Fall dieser Arbeit: PC-Nutzer), bezogen auf die Zielvariable gefunden wurden. Eine einfache Gewinnübersicht - bezogen auf die PC-Nutzer - sieht folgendermaßen aus:

ABBILDUNG 40 CART-ENTSCHEIDUNGSBAUM: GEWINNÜBERSICHT FÜR PC-NUTZUNG, ALTER UND BILDUNG[52]

Knoten	Knoten Anzahl	Knoten %	Gewinn Anzahl	Gewinn %	Treffer %	Index %
4	340	16.7	304	30.9	89.4	185.2
3	1073	52.6	605	61.5	56.4	116.8
6	92	4.5	36	3.7	39.1	81.0
5	533	26.2	39	4.0	7.3	15.2
	2038	100	984	100		

Die Spalten „Knoten Anzahl" und „Knoten Prozent" geben die jeweiligen Summen bzw. Anteile - bezogen auf die Gesamtzahl von 2038 Befragten - an. Knoten 3 hat hier den größten Anteil (N = 1073) von 52.6 %. Die Gewinnknoten beziehen das Ergebnis auf N = 984, d. h. die Zielvariable PC-Nutzer (605 : 984).

Die Spalte „Treffer %" vergleicht die Spalten „Knoten Anzahl" und „Gewinn Anzahl" und stellt also eine Relation zwischen allen Befragten im Knoten mit den gesuchten PC-Usern dar. Diese recht abstrakte Darstellung soll beispielhaft an der ersten Zeile der Tabelle verdeutlicht werden. Hierzu wird nochmal der Knoten dargestellt:

52. Bei den Spalten „Knoten Anzahl" und „Gewinn Anzahl" handelt es sich jeweils um klassifizierte Personen.

ABBILDUNG 41 CART-ENTSCHEIDUNGSBAUM: KNOTEN 4 FÜR PC-NUTZUNG, ALTER UND BILDUNG

Knoten 4			
	Kategorie	%	n
■	Non User	10,59	36
■	User	89,41	304
	Gesamt	(16,68)	340

Von den 340 Befragten sind 36 Non User und 304 PC-Nutzer, was einem Verhältnis von etwa 11 % zu 89 % entspricht.

ABBILDUNG 42 CART-ENTSCHEIDUNGSBAUM: GEWINNÜBERSICHT FÜR PC-NUTZUNG, ALTER UND BILDUNG - DARSTELLUNG DER KENNZAHLEN

Knoten	Knoten Anzahl	Knoten %	Gewinn Anzahl	Gewinn %	Treffer %	Index %
4	340	340 / 2038 = 16.7	304	304 / 984 = 30.9	304 / 340 = 89.4	30.9 / 16.7 = 185.2
	Prozentuierung auf die Gesamtstichprobe (N = 2038 („Wie hoch ist der Anteil des Knotens insgesamt an der Stichprobe?“)		Prozentuierung auf die PC-Nutzer (N = 984) („Wie hoch ist der Anteil der Zielkategorie (PC-Nutzer) in diesem Knoten?“)		(s. Knoten) („Wie hoch ist der Anteil der Zielkategorie im Knoten (hier: 4)?“	Prozentuierung Gewinn zu Knoten („Wie hoch ist der Gewinnanteil im Knoten bezogen auf den Gesamtanteil des Knotens?“)
	2038	100	984	100		

Im Knoten 4 finden sich die höchsten PC-Nutzer Anteile (!!!) mit 89.41 % (Treffer %, bezogen auf den Gesamtknoten). Knoten 3 hat einen höheren absoluten (Gewinne Anzahl), aber einen geringeren relativen (Gewinne %) Wert. Die Spalten „Knoten Anzahl“ und „Knoten %“ geben die jeweiligen Werte für die Gesamtknoten (also PC-Nutzer und Nichtnutzer) an. Im Knoten 4 finden sich 304 PC-Nutzer, im Knoten 3 605. Obwohl die Anzahl der Nutzer im Knoten 3 fast doppelt so hoch ist wie in Knoten 4, ist das prozentuale Verhältnis der PC User zu

den Nonusern im Knoten 4 mit rund 89 % gegenüber 56 % deutlich höher (Spalte „Treffer %“).[53] Dadurch sind die (kumulierten) Gewinnprozente unterschiedlich, da sie auf die Gesamtzahl der PC-Nutzer von 984 bezogen sind. Knoten 4 enthält 30.9 %, Knoten 3 61.5 % der PC-Nutzer. In den Knoten 3 und 4 finden sich somit rund 90 % der PC-Nutzer.

Die Spalte „Index %“ gibt das Verhältnis der jeweiligen Spalten „Gewinn %“ zu „Knoten %“ an. Im - auf PC-Nutzung basierten - wichtigsten Knoten 4 finden sich 340 von 2.038 Befragen, was einem Anteil von 16.7 % entspricht. Von den 340 Befragten sind 304 PC-Nutzer - bezogen auf 984 sind dies 30.9 %.

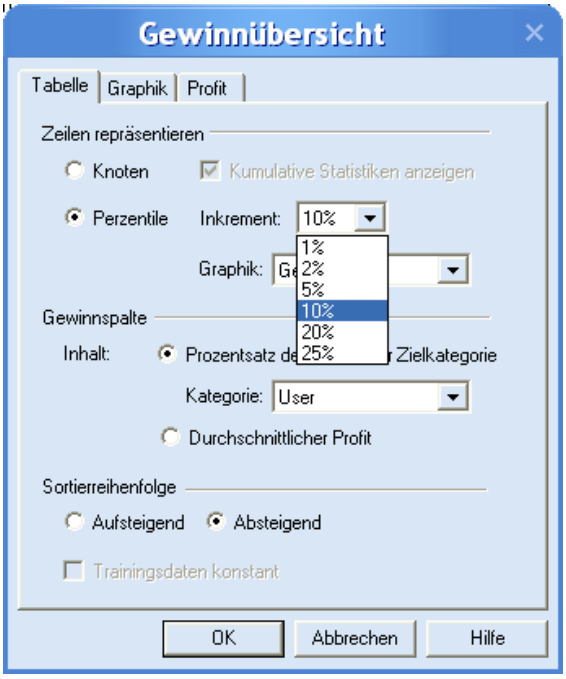
Die Spalte „Treffer %“ gibt an, wie hoch der Anteil der Treffer (in diesem Fall: PC-Nutzer) in den einzelnen Knoten ist. Für den Knoten 4 ergibt sich z. B.: 304 PC-Nutzer : 340 Befragte im Knoten 4 = 89.4 %.

Mit 16.7 % der Stichprobe werden rund 31 Prozent der Zielkategorie identifiziert - der Indexwert liegt mit 185.2 weit über 100 %. Anders im Knoten 6, der nur 4.5 % der Stichprobe enthält und davon nur 3.7 % PC-Nutzer. identifiziert. Der Wert unter 100 trägt somit wenig zur Erklärung der PC-Nutzung bei, da sich mehr Nichtnutzer als Nutzer im Knoten befinden.[54]

53. Diese Logik ist zu Beginn etwas verwirrend. Es geht jedoch bei der Gewinnübersicht darum, die „reinsten“ Knoten mit den höchsten Prozentanteilen - zum Beispiel der PC-Nutzer - zu identifizieren. BÜHL und ZÖFEL (2000: 31ff.) zeigen dies an einem Marketing-Beispiel mit Bestellern einer Zeitschrift. Hierbei interessiert vor allem die Frage, wie sich die Besteller am besten beschreiben lassen, um spätere Marketing-Aktionen darauf abzustimmen - also möglichst wenige potentielle Nichtabonnenten zu erreichen, da z. B. die Portokosten eingespart werden können. Würde man das Beispiel auf das potentielle Abonnement einer PC-Zeitschrift übertragen, könnte man zwar 605 PC-Nutzer aus Knoten 3 und nur 304 User aus Knoten 4 erreichen - es müßten aber bei der Wahl von Knoten 3 1073 Personen angeschrieben werden, wobei nur jeder zweite einen PC besitzt. Bei Knoten 4 sind rund 90 % (304 von 340 Personen) PC-Nutzer - die Fehlstreuung ist also wesentlich geringer.

Die Gewinnübersicht läßt sich auch als Grafik der Perzentile darstellen, die zwischen 1 und 25 Prozent liegen können, Selbstverständlich könnte man auch eine Gewinnübersicht über andere Ausprägungen der Zielvariablen (z. B. Nonuser) vornehmen.

ABBILDUNG 43 EINSTELLUNGSMÖGLICHKEITEN IN DER GEWINNÜBERSICHT BEI ENTSCHEIDUNGSBÄUMEN



Eine Darstellung der 10 % - Perzentile ergibt folgendes - tabellarisches - Ergebnis:

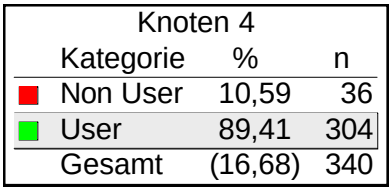
54. Hierbei ist zu berücksichtigen, dass die Prozentwerte in der Gewinnübersichtstabelle auf eine, die im Baumdiagramm jedoch auf zwei Kommastellen gerundet sind. Answertree verwendet bei den Berechnungen (z. B. des Indexwertes) die exakteren zweistelligen Werte, wodurch es bei den Berechnungen innerhalb der Gewinnübersicht zu geringfügigen Abweichungen kommt.

TABELLE 7 CART-ENTSCHEIDUNGSBAUM: GEWINNÜBERSICHT DER PC-NUTZUNG (UNABHÄNGIGE VARIABLEN: ALTER, BILDUNG)

Kno-ten	Perzentil	Perzentil: Anzahl	Gewinn: Anzahl	Gewinn: %	Treffer: %	Index: %
4	10	204	182	18,5	89,4	185,2
4;3	20	408	342	34,8	83,9	173,8
3	30	611	457	46,4	74,8	154,8
3	40	815	572	58,1	70,2	145,3
3	50	1019	687	69,8	67,4	139,6
3	60	1223	802	81,5	65,6	135,8
3;6	70	1427	914	92,9	64,1	132,7
6;5	80	1630	954	97,0	58,5	121,2
5	90	1834	969	98,5	52,8	109,4
5	100	2038	984	100,0	48,3	100,0

Die „wichtigsten“ 10 % der Verteilung - also der oder die Knoten, die die meisten Gewinne (hier: PC-Nutzer) enthalten, finden sich im Knoten 4. Der Datensatz besteht aus 2038 Fällen, also enthält das erste Perzentil 10 % davon, also (gerundet) 204 Befragte. 182 von 204 Befragten sind PC-Nutzer - bezogen auf die Gesamtzahl von 984 Nutzern sind dies 18.5 (Gewinn-) % bzw. 89.4 (Treffer-) % (182 : 204). Knoten 4 verdeutlicht dies nochmal:

ABBILDUNG 44 CART-ENTSCHEIDUNGSBAUM: KNOTEN 4 FÜR PC-NUTZUNG, ALTER UND BILDUNG (GEWINNÜBERSICHT)



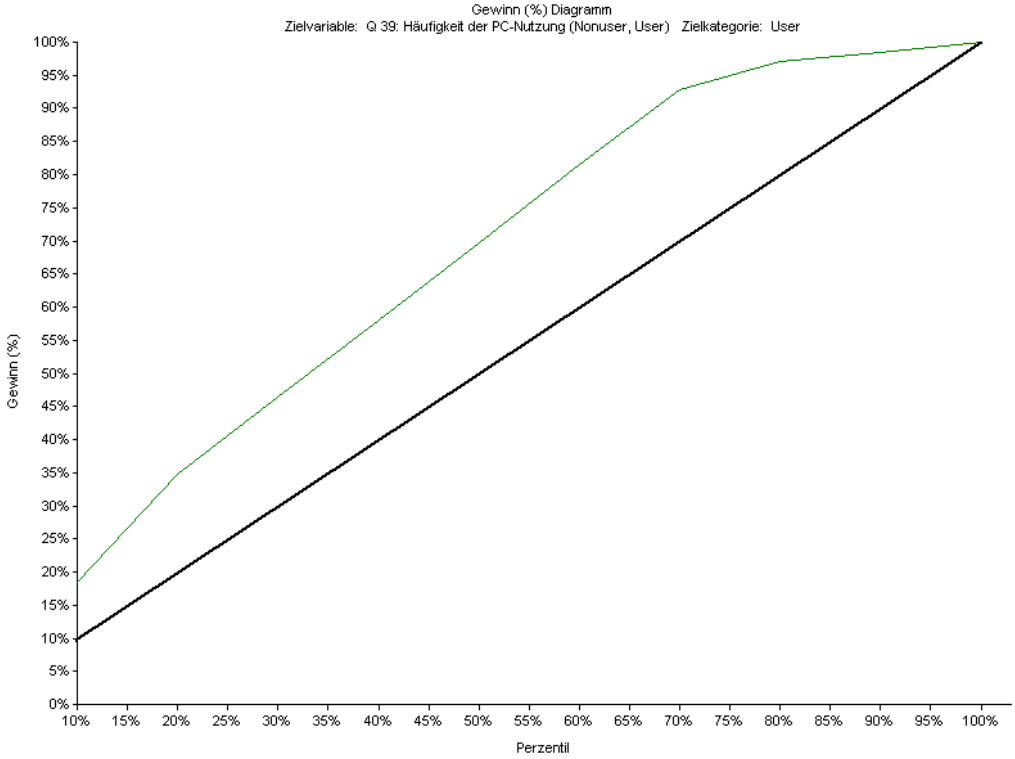
Knoten 4 enthält mehr als 204 Befragte - insgesamt 340. Deshalb ist das erste Perzentil darin völlig enthalten: es enthält sowohl die 182 Nutzer als auch die 22 Nichtnutzer.

4;3	20	408	342	34,8	83,9	173,8

Die zweite Zeile der Tabelle - das 20 %-Perzentil - zeigt, dass neben dem Knoten 4 auch Knoten 3 (der Knoten mit der zweithöchsten PC-Nutzeranzahl) herangezogen wird. 20 % der Verteilung werden durch 2038 * 0.2 = 408 Befragte abgebildet - davon sind insgesamt 342 Nutzer, was 83.9 % der Treffer entspricht. Knoten 4 enthält jedoch nur 304 Nutzer. Somit kommen noch 38 Nutzer aus Knoten 3 dazu - in der ersten Spalte wird ersichtlich, dass die PC-Nutzer aus den Knoten 4 und 3 kommen.

Dieses Ergebnis läßt sich auch grafisch darstellen - hierfür sind die Spalten „Perzentil“ und „Gewinn: %“ relevant:

ABBILDUNG 45 CART-ENTSCHEIDUNGSBAUM: GRAFISCHE GEWINNÜBERSICHT DER PC-NUTZUNG (UNABHÄNGIGE VARIABLEN: ALTER, BILDUNG)



Die diagonal verlaufende Linie unterstellt eine Gleichverteilung: aufgeteilt in „10-Prozent-Schritte" (Perzentile) würde das bedeuten, dass alle Knoten den gleichen Anteil an PC-Nutzern beisteuern. Dieses Ergebnis ist aber unbefriedigend und würde postulieren, dass es keine Unterschiede zwischen den Knoten gäbe. Je weiter nun die „trapezförmig" aussehende Linie von der Gleichverteilungslinie abweicht, desto interessanter ist das Ergebnis und desto mehr unterscheiden sich die Gruppen hinsichtlich des untersuchten Merkmals.[55]

Das nächste Perzentil bei 20 enthält 34.8 % der Nutzer, die sich aus den Knoten 4 und 5 ergeben. Die Kurve steigt etwas flacher an. Dies

55. Es ist irritierend, dass die x-Achse bei 10 % beginnt und nicht - wie in der Mathematik üblich - bei 0. Somit ist auch die Gerade „nach oben verschoben".

setzt sich fort, bis das Ende der Kurve wieder die Gerade schneidet - bei 100 %.

Diese Kurve erinnert stark an die sog. „Lorenzkurve“ - ein grafisches Verfahren, um den Gini-Index abzubilden. Dieses Maß ist vor allem in den Wirtschaftswissenschaften verbreitet und wird u. a. bei der Konzentrationsmessung verwendet. Ein Beispiel für Konzentrationsmessung liegt im Vergleich der (kumulierten) Marktanteile von Unternehmen zu deren jeweiligen Umsatz:

TABELLE 8 KONZENTRATIONSMESSUNG: IDEALTYPISCHER VERGLEICH ZWISCHEN POLYPOL UND MONOPOL

Polypol („alle Anbieter haben den gleichen Anteil am Umsatz“)		Monopol („ein Anbieter vereint den Gesamtumsatz auf sich“)	
Marktanteil	Umsatzanteil	Marktanteil	Umsatzanteil
10 %	10 %	99 %	1 %
20 %	20 %	1 %	99 %
30 %	30 %		
...	...		
100 %	100 %		

Die kleinsten 10 % der Unternehmen erzielen im Fall des Polypols 10 % des Umsatzes, usw. Es liegt somit eine Gleichverteilung von Marktanteil und Umsatz vor. Im Fall des Monopols ist es umgekehrt: 99 % aller Unternehmen auf dem Markt erzielen 1 % des Umsatzes, ein Unternehmen (bei betrachteten 100 Unternehmen) erzielt 99 %, konzentriert somit den gesamten Markt auf sich.

Es handelt sich hierbei um eine idealtypische Darstellung, um den Sachverhalt zu verdeutlichen: im Fall des Polypols erzielen 10 % der Unternehmen 10 % des Umsatzes, die nächsten 10 % (die Werte werden kumuliert!) ebenfalls wieder 10 %, etc. Einfach ausgedrückt: alle

Anbieter auf diesem Markt erzielen den gleichen Umsatz - der Umsatz ist gleichverteilt. Der Gini-Koeffizient, auf dessen Berechnung nicht eingegangen werden soll, kann Werte zwischen 0 und 1 annehmen: beim Wert 0 herrscht völlige Gleichverteilung (wie in diesem Fall). Anders in den rechten Spalten der Tabelle, dem Fall des Monopols, wo (idealtypischerweise) ein Unternehmen (bzw. 1 % der Unternehmen) 99 % des Umsatzes erzielt, die anderen Unternehmen anteilsmäßig so gut wie keinen Umsatz erzielen (1 %). Der Gini-Koeffizient liegt also hier nahe 1 und zeigt eine hohe Konzentration an.

Die Lorenzkurve - als grafische Umsetzung des Indexergebnisses - bildet die Achsen aus den betrachteten Kategorien und legt eine Diagonale durch den Nullpunkt. Diese Diagonale symbolisiert die Gleichverteilung (10 % Marktanteil = 10 % Umsatz). Je weiter nun die Ergebnisse von dieser Gleichverteilung abweichen, desto höher ist die Konzentration.

3.2 Entscheidungsbaum-Algorithmen

Der von SUNQUIST und MORGAN 1963 entwickelte AIID (Automatic Interaction Detector) Algorithmus ist vom Ablauf dem aszendierenden Verfahren GEIGERs nicht unähnlich:

> „The algorithm performs stepwise splitting. It begins with a single cluster of cases and searches a candidate set of predictor variables for a way to split this cluster into two clusters. Each predictor is tested for splitting as follows: sort all the n cases on the predictor and examine all n - 1 ways to split the cluster in two. For each possible split, compute the within-cluster sum of squares about the mean of the cluster on the dependet variable. Choose the best of the n - 1 splits to represent the predictor's contribution. Now do this for every other predictor. For the actual split, choose the predictor and its cut point which yields the smallest overall within-cluster sum of squares." (WILKINSON (1992: 4))

Dieses Vorgehen wird für jede Vorhersagevariable wiederholt. Das Ziel für jeden Split ist die Minimierung der Varianz („sum of squares", kleinste Quadrate).

Während allerdings bei GEIGER eher Mentalitätskategorien im Vordergrund standen, sind es hier statistische Tests in Form der Kleinsten Quadrate-Statistik. Im Verlauf des Kapitels wurde beim Eta-Wert die Bedeutung der Varianz bereits erläutert: es ist ein Maß, dass die Abweichung vom Mittelwert kennzeichnet und damit eine Beurteilungsgrundlage liefert, ob der gruppenspezifische Mittelwert eine Verteilung gut oder weniger gut beschreibt: je kleiner die Varianz, desto besser, je größer die Varianz, desto schlechter wird der Mittelwert charakterisiert.

Diese Überlegung macht sich der AID-Algorithmus zunutze und überprüft, für welche Prädiktorvariable die Varianz für alle möglichen Cluster am minimalsten ist - diese wird dann ausgewählt.

Um dies zu veranschaulichen soll ein einfacher Algorithmus in das Denken mit Entscheidungsbäumen einführen:

Gegeben sind mehrere unabhängige (z. B. Alter, Geschlecht, Einkommen, Schulbildung) und eine abhängige Variable (z. B. PC-Nutzung) aus einer Untersuchung mit 2.000 Befragten. Ziel ist es, homogene Untergruppen zu bilden. Folgender einfacher Algorithmus könnte verwendet werden (vgl. NEVILLE (1999: 9)):

- Finde diejenige unabhängige Variable X, die am höchsten mit PC-Nutzung korreliert
- Um die Unterschiede zwischen den Gruppen zu messen, verwende die F-Statistik für metrische, die Chi-Quadrat-Statistik für Nominaldaten. Finde die optimale Trennung
- Wiederhole den Prozess solange, bis eine Stopregel (z. B. Gruppengröße < 30) erreicht ist

Dies ist ein Beispiel für ein typisches Klassifikationsproblem: welche Nutzergruppen lassen sich anhand der Daten segmentieren? Mit den Grundannahmen (Variablenauswahl, z. B. Alter, Geschlecht, Bildung, Einkommen) liegen bereits Hypothesen zugrunde.

Es würde den Rahmen der Arbeit sprengen, alle jemals entwickelten Entscheidungsbaum-Algorithmen darzustellen und zu diskutieren. Neben den in dieser Arbeit verwendeten Algorithmen CHAID, EXHAUSTIVE CHAID, C&RT (= CART[56]) und QUEST wird aufgrund des historischen Verständnisses von Entscheidungsbaumalgorithmen noch kurz auf den AID-Algorithmus von MORGAN und SUNQUIST (1963) eingegangen. Zu dem von QUINLAN entwickelten C5-Algorithmus[57], der bis zu C4.5 noch Freeware war und jetzt kommerziell vermarktet wird (vgl. http:/www.rulequest.com), siehe u. a. WILKINSON (1992), QUINLAN (1993))[58].

Der AID-Algorithmus wird von WILKINSON (1992: 4f.) charakterisiert:

> „Morgan and Sonquist (1963) proposed a simple method for fitting trees to predict a quantitative variable. They called the method AID, for Automatic Interaction Detection. ... because it naturally incorporates interaction among predictors. Interaction is not correlation. ... In the analysis of variance, Interaction means that a trend within one level of a variable is not parallel to a trend within annother level of the same variable.

LEWIS (2000: 2ff.) nennt vier Hauptbestandteile für Klassifikationsprobleme: eine abhängige Variable, verschiedene unabhängige Variablen, der zugrundeliegende Datensatz und die Vorhersagekraft für die untersuchten Einheiten (Modellfit).

Neben den Klassifikationsproblemen gibt es eine Reihe weiterer Anwendungsgebiete für Entscheidungsbäume, z. B. Vorhersagen, Datenreduktion und „Datenscreening“, Entdecken von statistischen Interaktionseffekten, gruppenspezifische Bewertung (z. B. in hohes, durchschnittliches und niedriges Risiko), usw. (vgl. SPSS (2001a: 5)).

56. CART und C&RT werden synonym verwendet.

57. Dieser Algorithmus ist im Datamining-Programm Clementine enthalten.

58. QUINLAN, der aus dem Bereich des maschinellen Lernens kommt, hat ebenfalls sehr früh schon Baumalgorithmen entwickelt.

In den letzten Jahrzehnten wurden eine Vielzahl von Algorithmen entwickelt, deren Darstellung eher zur Verwirrung denn zur Klärung beitragen würde. Ein Vergleich von über 30 älteren und neueren Algorithmen findet sich bei LIM und LOH (2000). Interessanterweise spielt in der Literatur der letzten Jahre vor allem der CART-, aber auch der QUEST-Algorithmus eine große Rolle, wohingegen es scheint, dass die CHAID-basierten Algorithmen unverdienterweise eher in den Hintergrund getreten sind. Unabhängig von der statistischen „Leistungsfähigkeit" segmentieren sie mehr als zwei Unterknoten.

Diese Sichtweise mag zugegebenermaßen eher eine soziologische denn eine statistisch-mathematische sein: im Mittelpunkt der Arbeit stehen jedoch weniger statistische Algorithmen als vielmehr ihr soziologischer und damit inhaltlicher Informationsgehalt. Zu ersterem Thema gibt es ausreichende Literatur, die die entsprechenden Verfahren ausführlich untersuchen.

Um das vorliegende Klassifikationsproblem aus unterschiedlichen Blickwinkeln zu beleuchten, wurde auf das Programm Answertree in der Version 3.1 von SPSS zurückgegriffen. Es enthält zwei Binärbaumverfahren (CART und QUEST) und eine Methode, mehr als zwei Unterknoten zu bilden (CHAID und eine verbesserte Variante, EXHAUSTIVE CHAID).

Rein rechnerisch wäre es auch kein Problem, weitere Verfahren hinzuzuziehen - Ziel der Arbeit ist jedoch der Vergleich der Entscheidungsbaumalgorithmen miteinander - und mit anderen statistischen Verfahren. Da zu erwarten ist, dass jede Methode ein wenig anders klassifiziert und zu etwas anderen Ergebnissen kommt, würde es den Leser eher verwirren als zur Klarheit beitragen.

Die Entscheidung für oder gegen bestimmte Algorithmen ist zum einen abhängig von der eingesetzten Software, zum anderen vom er-

hofften Erkenntnisgewinn. Der Einsatz der o. g. Algorithmen in dieser Arbeit ist keine inhaltliche Ablehnung von anderen Verfahren - es ist die Konzentration auf in der Literatur wesentliche, wenige, aber überschaubare Algorithmen.

TABELLE 9 ALLGEMEINE KENNZEICHEN AUSGEWÄHLTER BAUMALGORITHMEN (VGL. WILKINSON (1992), SPSS (2001A: 185FF.))

Verfahren	Splits	Statistische Verfahren	Zielvariable	Besonderheiten
CHAID[a] KASS (1980)	2 und mehr	Chi-Quadrat (nominal) likelihood-ratio (ordinal) F-Test (metrisch)	nominal ordinal metrisch	
Exhaustive Chaid BIGGS, DE VILLE und SUEN (1991)	2 und mehr	Chi-Quadrat (nominal) likelihood-ratio (ordinal) F-Test (metrisch)	nominal ordinal metrisch	
C&RT BREIMAN, FRIEDMAN, OLSHEN, STONE (1984)	2	Gini, Twoing (nominal) ordered Twoing (ordinal) Least-Square-Deviation (metrisch)	nominal ordinal metrisch	Pruning A prioris Ersatzprädiktoren
QUEST LOH und SHISH (1997)	2	Chi-Quadrat (nominal) F-Test (ordinal, metrisch)	nominal	Pruning A prioris Ersatzprädiktoren

a. SPSS weist darauf hin, dass CHAID ursprünglich nur für die Untersuchung nominaler Merkmale von KASS (1980) entwickelt wurde. Der in Answertree enthaltene Algorithmus kann jedoch mit allen Skalenniveaus eingesetzt werden (vgl. SPSS (o. J. 2)).

Die eingesetzten Verfahren orientierten sich am Skalenniveau der zu untersuchenden Zielvariablen: bei nominalem Niveau (ist also eine Rangfolge oder Quantifizierung der Ausprägungen („a ist doppelt so

hoch wie b“) nicht möglich) wird der **Chi-Quadrat-Test** herangezogen.

Der **Likelihood-Wert** wird für ordinale Daten, für die eine Rangfolge, aber die Differenz der Abstände unbekannt sind, wird in Form eines Chi-Quadrat-Werts herangezogen. Ähnlich wie der F-Test bei der Varianzanalyse können einzelne Effekte betrachtet werden (vgl. BROSIUS (1989: 247)). Die einzelnen Werte werden hier auf der 10er-Basis logarithmiert - dadurch entstehen sehr große Wertebereiche bzw. Differenzen zwischen den beobachteten Werten, was eine gute Unterscheidung der Ausprägungen ermöglicht (zu einer ausführlichen Darstellung des Logarithmus vgl. PAMPEL (2000: 74ff.)). Dies ist natürlich nur mit ordinalen (oder metrischen) Daten sinnvoll, da Nominaldaten keine Rangfolge besitzen.

Am Beispiel der PC-Nutzung (abhängig) und den unabhängigen Variablen Alter und Bildung wird dies deutlich:

ABBILDUNG 46 PC-Nutzung: Likelihood-Wert bei unabhängigen Variablen Alter und Schulbildung

Informationen zur Modellanpassung

Modell	-2 Log-Likelihood	Chi-Quadrat	Freiheitsgrade	Signifikanz
Nur konstanter Term	1502,338			
Endgültig	605,331	897,007	78	,000

Likelihood heißt einfach übersetzt „Wahrscheinlichkeit“. Für multivariate Verfahren wie die logistische Regression gibt Likelihood die Wahrscheinlichkeit an, dass das errechnete Modell realitätsgerechter ist als das Ausgangsmodell ohne Variablen. Je mehr sich Ausgangsmodell und Zielmodell unterscheiden, umso realistischer ist es. Der Begriff der „Größe“ ist allerdings sehr relativ und eignet sich weni-

ger für den Vergleich von Modellen. Aus diesem Grund wird Likelihood logarithmiert. Es entsteht der sog. „-2 log likelihood-Wert" (=-2LL). BALTES-GÖTZ (2004b: 30) bemerkt:

> „Je besser ein Modell zu den Daten paßt, desto höher wird seine [sic!] Likelihood, und desto kleiner folglich die Größe -2LL, die somit als Fehlermaß aufgefaßt werden kann."

Somit erfolgt eine gewisse Standardisierung, die einen besseren Vergleich ermöglicht. In diesem Beispiel ist der Wert 605 deutlich kleiner als 1502 - somit haben die beiden unabhängigen Variablen Alter und Bildung einen deutlichen Einfluß auf die PC-Nutzung.

Die Differenz zwischen Ausgangs- und Zielmodell wird als Chi-Quadrat-Wert ausgegeben - und folgt der gleichen Logik wie im nominalen Fall: je mehr sich der Chi Quadrat-Wert von 0 unterscheidet, desto mehr unterscheiden sich die erwarteten von den beobachteten Häufigkeiten.

Ein ausführliches Beispiel findet sich bei der Darstellung der logistischen Regression, wo auch alle Werte erläutert werden.

Bei metrischen Variablen wird der **F-Test** herangezogen. Der F-Test basiert auf dem Mittelwert und der Standardabweichung und ist daher nur für intervall- und ratioskalierte Variablen geeignet (die Ausprägungen einer Variablen lassen sich ordnen und die Abstände können in ein Verhältnis zueinander gesetzt werden[59]). Der F-Wert untersucht die Streuung innerhalb und zwischen empirisch untersuchten Gruppen:

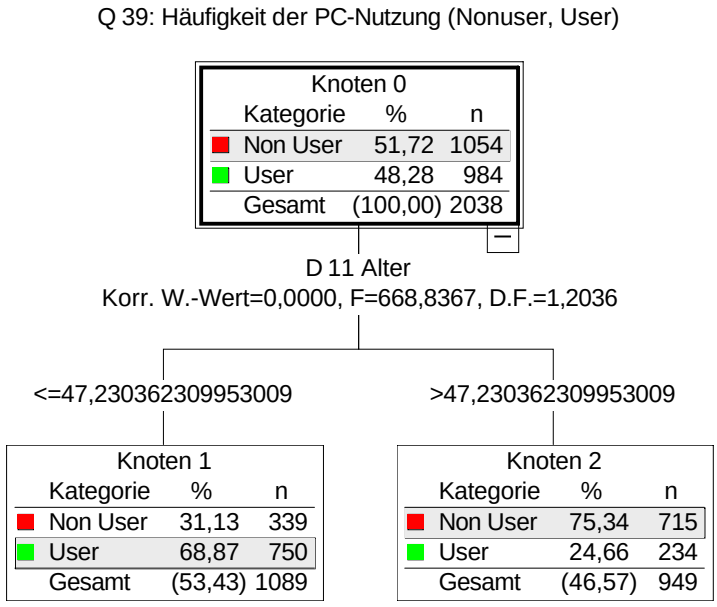
> „Ist die Streuung innerhalb der Gruppen gleich null, während gleichzeitig eine große Streuung zwischen den Gruppen vorliegt, ist dies

59. Am Beispiel „Lebensalter" läßt sich dies sehr gut zeigen: ein Dreißigjähriger ist nicht nur älter als ein Fünfzehnjähriger, sondern er ist auch doppelt so alt. Bei ordinalem Skalenniveau (z. B. Bildungsabschluss) läßt sich zwar eine Rangfolge herstellen, aber kein Verhältnis berechnen. Die Mittlere Reife ist ein höherer Bildungsabschluss als der Hauptschulabschluss, er ist aber nicht „doppelt" oder „vierfach" so „gut".

> gleichbedeutend damit, dass die einzelnen Gruppen sehr unterschiedliche Mittelwerte aufweisen, innerhalb der Gruppen jedoch alle Werte gleich sind." (BROSIUS (1998: 484))

Neben dem nachfolgenden Beispiel sei auf die umfangreiche Literatur (z. B. NORUSIS (1998: 292)) zu diesem Thema verwiesen.

ABBILDUNG 47 QUEST-Entscheidungsbaum: PC-Nutzung nach Alter (N = 2038)



Bei rund 47 Jahren wird der Baum aufgesplittet: links sind rund 69 %, rechts knapp ein Viertel der PC-User enthalten. Der korrigierte W-Wert gibt die Signifikanz für dieses Ergebnis an - in diesem Falle höchstsignifikant. Der F-Wert spiegelt mit rund 689 die Unterschiede des Mittelwertes wieder. Die Frage lautet hier: unterscheiden sich PC-Nutzer hinsichtlich des Alters deutlich von Nichtnutzern? - Nach dem Ausschnitt des obigen Entscheidungsbaums tun sie das, was auch recht anschaulich durch die Prozentanteile wiedergegeben wird: war im Wurzelknoten 0 das Verhältnis der Nichtnutzer zu den Nutzern 52 : 48, verstärkt sich dies in Knoten 2 (75 : 25) bzw. schwächt sich in Knoten 1 (31 : 69) deutlich ab.

Dieses Beispiel ist das Ziel eines jeden Segmentierungsverfahrens (egal ob Cluster-, Entscheidungsbaumanalyse, etc.): möglichst homogene Untergruppen zu erhalten, die sich maximal kontrastieren. In diesem Fall wird der F-Wert hochsignifikant sein. Im entgegengesetzten Fall, wenn also die Mittelwerte für die einzelnen Gruppen sehr nahe beieinanderliegen, wird sich die Übertragung auf die Grundgesamtheit häufig nicht ziehen lassen. Alle Algorithmen können mit jedem Skalenniveau rechnen - nur QUEST setzt eine nominale (auch dichotome) Zielvariable voraus.

CHAID und **EXHAUSTIVE CHAID** sind Algorithmen, die auf dem Chi-Quadrat-Wert beruhen und zwei oder mehr Baumsplits pro Ebene durchführen. Vorteil ist hierbei möglicherweise ein höherer Informationsgewinn. Ebenso wie C&RT stellen CHAID und EXHAUSTIVE CHAID keine Anforderungen an das Skalenniveau.

CHAID untersucht die Signifikanz der Vorhersagevariablen und bewertet alle Ausprägungen. Sind die Werte homogen, werden sie zu einer Gruppe zusammengefaßt und ein Baumsplit errechnet, wobei zwei oder mehr Verzweigungen vorgenommen werden können. Bei nominalen Daten verwendet CHAID für die Baumsplits den Chi-Quadrat-Wert, bei Ordinaldaten den Likelihood-Wert, bei metrischen Daten den F-Wert (vgl. SPSS (2001a: 188)).

Der **CHAID**-Algorithmus (Chi-Squared Automatic Interaction Detector) wurde bereits 1980 von KASS entwickelt. Er sollte heute eher bei langsameren Rechnern eingesetzt werden, da der Exhaustive Chaid-Algorithmus in einigen Fällen exakter arbeitet (s. u.), jedoch mehr Rechenzeit in Anspruch nimmt.

EXHAUSTIVE CHAID ist eine Weiterentwicklung des CHAID-Algorithmus von BIGGS, DE VILLE und SUEN (1991). Ebenfalls wie CHAID können zwei oder mehr Entscheidungsbaumäste modelliert werden.

EXHAUSTIVE CHAID versucht jedoch, einen Nachteil des CHAID-Algorithmus zu verbessern:

> „In particular, sometimes CHAID may not find the optimal split for a variable, since it stops merging categories as soon as it finds that all remaining categories are statistically different. Exhaustive CHAID remdies this by continuing to merge categories of the predictor variable until only two supercategories are left. It then examines the series of merges for the predictor and finds the set of categories that gives the strongest association with the target variable, and computes an adjusted p value for this association." (SPSS (2001a: 188f.))

Die grundsätzliche Idee, die hinter den beiden Verfahren steht, ist jedoch gleich und so ist es auch möglich, dass beide Verfahren - je nach Datensatz - zu den gleichen Ergebnissen kommen können. Bei EXHAUSTIVE CHAID liegen die gleichen Verfahren (Chi-Quadrat, Likelihood, F-Wert) und die gleiche Vorgehensweise beim Baumaufbau wie CHAID zugrunde (vgl. BALTES-GÖTZ (2004a: 28ff.)). Der Unterschied liegt darin, dass EXHAUSTIVE CHAID an dem Punkt, wo statistisch keine weiteren Unterschiede (von CHAID) gefunden werden, versucht, die Kategorien zu zwei Gruppen zusammenzufassen, um erneut zu prüfen, ob sich statistische Unterschiede finden lassen. Das CHAID-Verfahren wendet diesen Schritt nicht an. Je nach Datensatz kommen die beiden Algorithmen entweder zu den gleichen oder auch zu unterschiedlichen Ergebnissen. Im letzteren Fall ist, bei inhaltlich besserer Interpretierbarkeit, der EXHAUSTIVE-CHAID-Algorithmus CHAID vorzuziehen.

Im Gegensatz hierzu bilden **CART-** und **QUEST-**Bäume immer zwei Unterknoten pro Ebene, was evtl. zu einem Informationsverlust führen könnte. Dafür sind die beiden Algorithmen in der Lage, komplexe Bäume zu kürzen („pruning", beschneiden), ohne den Anteil der Fehlklassifikation wesentlich zu verschlechtern. Ein weiterer Vorteil liegt darin, dass bei fehlenden Werten von Variablen („missing values") Ersatzvariablen herangezogen werden. Wenn es beispielsweise bei der Angabe zum Haushaltsnettoeinkommen viele fehlende Wer-

te gibt, wird die Analyse nicht abgebrochen, sondern die nächstwichtige Variable (anhand des höchsten statistischen Kennwerts) herangezogen.

C&RT (Classification and Regression Trees), 1984 von BREIMAN, FRIEDMAN, OLSHEN und STONE entwickelt, untersuchen alle Variablen rekursiv auf jeder Stufe neu. Dies kann zu einem Baummodell mit sehr vielen Ebenen führen (im Gegensatz dazu können mit CHAID und EXHAUSTIVE CHAID sehr „breite" Bäume mit mehr als zwei Unterknoten modelliert werden). Dies trifft vor allem dann zu, wenn die Variablen viele Ausprägungen haben (z. B. Alterskategorien in 5-Jahres-Schritten). Während QUEST und C&RT immer zwei Gruppen pro Ebene bilden und diese Gruppen evtl. immer weiter pro Ebene zerlegen, kann es bei (EXHAUSTIVE) CHAID der Fall sein, dass auf einer Ebene - wie oben - z. B. zehn Alterskategorien zusammengefasst werden, wodurch die Gruppengröße der segmentierten Gruppen sehr klein werden kann. Dies ist aber eine explizite Abbruchregel, denn wenn die Gruppengröße unter 30 Befragte sinkt, ist es schwierig, signifikante Ergebnisse zu erhalten). Dies spricht für binäre Baumalgorithmen, denn wie WILKINSON (1992: 3) überzeugend darstellt, können binäre sehr einfach in nichtbinäre Entscheidungsbäume konvertiert werden: während nichtbinäre Algorithmen mehrere Ausprägungen auf einer Ebene trennen, geschieht dies bei binären Verfahren auf verschiedenen Ebenen.

QUEST (Quick, Unbiased, Efficient Statistical Tree) ist der neueste der vorgestellten Baumalgorithmen und wurde 1997 von LOH und SHISH entwickelt. Auch dieser Binärbaum bietet die gleichen Vorteile mit Ersatzprädiktoren, A-prioris und Angabe von Fehlklassifizierungskosten wie CART. Im Unterschied zu allen anderen vorgestellten Algorithmen setzt QUEST eine nominale Zielvariable voraus.

BALTES-GÖTZ (2004a: 57) benennt zwei „potentielle Probleme“ des C&RT-Verfahrens, die QUEST vermeidet: einen hohen „Rechenaufwand“ und die Neigung, „Prädiktoren mit vielen Ausprägungen“ zu bevorzugen. Das letzte Problem läßt sich durch einen Ergebnisvergleich der verschiedenen Algorithmen kontrollieren, das erste Problem ist spürbar, sollte aber für die Auswertungen dieser Arbeit keine Rolle spielen. Die statistischen Regeln, nach denen QUEST die Bäume aufbaut, finden sich z. B. bei BALTES-GÖTZ (2004a: 57f.).

Bei QUEST werden für nominale Prädiktoren der Chi-Quadrat-Test, für ordinale und metrische Variablen der F-Test der einfaktoriellen Varianzanalyse herangezogen (vgl. BALTES-GÖTZ (2004a: 57)).

Im Gegensatz zu den anderen Algorithmen geht QUEST in zwei Schritten vor: im ersten Schritt wird ein Prädiktor ausgewählt und - wie nachfolgend beschrieben - untersucht. Erst dann wird die Stichprobe gesplittet. Die Analyse läuft in folgenden Schritten ab (vgl. BALTES-GÖTZ (2004a: 59ff)):

1. Für jeden Prädiktor wird geprüft, welchen Informationsgehalt er hinsichtlich der Zielvariablen besitzt. Ausgewählt wird diejenige Variable mit den höchsten Werten (anhand Chi-Quadrat bzw. F-Wert). Wird kein signifikanter Prädiktor (voreingestellt, aber änderbar: ≤ 0.05) gefunden, werden für mindestens ordinalskalierte Prädiktoren Varianzunterschiede nach dem Levene-Test[60] untersucht. Findet sich auch hier kein signifikanter Wert, wird der größte F- bzw. Chi-Quadrat-Wert herangezogen.
2. Wenn mehr als zwei Kategorien einer unabhängigen Variablen vorhanden sind, werden diese zu zwei „Superklassen“ zusammengefaßt. Hierzu wird ein Clusterverfahren herangezogen, das die beiden am weitesten auseinanderliegenden Mittelwerte miteinander vergleicht.
3. Für die Trennung eines metrischen Prädiktors wird das Konzept der Diskriminanzanalyse herangezogen. Eine Normalverteilung wird unterstellt, die Mittelwerte bzw. Varianzen werden geschätzt.
4. Sind die Verteilungen bekannt, können auch a prioris herangezogen werden

60. vgl. NORUSIS (1998: 240ff.). Der LEVENE-Test ist ein Test auf Varianzhomogenität. Die Fragestellung lautet: „Unterscheiden sich die zu untersuchenden Gruppen hinsichtlich der Varianz signifikant voneinander oder kommt das Ergebnis zufällig zustande?“ Die Ergebnisse werden als Signifikanzniveau ausgegeben. Werte von $p > 0.05$ sind nicht signifikant.

Von den errechneten statistischen Werten unterscheidet sich QUEST kaum von den CHAID-basierten Verfahren, kann aber zu einem realitätsgerechteren Baum führen:

> „Die Variablenauswahl und die Trennungspunktauswahl erfolgen bei dieser Methode getrennt voneinander. Die univariate Trennung in QUEST führt eine annähernd unvoreingenommene Variablenauswahl durch. Wenn alle Prediktorvariablen in Bezug auf die Zielvariable gleichermaßen informativ sind, wählt QUEST also mit gleicher Wahrscheinlichkeit eine beliebige Prediktorvariable aus." (SPSS (2001b: 222))

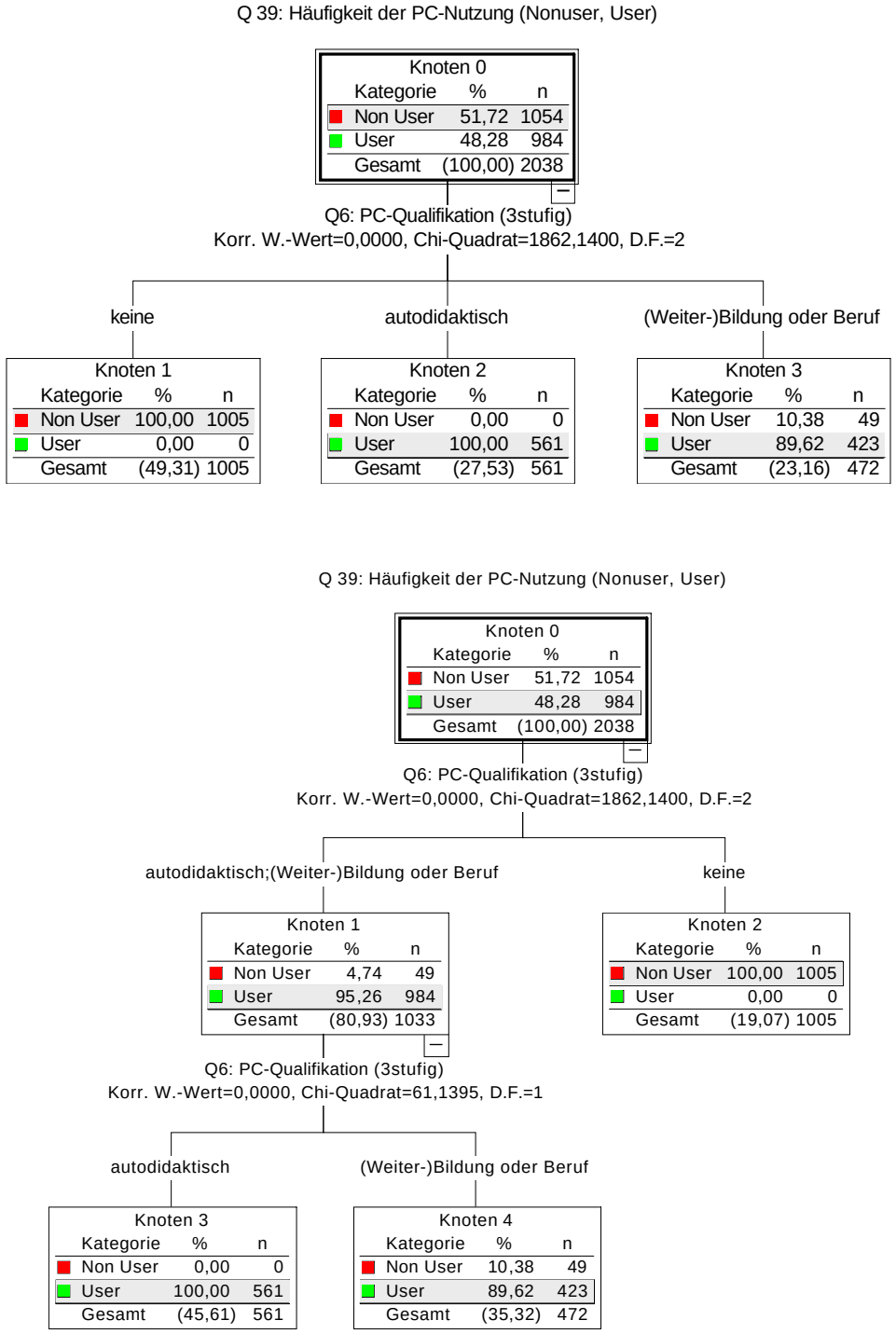
Deutlich wird, dass es nicht einen „richtigen" Algorithmus gibt, sondern die Wahl des Verfahrens eher vom Ergebnis abhängt: CHAID und EXHAUSTIVE CHAID segmentieren eher breitere, C&RT und QUEST eher tiefere Entscheidungsbäume. Bei metrischen Variablen mit vielen Ausprägungen (z B. Alter) könnte der (EXHAUSTIVE) CHAID-Algorithmus viele Untergruppen, C&RT bzw. QUEST immer zwei Unterknoten bilden. Möglicherweise führen letztere Verfahren zu einer zu starken Informationsreduktion, es ist aber vielleicht sinnvoll, bei vielen Ausprägungen „Grundsegmentierungen" zu modellieren. Bei kleineren Datensätzen mit geringen Fallzahlen könnten bei QUEST und C&RT eher Abbruchregeln und somit Informationsverlust eintreten, da die Bäume eher schmal aber tief werden. Diese Beispiele illustrieren, dass für jede Fragestellung unterschiedliche Entscheidungsbäume in Frage kommen können - abhängig von den zu untersuchenden Daten. Möchte man z. B. das Lebensalter in Jahren untersuchen und vorab keine Recodierungen vornehmen, so ist zu erwarten, dass bei (EXHAUSTIVE) CHAID sehr „breite" Bäume klassifiziert werden, die schnell zu einer kleinen Fallzahl für die weiteren Untersuchungsebenen (und damit zum Abbruch der Analyse) führen können. Wenn man keine theoretischen Vorgaben hat, wäre es unter Umständen sinnvoller, C&RT oder QUEST heranzuziehen. Andererseits kann die Erzwingung einer Zwei-Knoten-Lösung pro Ebene zu einem Informationsverlust führen. Deshalb ist es von Fragestellung zu

Fragestellung sinnvoll, den passenden Algorithmus anzuwenden - die Fehlklassifikationen werden sich kaum unterscheiden.

Aus diesem Grund werden drei der vier Algorithmen - EXHAUSTIVE CHAID, CART und QUEST - in dieser Arbeit herangezogen, um das Problem der PC-Nutzung zu untersuchen. Wahrscheinlich werden die Ergebnisse nicht identisch sein und somit wird es für diese Fragestellung vielleicht einen Algorithmus geben, der den größten Informationsgewinn erzielt. Die Schlußfolgerung, dass alle anderen Algorithmen zu verwerfen sind, ist jedoch nicht gegeben. Für andere Fragestellungen könnten sie zu besseren Lösungen führen - abhängig von der Struktur der Daten, den Skalenniveaus und der Anzahl der Ausprägungen.

Die Transformation von nichtbinären in Binärbäumen soll an einem kleinen Beispiel anhand des EXHAUSTIVE CHAID- und des QUEST-Algorithmus, die beide auf der Chi-Quadrat-Statistik beruhen, verdeutlicht werden. Als abhängige Variable wurde die dichotomisierte PC-Nutzung, als unabhängige Variable die Art der PC-Qualifikation (keine, durch Bildungseinrichtungen/Beruf, autodidaktisch) gewählt.

ABBILDUNG 48 Umwandlung von nichtbinären Bäumen in Binärbäumen anhand des EXHAUSTIVE CHAID (oben) und des QUEST-Algorithmus (unten)



Der Informationsgehalt ist gleich hoch, die Gruppen werden ähnlich gut anhand der Chi-Quadrat-Statistik segmentiert. Die Chi-Quadrat-

Werte weichen für alle Baumsplits deutlich von 0 ab und sind mit 1862.14 identisch. Deutlich wird aber auch, dass durch die ausführliche Trennung bei EXHAUSTIVE CHAID (oberer Baum) in drei Unterkategorien möglicherweise Informationen verlorengehen, da schneller Stopkriterien erreicht werden. Die zusätzliche Information, die hier QUEST liefert, bezieht sich darauf, dass sich die Nutzergruppen ähnlicher sind als die Nichtnutzergruppen (es hätten ja auch die Nichtnutzer und diejenigen mit Weiterbildungs-, Berufskenntnissen zu einer Gruppe zusammengefaßt werden können). Das ist in diesem Falle ein eher alltägliches Ergebnis, soll aber die möglichen Gefahren von nichtbinären Verfahren verdeutlichen.

Weiter wird die Rolle der 49 aktuellen Nichtnutzer, die aber schon irgendwann einmal mit dem PC in Berührung kamen (in Form von (Weiter-)Bildung oder Beruf) ‚verdeutlicht: durch die PC-Kenntnisse werden sie in diesem Fall falsch zugeordnet.

Der Vorteil bei QUEST (und auch C&RT) kann darin liegen, sehr bedeutsame Variablen zu identifizieren: wird bei einer größeren Anzahl unabhängiger Variablen (z. B. Alter, Bildung, berufliche Stellung, Einkommen, Musikgeschmack, Lesegewohnheiten, ...) das Alter als unabhängige Variable auf der ersten und der zweiten Baumebene als Splitkategorie - und damit als wichtigste Variable ermittelt - zeigt dass die Bedeutung des Alters, bezogen auf die Zielvariable nicht nur auf der ersten, sondern auch auf der zweiten Ebene den größten Einfluß ausübt. Dieser Vorteil kann sich allerdings auch zum Nachteil verkehren, denn es existieren bei jeder Analyse auch sog. Abbruchregeln (z. B. Größe der Gruppen, Zahl der Ebenen, etc.). Sind die Datensätze einer untersuchten Datei nicht sehr umfangreich, kann es passieren, dass nur eine unabhängige Variable auf zwei oder drei Ebenen als Splitkriterium errechnet wird - ein Ergebnis, das nicht sehr hilfreich ist.Binäre Splits können auch zur Informationsreduktion führen, denn

C&RT fasst - wie QUEST - alle Variablen immer zu zwei Unterknoten zusammen - auch wenn andere Verfahren wie EXHAUSTIVE CHAID mehrere Splits finden würden (vgl. hierzu auch SPSS (2001a: 189)).

C&RT verwendet für die Baumsplits Gini bzw. twoing, für Ordinaldaten ordered twoing und für metrische Daten Least Square Deviation (Kleinste Quadrate).

„Twoing" (das man ins Deutsche mit „in zwei gleich große Teile aufteilen" übersetzen könnte) ist eine Splitregel, die versucht, einen Knoten in etwa zwei ähnlich große Splits aufzuteilen.[61] Die Gini-Splitregel versucht, die größte(n) Gruppe(n) der betrachteten Daten zu isolieren (vgl. LAST (2002c: 4)), vor allem bei dichotomen Variablen ist Gini besonders geeignet (vgl. LAST (2002c: 5).

Beispiel: Die Variable „Familienstand" hat folgende Ausprägungen: ledig 30 %, verheiratet 40 % geschieden 10 %, verwitwet: 20 %. Nach der twoing-Regel würde die Gruppe der Ledigen und Verwitweten (= 50 %) und die Gruppe der Verheirateten und der Geschiedenen (= 50 %) zusammengefaßt, nach der Gini-Splitregel die Verheirateten, die die größte Gruppe darstellen (40 %) von den anderen Gruppen isoliert werden.

Der Gini-Koeffizient wird insbesondere in den Wirtschaftswissenschaften zur Untersuchung von Konzentrationen herangezogen. Hierbei werden kumulierte Häufigkeitstabellen der zu untersuchenden Merkmale gebildet (z. B. Anteil der Unternehmen - Anteil des Gesamtumsatzes, Anteil der Bevölkerung - Anteil der Einkommen, ...). Dieses

61. Dahinter steht die Idee eines möglichst hohen Informationsgehaltes durch die Einbeziehung möglichst vieler aussagekräftiger unabhängiger Variablen. Wenn auf einer Seite des Astes nur 10 %, auf der anderen Seite 90 % segmentiert werden, können vielleicht keine weiteren Segmentierungen mehr vorgenommen werden, weil die Gruppengröße unter 30 sinkt.

Thema wurde bereits ausführlich bei der Beschreibung der Gewinnübersicht behandelt (vgl. Abbildung 45 auf Seite 128).

Der CART-Algorithmus macht sich dieses Maß insofern zunutze, dass er nach Konzentrationen sucht. In der Fachliteratur wird dies häufig als „Verbesserung der Inhomogenität" bezeichnet, was ein sprachlich etwas irreführender Begriff ist, da man durch Kürzen von „Verbesserung" und „In-" (von Inhomogenität) nur noch „Homogenität" erhält. Ziel ist es also, homogenere (Unter-)Gruppen zu finden. Möglicherweise verbindet der eine oder andere Leser mit dem Wort der Konzentration auch etwas Negatives - bezogen zum Beispiel auf Unternehmenskonzentration. Dies ist aber genau das Ziel von Entscheidungsbäumen: Gruppen mit gleichen Merkmalen (z. B. Alter) zu „konzentrieren" und in Gruppen zusammenfassen. Da bei Entscheidungsbäumen nicht eine, sondern mehrere Gruppen betrachtet werden, liegt der Gini-Wert z. B. für zwei Gruppen zwischen 0 (keine Konzentration) und 0.5 (vollständige Konzentration), für drei Gruppen zwischen 0 (keine Konzentration) und 0.33 (vollständige Konzentration), usw..

3.2.1 A-priori-Wahrscheinlichkeiten (nur CART und QUEST)

CART und QUEST bieten die Möglichkeit, sog. „A priori"-Wahrscheinlichkeiten für kategoriale Zielvariablen zur Verfügung:

> „A-priori-Wahrscheinlichkeiten sind numerische Werte, die die Fehlklassifizierungsraten für Kategorien der Zielvariablen beeinflussen. Sie geben den Anteil der Fälle an, die bereits vor der Analyse mit jeder Kategorie der Zielvariablen verknüpft waren." (SPSS (2001b: 233))

„A prioris" sind somit Anweisungen, ganz bestimmte Variablenausprägungen mit bestimmten Wahrscheinlichkeiten zu erwarten:

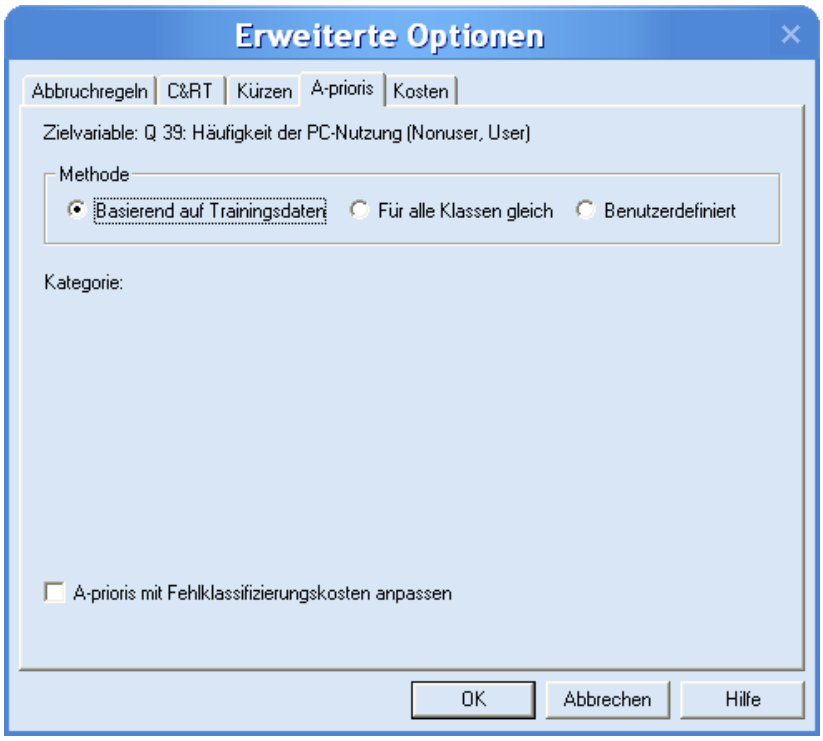
> „Durch die festgelegten A-priori-Wahrscheinlichkeiten wird AnswerTree angewiesen, Fälle mit der jeweils zugeordneten Wahrscheinlichkeit zu erwarten." (SPSS (2001b: 180))

Ziel ist es, durch eine Gewichtung realitätsgerechtere Bäume zu generieren und Fehlklassifikationen zu vermindern. Wenn bei einer Fragestellung von vornherein Informationen vorliegen, die z. B. eine Gleichverteilung unterstellen, dann kann dies explizit angegeben werden.

Ein mögliches Beispiel wäre eine Landtagswahl in einem westdeutschen Bundesland. In der Regel kann davon ausgegangen werden, dass es zwei größere (CDU/CSU bzw. SPD) und zwei kleinere (FDP, GRÜNE) Parteien gibt, die Chancen haben, über die 5 %-Hürde zu kommen. Diese Informationen könnten hier eingebracht werden: grundsätzlich würde davon ausgegangen, dass alle Parteien alle die gleiche Wahrscheinlichkeit haben, stärkste Kraft im jeweiligen Landesparlament zu werden. Dies ist jedoch nicht der Fall, da es weder FDP- noch GRÜNEN-Ministerpräsidenten gibt. Diese Informationen können also verwendet werden, um eine Art Gewichtung vorzunehmen: kleinere Parteien werden geringer, größere Parteien höher gewichtet.

Hierzu gibt es in Answertree verschiedene Möglichkeiten:

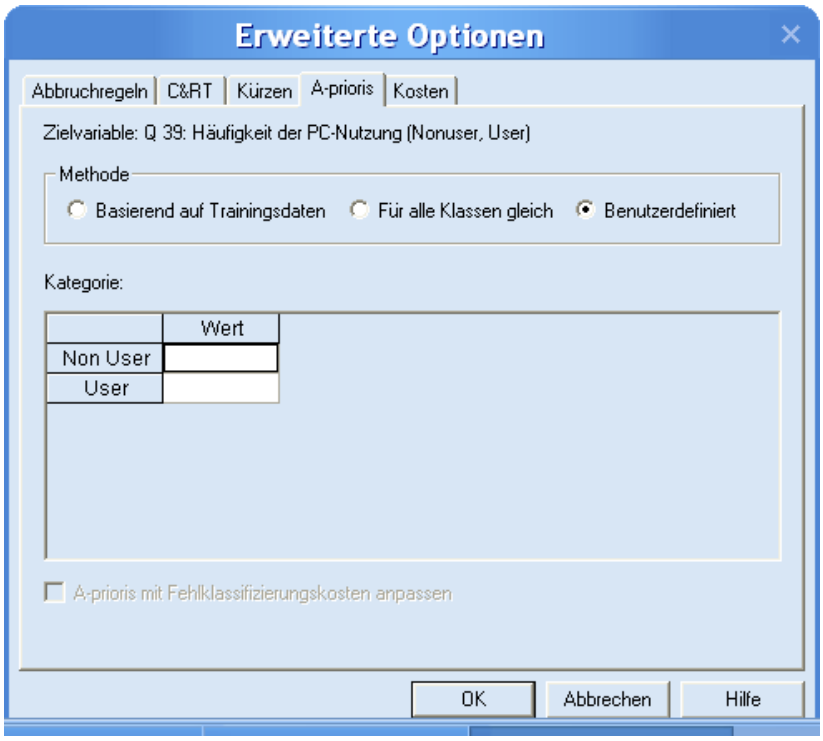
ABBILDUNG 49 CART-Algorithmus: erweiterte Optionen - A prioris (Grundeinstellungen)



Die voreingestellte Methode basiert auf dem jeweils eingesetzten Datensatz (basierend auf Trainingsdaten) - am Beispiel der PC-Nutzer 52 % Nichtnutzer vs. 48 % Nutzer.[62] Alternativ kann er für alle Klassen gleichgewichtet vorgenommen oder auch als benutzerdefinierter Erwartungswert angegeben werden:

62. Als unabhängige Variablen wurden Alter (in Lebensjahren) und Bildungsabschluss herangezogen. Die einführenden Beispiele basieren - der besseren Verständlichkeit halber - zu Beginn auf zwei Ebenen.

ABBILDUNG 50 CART-Algorithmus: erweiterte Optionen - A prioris (erweiterte Einstellungen)



„A prioris" führen nicht unbedingt zu besseren Bäumen - vor allem kann auch eine Reduzierung der Fehlklassifikation zu Lasten der Verbesserung des Baums gehen:

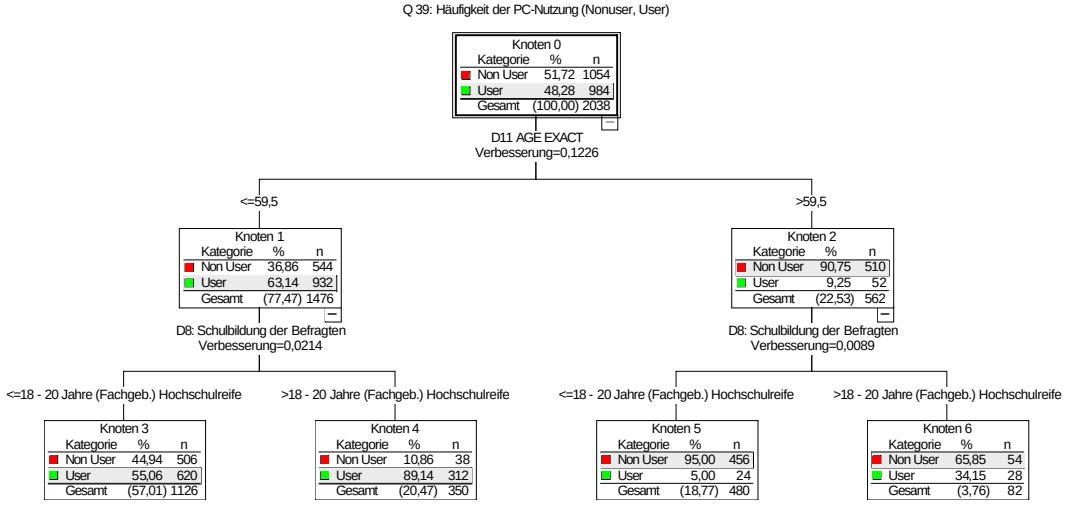
ABBILDUNG 51 CART-Algorithmus: Prediktoren für PC-Nutzung

Der Baum ohne Anpassung hat eine Fehlklassifikation von 28.4 % mit der oben dargestellten Verbesserung. Wenn nun größere Anstrengungen unternommen werden sollen, PC-Nutzer richtig zu klassifizieren (z. B. mit 0.6, Nichtnutzer folglich mit 0.4), ergeben sich folgende Werte:

ABBILDUNG 52 CART-Algorithmus: Prediktoren für PC-Nutzung (A priori Einstellung: 0.6 PC-User, 0.4 Non User)

Bei der Variable Alter findet eine weitere Verbesserung von 0.1165 auf 0.1226 statt - die Verbesserung der Schulbildung fällt leicht auf 0.0454. Die Fehlklassifikation liegt bei rund 23.8 % und sinkt somit um nahezu 5 %. Auch der Baum hat sich geringfügig verändert:

ABBILDUNG 53 CART: 2stufiger Entscheidungsbaum für PC-Nutzung (abhängig), Alter und Schulbildung (unabhängig) - A prioris: 0.6 : 0.4



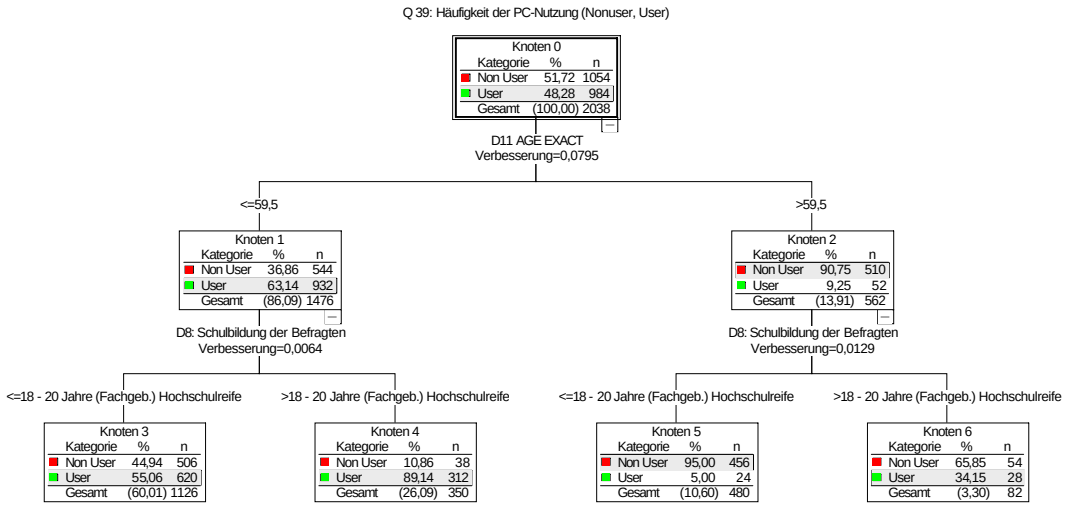
Wurde im ursprünglichen Beispiel die Alterskategorie bei 57.5 Jahren getrennt, so findet hier die Trennung bei rund 60 Jahren statt. Der Anteil der Nutzer im rechten Teil fällt durch die Anhebung der Altersgrenze von 12 % auf 9.25 %. Die Trennung des Schulabschlusses bleibt gleich, die Anteile der Nutzer sinken jedoch bei den Jüngeren durch die Altersverschiebung von rund 7 % bei den Jüngeren auf 5 % und von 39 % auf 34 % zugunsten der linken Aststruktur. Dieser Baum ist also - was Verbesserung und Fehlklassifikation (mit Ausnahme des Bildungsgrads, der leicht sinkt) angeht - etwas positiver zu bewerten.

Ein Erhöhen der Werte auf 0.8 (PC-Nutzer) und 0.2 (Nichtnutzer) führt zwar zu einer Fehlklassifikation von nur noch 13 %, dafür sinken die Verbesserungen drastisch.

ABBILDUNG 54 CART-Algorithmus: Prediktoren für PC-Nutzung (A priori Einstellung: 0.8 PC-User, 0.2 Non User)

Der Entscheidungsbaum liefert die exakten Werte und Prozentanteile wie der vorangegangene Baum - nur die Verbesserungswerte sind schlechter als oben:

ABBILDUNG 55 CART: 2stufiger Entscheidungsbaum für PC-Nutzung (abhängig), Alter und Schulbildung (unabhängig) - A prioris: 0.8 : 0.2)

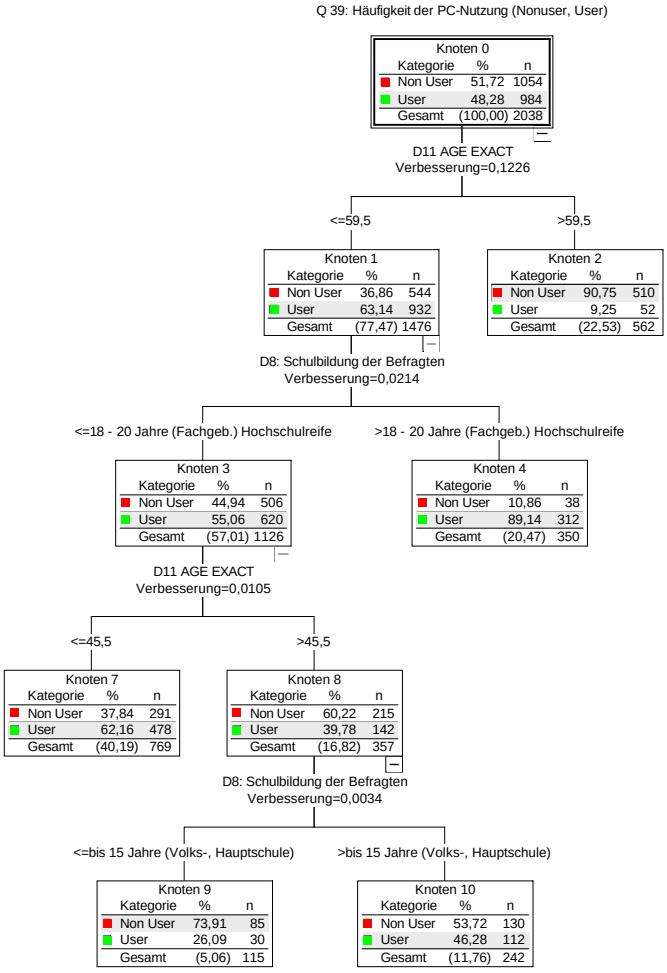


Dies läßt sich erklären, wenn man einen Blick auf den - gekürzten - Gesamtbaum wirft, der im folgenden Unterkapitel erläutert wird.

3.2.2 Pruning (nur CART und QUEST)

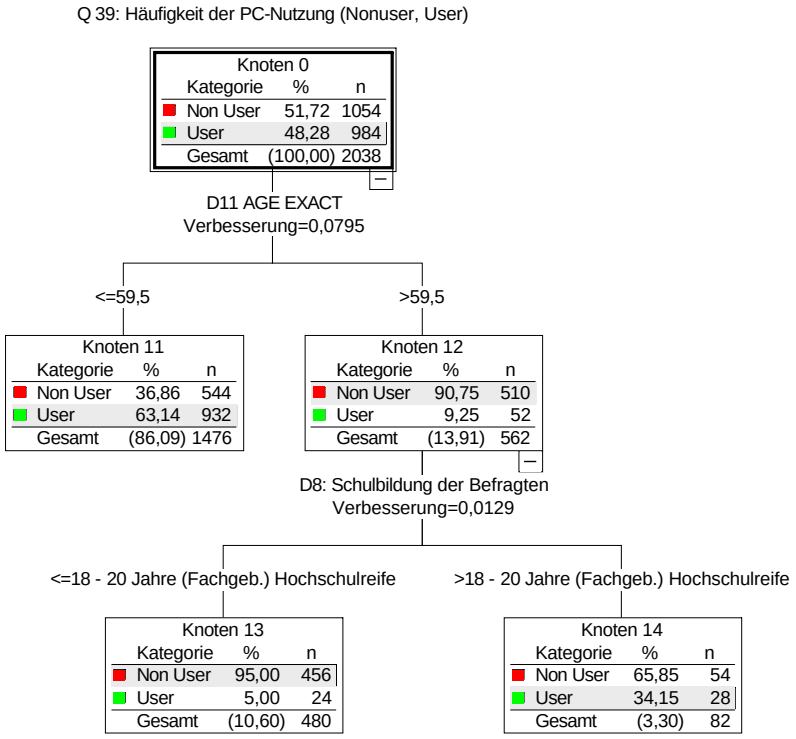
Binäre Entscheidungsbäume neigen dazu, sehr schnell unübersichtlich zu werden, da die Ausprägungen der unabhängigen Variablen dichotomisiert und auf unterschiedlichen Ebenen zur Trennung herangezogen werden können. Aus diesem Grund gibt es bei CART und QUEST die Möglichkeit des sog. „pruning“ (Beschneiden bzw. Kürzen) der Bäume. Äste, die kaum einen Beitrag zur Verbesserung bzw. Fehlklassifikation beitragen bzw. eine recht geringe Fallzahl besitzen, werden „beschnitten“. Die im vorangegangenen Unterkapitel beschriebene (0.6 Nutzer : 0.4 Nichtnutzer) Apriori-Lösung sieht - beschnitten - folgendermaßen aus:

ABBILDUNG 56 CART-Entscheidungsbaum (beschnitten) für PC-Nutzung (abhängig), Alter und Schulbildung (unabhängig) - A prioris: 0.6 : 0.4)



Im Gegensatz dazu liefert die Apriori-Lösung mit (0.8 Nutzer, 0.2 Nichtnutzer) folgende Lösung:

ABBILDUNG 57 CART-Entscheidungsbaum (beschnitten) für PC-Nutzung (abhängig), Alter und Schulbildung (unabhängig) - A prioris: 0.8 : 0.2)



Beide Bäume unterscheiden sich grundlegend: während beim oberen Baum der linke Ast mit den jüngeren Befragten weiter differenziert wird, wird beim unteren Beispiel der Ast der Älteren erweitert - obwohl der linke Ast von der Höhe der Fallzahlen (1476 Befragte) durchaus weiter gesplittet werden könnte. Die weitere - rechte - Aufsplittung bringt mengenmäßig (N = 52 User) kein Ergebnis, das inhaltlich besonders bedeutsam wäre.

Für die inhaltliche Bedeutung der Bäume kann das automatische Kürzen zu einer - zumindest soziologischen - verkürzten Darstellung der Variablen führen, sicherlich jedoch nicht zu einer mathematischen. Der Einsatz von „A prioris“ ist deshalb theoretisch gut zu überdenken und nicht automatisch durchzuführen.

3.2.3 Ersatzprädiktoren (nur CART und QUEST)

In einigen Fällen kann es bei Erhebungen zu Antwortverweigerungen kommen. Typische Beispiele sind die Frage nach der Höhe des Einkommens oder persönliche Fragen.

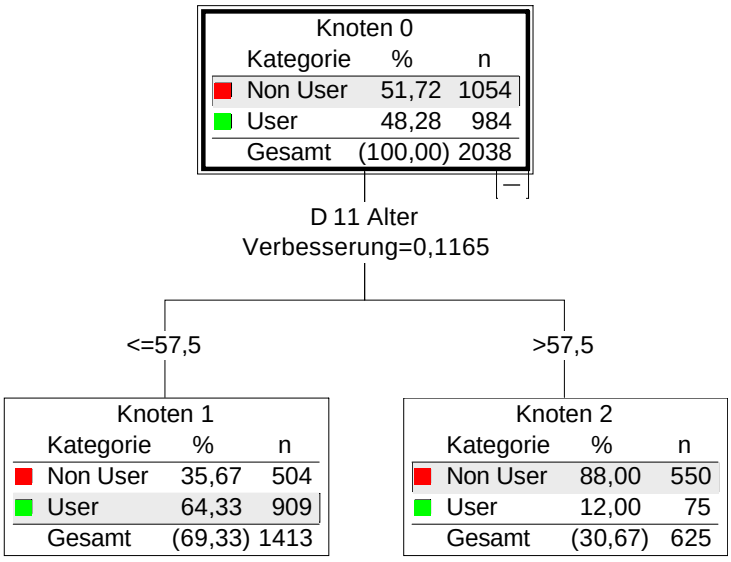
Die binären Algorithmen stellen hier die Auswahl von Ersatzprädiktoren zur Verfügung: wenn es viele Befragte mit fehlenden Werten gibt, so werden andere unabhängige Variablen herangezogen. Auch wenn dies nicht unbedingt die beste Lösung ist, wäre die Alternative ein Abbruch der Analyse bzw. ein Knoten mit fehlenden Werten - Alternativen, die gerade vermieden werden sollen, da sie keinerlei Informationsgehalt besitzen. Die Zahl der Ersatzprädiktoren kann voreingestellt werden - es ist auch möglich, keine anderen Ersatzvariablen heranzuziehen.

3.3 Die Interpretation von Entscheidungsbäumen - ein praktisches Beispiel

Der nachfolgende CART-Entscheidungsbaum führt anwendungsorientiert und praktisch anhand des gewählten Beispiels in das Denken mit dieser Methode ein. Aus didaktischen Gründen wurde auf den CART-Algorithmus zurückgegriffen, da die statistischen Werte und deren Berechnung noch nicht erläutert wurden - denn im Gegensatz zu EXHAUSTIVE CHAID und QUEST basiert CART nicht auf dem Chi-Quadrat-Wert und dem F-Test.

ABBILDUNG 58 CART: Einstufiger Entscheidungsbaum für PC-Nutzung, Alter und Bildung (N = 2038)

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)



In diesem Beispiel ist die abhängige Variable nominal-dichotom skaliert (PC-Nutzer: ja - nein). Anhand der unabhängigen Variablen Alter (metrisch, in Jahren) und Schulbildung (ordinal) wird untersucht, welche der beiden Variablen den größten Einfluss auf die abhängige Variable PC-Nutzung besitzt.[63]

Die Verbesserung der Inhomogenität durch das Alter liegt im Gesamtmodell bei 0.1165 (Sig. = .000). Eine mathematische Erläuterung am Beispiel des bekannten Iris-Datensatzes von FISHER findet sich z. B. bei BALTES-GÖTZ (2004a: 34ff).

Ziel eines Entscheidungsbaums ist es, den inhomogenen Wurzelknoten zu „verbessern", d. h. homogenere Unterknoten zu segmentieren. Dabei wird der („unreinere") Hauptknoten in Beziehung gesetzt zu den - hoffentlich - „reineren" Unterknoten. Der Hauptknoten im Beispiel enthält ca. 52 % Nichtnutzer und ca. 48 % Nutzer. Die darunter-

63. Für den Vergleich wurde auf Aprioris und Kürzen verzichtet, um den direkten Vergleich mit den CHAID-basierten Algorithmen zu ermöglichen.

liegenden Knoten sind - hinsichtlich den PC-Nutzer-Anteilen - homogener: links ist der Anteil der PC-Nutzer mit 64 % höher, rechts niedriger (12 %). Somit findet eine „Verbesserung der Inhomogenität", also eine „Konzentration" der Variablenausprägungen der abhängigen Variable statt. Um diese Konzentration zu messen, wird das Maß von Gini herangezogen.[64]

$$\textbf{Knoten 1: } 1 - 0.3567^2 - 0.6433^2 = 0.4590$$

$$\textbf{Knoten 2: } 1 - 0.88^2 - 0.12^2 = 0.2112.$$

Allgemein ausgedruckt:

$$\textbf{Knoten n: } 1 - \textbf{Anteil Nonuser}^2 - \textbf{Anteil User}^2.$$

Der Wert von .4590 für Knoten 1 (Verhältnis: 35 : 65) ist höher als für Knoten 2 (Verhältnis: 88 : 12). Somit sind die Gruppen in Knoten 1 inhomogener als in Knoten 2. Bei einem Verhältnis von 50 % : 50 % - also bei einer reinen Gleichverteilung - ergäbe sich für zwei Ausprägungen:

$$1 - 0.5^2 - 0.5^2 = 0.5.$$

Bei einem Wert von 0.5 herrscht völlige Inhomogenität (50 % Nichtnutzer stehen 50 % Nutzern gegenüber). Je mehr sich dieser Wert gegen 0 verschiebt, desto homogener und konzentrierter (hinsichtlich der abhängigen Variablen) wird der Entscheidungsbaum. Bei einem Wert von 0 würden alle PC-Nutzer bzw. Nichtnutzer korrekt im Entscheidungsbaum segmentiert.[65]

Die Ergebnisse für Knoten 1 und 2 (vgl. Abbildung 58 auf Seite 158) zeigen nur die „Konzentrationen" (Anteile der PC-(Nicht-)Nutzer) an.

64. Das Maß von Gini gilt als wichtigstes Verfahren für CART-Bäume und kommt auch in dieser Arbeit zum Einsatz. Alternativ kann bei nominalen Daten twoing herangezogen werden, die Ergebnisse sollten sich jedoch nicht wesentlich unterscheiden.

65. Die „Konzentration" wäre also hier am höchstren. Dies heißt aber nicht, dass es keine Fehklassifikationen gibt.

Diese Knoten sollen aber mit den jeweiligen Knotenanteilen (69.33 bzw. 30.67) zum Hauptknoten (100 %) gewichtet werden, um zu prüfen, ob die Homogenität zugenommen hat:

Gewichtetes Mittel: 0.6933 * 0.459 + 0.3067 * 0.2112 = 0.383.

Würde überhaupt keine Konzentration unterstellt, d. h. würde sich die abhängige Variable nicht durch die unabhängigen Variablen segmentieren lassen, wären bei zwei Ausprägungen die Verhältnisse 50 : 50. Dies ist nicht der Fall. Die PC-Nutzung läßt sich sehr wohl durch das Alter (und andere Variablen) erklären.

Verbesserung = 0.5 - 0.383 = 0.117 (gerundet).

Gäbe es keine Segmentierung, so wäre die Verbesserung = 0 - am Hauptknoten würde sich nichts ändern. Der Wert kann also bei zwei Variablen zwischen 0 und 0.5 liegen. Ein Wert von 0.5 zeigt eine perfekte, ein Wert von 0 keine Segmentierung der Gruppen. Die Verbesserungswerte für ein Gesamtmodell werden bei Answertree in einem eigenen Fenster angezeigt:

ABBILDUNG 59 CART-Algorithmus: Verbesserungswerte für Alter und Schulbildung bei abhängiger Variable PC-Nutzung

Würde man den Verbesserungswert als eine Art von Varianz[66] (besser: Informationsaufklärung) begreifen, so gäbe es im Beispiel (bezogen auf 0.5!) einen erklärten Anteil (durch das Alter) von 0.117, einen nicht erklärten Anteil von 0.383 (die Addition ergibt wieder 0.5). Unter zusätzlicher Berücksichtigung weiterer Variablen (hier: des Bildungsabschlusses) ergibt sich eine zusätzliche Verbesserung von 0.0541:

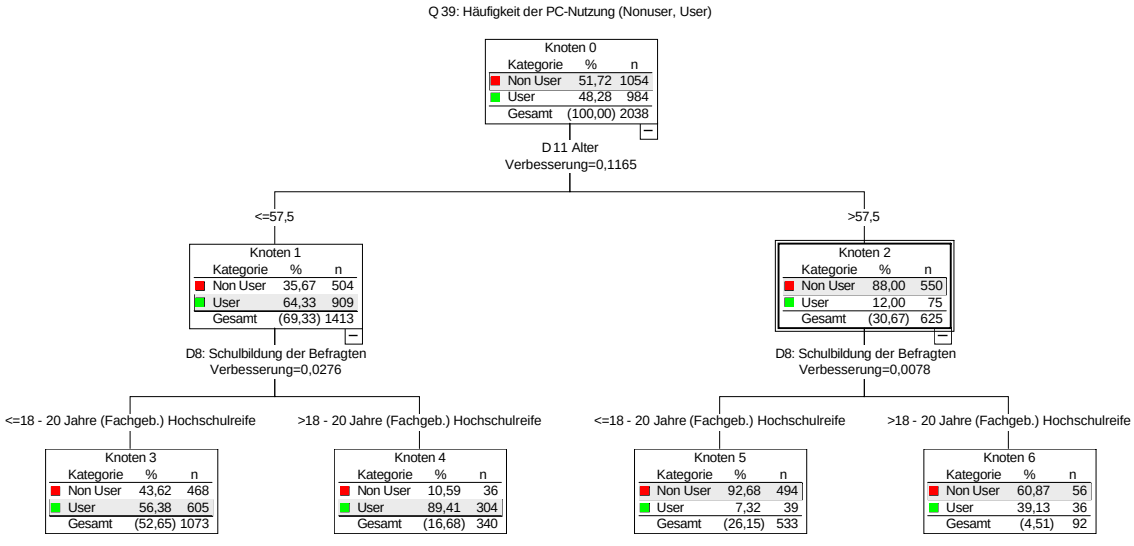
Verbesserung = 0.5 - 0.1165 (Alter) - 0.0541 (Bildung) = 0.3294.

Leider wird dieser Wert nicht in Answertree berechnet, sondern auch hier muss man manuelle Berechnungen durchführen. Berechnet man zusätzlich den Anteil der durch die beiden unabhängigen Variablen „erklärten" Anteile, ergibt sich:

(0.1165 + 0.541) / 0.5 = 0.3412.

34.12 % der „Varianz" werden durch die beiden Variablen Alter und Bildungsgrad „erklärt" - was für zwei Variablen kein allzuschlechtes Ergebnis ist.

ABBILDUNG 60 CART-Algorithmus: 2stufiger Entscheidungsbaum für PC-Nutzung, Alter, Bildung



66. Hierbei handelt es sich natürlich nicht um die „Varianz" einer metrischen Variablen, sondern vielmehr um eine Annäherung auf Nominalniveau ohne rechnerischen Exaktheitscharakter.

Das Baummodell zeigt die einzelnen Segmentierungen: neben der Trennung des Alters bei rund 58 Jahren und des Studienabschlusses (bis Fachabitur einerseits bzw. (abgeschlossenes) Studium andererseits). Die PC-Nutzer liegen eher auf der Seite der jüngeren Befragten (Knoten 4) mit höheren Bildungsabschlüssen (Knoten 6 bei den Älteren).[67]

Neben dem Verbesserungswert ergibt sich eine Fehlklassifikation von rund 0.26. Zwischen diesen beiden Kennzahlen gibt es einen gewissen, aber nicht stringenten Zusammenhang: die Verbesserung gibt an, wie gut die unabhängigen Variablen die abhängige Variable trennen - unabhängig von der Fehlklassifikation, einzig bezogen auf den Baum. Die Fehlklassifikation betrachtet nicht die Verbesserung, sondern nur die nicht richtig zugeordneten Fälle.

Im Gegensatz zur Regression und der Diskriminanzanalyse werden von Answertree konkrete Gruppen gebildet, die anhand der unabhängigen Variablen weiter differenziert werden können.[68]

Die CHAID-basierten Algorithmen bzw. QUEST beruhen nicht auf Inhomogenitätsmaßen, sondern - für nominale Daten - auf der Chi-Quadrat-Statistik.

Für CHAID (Fehlklassifikation: 0.26) ergibt sich:

67. Legt man das Alter ordinalskaliert in 10-Jahres-Abständen zugrunde, erfolgt die Dichotomisierung in der Altersgruppe 50 - 59 Jahre bzw. älter. Deutlich wird bei beiden Segmentierungen, dass die Trennung der Gruppen ungefähr um das Renteneintrittsalter beginnt. Dies zeigt auch die enorme Bedeutung, die der PC heute in vielen Branchen besitzt.

68. Es ist jedoch bei Entscheidungsbäumen nicht zwingend notwendig, dass in den jeweiligen Segmenten die gleichen Splitvariablen verwendet werden. Die Jüngeren könnten nochmals nach dem Alter weiter differenziert, während die Älteren nach der Schulbildung aufgesplittet werden. Hier liegt ebenfalls ein entscheidender Vorteil gegenüber anderen Verfahren (z. B. Diskriminanz-, Faktoren-, Clusteranalyse), die nur die allgemeine Stärke der unabhängigen Variablen ausgeben. Somit können Baumalgorithmen Untergruppen prägnanter untersuchen.

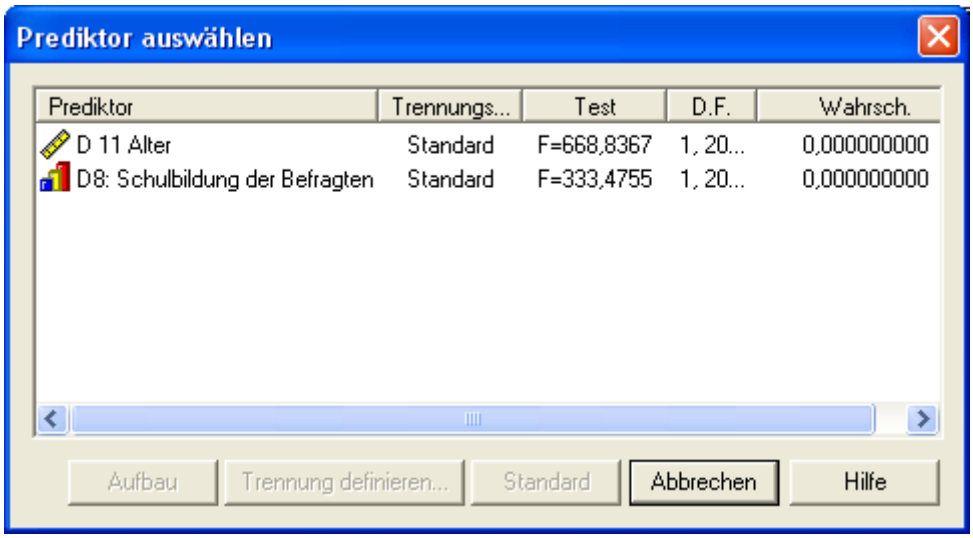
ABBILDUNG 61 CHAID-Algorithmus: Chi-Quadrat-Werte für Alter und Schulbildung bei abhängiger Variable PC-Nutzung

EXHAUSTIVE CHAID (Fehlklassifikation: 0.27) liefert nahezu das gleiche Ergebnis:

ABBILDUNG 62 EXHAUSTIVE CHAID-Algorithmus: Chi-Quadrat-Werte für Alter und Schulbildung bei abhängiger Variable PC-Nutzung

Auch QUEST (Fehlklassifikation: 0.26) unterscheidet sich im Ergebnis kaum - hier wie auch weiterhin fällt auf, dass trotz eindeutig ordinaler Variable (Schulbildung) der F-Wert herangezogen wird:

ABBILDUNG 63 QUEST-Algorithmus: F-Werte für Alter und Schulbildung bei abhängiger Variable PC-Nutzung

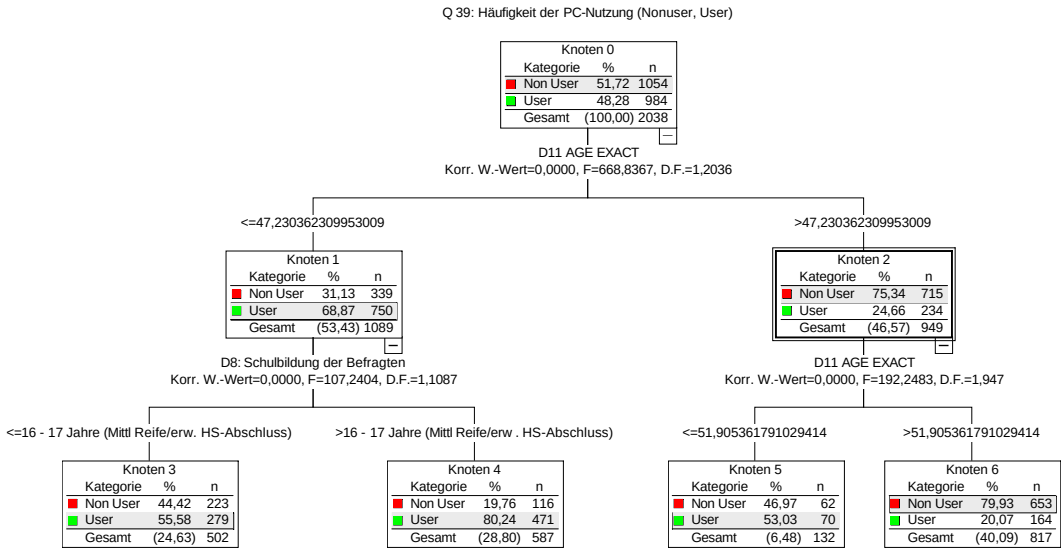


Das Alter ist in allen vier Algorithmen die wichtigste unabhängige Variable. Allerdings unterscheiden sich die binären Algorithmen hinsichtlich der Splits. Während die CART-Segmentierung bei ca. 58 Jahren und nichtakademischen Abschlüssen liegt, trennt QUEST bereits bei 47 Jahren auf der ersten Ebene. Auf der zweiten Stufe wird bei den Jüngeren nach Bildungsabschluss, bei den Älteren nochmals nach dem Alter differenziert.

Überraschenderweise verwendet QUEST bei log-likelihood oder metrischen Daten immer den F-Wert (vgl. SPSS (2001b: 230)). Während die Unterknoten 1 und 2 mit rund 50 % nahezu gleichverteilt sind, was für den weiteren Baumaufbau vorteilhaft sein kann, segmentiert CART (vgl. Seite 161) hier stärker (70 : 30). Dies führt zu einer deutlichen Verbesserung. Dieses Manko wird bei QUEST auch durch die

weitere Alterssegmentierung bei den Älteren und durch die nicht sehr effiziente Trennung in den Knoten 3 und 5 erschwert.

ABBILDUNG 64 QUEST-Entscheidungsbaum: PC-Nutzung, Alter, Bildung (zweistufig)



Um einen - annähernden - Vergleich der beiden Algorithmen durchzuführen, wird - zumindest für die erste Ebene, die sich besonders unterscheidet - die Verbesserung berechnet[69]:

Knoten 1: 1 - 0.3113² - 0.6887² = 1 - 0.0969 - 0.4743 = 0.4288.

Knoten 2: 1 - 0.7534² - 0.2466² = 1 - 0.5676 - 0.0608 = 0.3716.

Gewichtetes Mittel: 0.5343 * 0.4288 + 0.4657 * 0.3716 = 0.4022.

Verbesserung = 0.5 - 0.4022 = 0.0978.

Die Verbesserung fällt folglich hier mit 0.0978 schlechter aus als beim CART-Baum (rund 0.117).[70] Damit ist - für diesen Datensatz bzw. das Beispiel! - mathematisch der CART-Algorithmus etwas besser geeig-

69. Der Verbesserungswert wird nicht für QUEST-Bäume berücksichtigt. Meines Wissens ist er auch noch nie in einem Vergleich so dargestellt worden.

70. Die Knoten des QUEST-Baums sind inhomogener als die des CART-Baums (z. B. CART bei den Älteren: 88 % Nichtnutzer, 12 % Nutzer, QUEST: 75 : 25). Da das Ziel der Verbesserung immer „reinere" Unterknoten hinsichtlich der abhängigen Variablen ist, muss die Verbesserung der Altersvariablen bei QUEST schlechter sein.

net, vor allem, da sich die weiteren Segmentierungen auf die wichtigste unabhängige Variable bzw. Trennung beziehen. Da die Altersvariable am bedeutsamsten ist, wird in dieser Arbeit von den binären Entscheidungsbäumen der CART-Algorithmus bevorzugt.[71]

Gerade bei metrischen Variablen mit vielen Ausprägungen - wie dem Alter in Lebensjahren - tendieren (EXHAUSTIVE) CHAID-Bäume dazu, sehr breit zu werden, wodurch eine direkte Vergleichbarkeit mit den binären Algorithmen erschwert wird. Ein weiterer Nachteil liegt hier darin, dass bei zahlreichen Variablenausprägungen auf der ersten Ebene die Fallzahlen sehr schnell dazu tendieren, Werte unter 30 anzunehmen. Der Entscheidungsbaum wird zwar für die ersten Variablen möglicherweise sehr aussagekräftig, eine weitere Aufteilung kann aber nicht mehr erfolgen. Im schlechtesten Fall finden sich so viele Altersgruppen, die nicht mehr weiter aufgeteilt werden können, weil die Zahl von n = 30 unterschritten wird. Aus diesem Grund sind sie für die ausgewählten Variablen - vor allem das Alter in Lebensjahren - dieser Arbeit nicht die erste Wahl. Möglicherweise können sie aber interessante Beiträge zur Interpretation von Subpopulationen hinsichtlich Kultur- und Freizeitvariablen leisten, da sich in kleineren Gruppen möglicherweise diese Variablen deutlicher niederschlagen. So könnte unterstellt werden, dass insbesondere sehr junge Befragte (bis 25 Jahre) besonders häufig ins Kino gehen. Dies ist aber aus einem binären Baum, der das Alter z. B. bei 58 Jahren trennt, nicht ersichtlich. Hier würde man unterstellen, dass die Kinobesucher bis 58 Jahre alt sind, obwohl sich die Teilpopulation der bis 25jährigen dahinter verbirgt.

71. CART ist nicht grundsätzlich ein besserer Algorithmus als QUEST, er scheint aber für den Datensatz besser geeignet, da er für die wichtigste unabhängige Variable bessere Ergebnisse erzielt und wird folglich auch für den weiteren Gang der Untersuchung herangezogen.

Abschließend sollen nochmals die Merkmale der Algorithmen in einer Tabelle zusammengefaßt werden. Ich möchte ganz bewußt auf die Begriffe „Vorteile“ und „Nachteile“ verzichten, da sie mir an dieser Stelle nicht angemessen erschienen. Jeder Algorithmus hat seine spezifischen Vorzüge. Auch wenn in dieser Arbeit dem CART-Algorithmus der Vorzug gegeben wird, ist dies keinesfalls ein Hinweis, dass die anderen Algorithmen „schlechter“ oder „ungeeigneter“ sind:

TABELLE 10 ZUSAMMENFASSENDE MERKMALE DER IN ANSWER-TREE IMPLEMENTIERTEN ALGORITHMEN

Algorithmus	Merkmale
(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)
CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und „A prioris“
QUEST	• vieldiskutierter, relativ neuer Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus (unabhängige Variablen) geeignet • nur für dichotome Zielvariablen geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, F-Test) • erlaubt Ersatzprädiktoren, Pruning und „A prioris“

4 Multinominale logistische Regression

Regressionsverfahren erfreuen sich - besonders in der „hypothesengeleiteten“, deduktiv-nomologischen Soziologie - einer hohen Beliebtheit. Wie bei den Entscheidungsbäumen soll eine abhängige Variable möglichst gut durch die unabhängigen Variablen erklärt werden.[72] Es gibt jedoch einige bedeutsame Unterschiede bei den beiden Methoden: zum einen werden bei der Regressionsanalyse keine Subpopulationen gebildet, die dann anhand anderer Variablen weiter segmentiert werden können, zum anderen wird ein Modell entwickelt, dass - unter Berücksichtigung aller in die Analyse eingehenden unabhängigen Variablen und gegebenenfalls deren Wechselwirkungen - Abhängigkeitsbeziehungen in Form von (standardisierten) Regressionskoeffizienten und einen Modellfit errechnet. Auch lassen sich direkte und indirekte Effekte errechnen und im Modell berücksichtigen, was bei Entscheidungsbäumen nicht so ohne weiteres möglich ist:

> „Das Interessante an der logistischen Regression ist nun, dass nicht nur die Unterschiede zwischen verschiedenen untersuchten Gruppen bestimmt werden können. Indem sie (Gruppen-) Zugehörigkeitswahrscheinlichkeiten ermittelt, werden auch Aussagen möglich bezüglich der Veränderung eben dieser Wahrscheinlichkeit, wenn eine beobachtete Variable einen anderen Wert annimmt.“ (BACKHAUS (2000: 105))

Deutlich wird dies an einem Beispiel für die lineare multiple Regression: die Funktion einer linearen Regressionsgeraden hat stets die Form aller linearen Geraden, nämlich y (x) = a + bx. Während a die Konstante ist, gibt b den Steigungsgrad der Geraden an. Bei der Abhängigkeit des Einkommens von der Anzahl der Jahre im Berufsleben könnte sich folgendes ergeben: konstant (a) = 500 Euro, pro Berufs-

72. Im Gegensatz zu den Entscheidungsbäumen gibt es eine Vielzahl von Literaturverweisen für Regression und Diskriminanzanalyse - nahezu in jeder Veröffentlichung über multivariate Verfahren (z. B. BACKHAUS et al. (2004)). Aus diesem Grund und da diese beiden Verfahren eher eine Kontrollfunktion besitzen, werden sie in aller Kürze behandelt.

jahr (b) steigt der Verdienst um 200 Euro. Dies würde bedeuten: ohne Berufsjahre ergibt sich ein monatliches Einkommen von 500 Euro, nach dem ersten Berufsjahr 500 + 200 * 1 = 700 Euro, nach 2 Berufsjahren 500 + 200 * 2 = 900 Euro, usw. Kritikpunkt an dieser Art der Regression ist eine gewisse Plausibilität der Werte - z. B. ergibt sich nach 40 Jahren ein Einkommen von 500 + 200 * 40 = 8500 Euro, was etwas realitätsfremd ist. Deshalb ist die lineare Regression nur für Zusammenhänge geeignet, die auch tatsächlich linear sind.

Bei der logistischen Regression können die Werte nicht „ins Unendliche“ gehen, da sie durch den Logarithmus in die Grenzen 0 und 1 transformiert werden (vgl. URBAN (1993: 29), BACKHAUS (2000: 110f)). Unterstellt wird hier eine Wahrscheinlichkeitsfunktion, die eine Summenfunktion der Normalverteilung darstellt (vgl. PAMPEL (2000: 56)): am Wendepunkt ändern sich die Werte stark, an den „Rändern“ weniger stark. BACKHAUS et al. belegen dies an folgendem Beispiel:

> „Wenn also jemand eingeschworener Fan einer Fußballmannschaft ist (unabhängige Variable 1) werden ihn auch einige verlorene Spiele (unabhängige Variable 2) nicht dazu bewegen, seine Besuche im Stadion (abhängige kategoriale Variable) zu beenden. Betrachten wir hingegen einen sehr viel weniger fanatischen Fußballfan, kann eine gewisse Zahl verlorener Spiele sehr wohl aus einem Stadionbesucher einen Sportschaugucker machen.“ (BACKHAUS et al. (2000: 110))

Umgekehrt wird jemand, der völlig desinteressiert an Fußball ist, kein Stadion besuchen - unabhängig davon, ob eine Mannschaft gewinnt oder nicht.

Dieses Beispiel läßt sich auch auf alle möglichen Dinge des Alltags anwenden: überzeugte Wähler einer Partei werden eher nicht zu Wechselwählern wie Personen, die einer Partei nicht so nahestehen, Patienten mit hohem Blutdruckwert neigen häufiger zu Folgeerkrankungen als Befragte mit niedrigem, usf. Die Annahme der linearen Regressionsfunktion, die einen je ... desto-Zusammenhang unterstellt,

ist hier realistischer gefasst: die Wahrscheinlichkeit der Folgeerkrankungen ist z. B. bei hohem Blutdruck grundsätzlich höher als bei niedrigem. Ob der Wert nun 10 %, 15 % oder 20 % über dem normalen (altersabhängigen) Wert liegt, ist für den Ausbruch der Folgeerkrankung nicht ausschlaggebend, sondern von der Person. Er steigt nicht linear an. Man könnte sagen, dass bei stark erhöhtem Blutdruck die Wahrscheinlichkeit von Folgeerkrankungen eher gegeben ist als bei niedrigen Werten. Somit liefert diese Art der Regression - gerade für soziologische oder medizinische Fragestellungen, die sich nicht deterministisch fassen lassen - ein realistischeres Bild als die lineare Version.

PAMPEL (2000: 74ff.) beschreibt die Auswirkungen des Logarithmus an einem einfachen Beispiel. Er zeigt, dass die prozentuale Erhöhung beim 10er-Logarithmus immer gleich hoch ist, da der prozentuale Abstand zwischen logX 1 = 10 bzw. logX 2 = 100 mit 900 % genauso hoch ist wie zwischen logX 4 = 10.000 und logX 5 = 100.000.

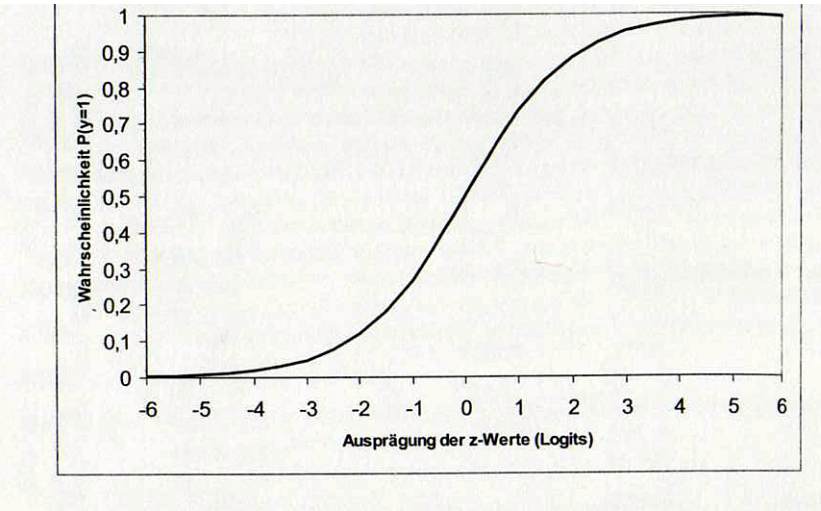
TABELLE 11 LOGARITHMUS ZUR BASIS 10

X	log X	Anstieg pro Einheit
10	1	1.04
100	2	2.004
1000	3	3.0004
10000	4	4.00004
100000	5	5.000004

Mit anderen Worten: der s-förmige Kurvenverlauf, basierend auf der Normalverteilung kommt durch die, zu Beginn leicht, dann immer höher ansteigenden Werte der „Glockenkurve“ zustande, welche den Anstieg in der Summenkurve symbolisiert. Mit zunehmenden Werten steigt auch die Logarithmusfunktion deutlich an („Anstieg pro Ein-

heit“). Allerdings fällt die Normalverteilungskurve mit sinkenden Werten schnell ab, was in der Summenkurve durch die Annäherung an 1 charakterisiert wird.

ABBILDUNG 65 Summenkurve der logistischen Regression (vgl. BACKHAUS et al (2004: 424))



Somit liegt ein Vorteil dieser Verteilung in der geringen Gewichtung von „Ausreißern“, also Extremwerten, die entweder nahe 0 oder nahe 1 der Summenkurve liegen.

Die multinominale logistische Regression soll ebenfalls an dem gewählten Beispiel erläutert werden. Der Vorteil dieses Verfahrens gegenüber der binär-logistischen Regression besteht in einer Verwendung binärer und polytomer Variablen. Der Unterschied zwischen beiden Verfahren liegt darin, dass die binär-logistische Regression auf Einzelfallebene, während die multinominale logistische Regression mit den Teilgesamtheit en (in diesem Fall: Nutzer - Nichtnutzer) rechnet. Bei den Ergebnissen scheinen sich jedoch die Verfahren kaum zu unterscheiden (vgl. LUDWIG-MAYERHOFER (1990: 80)).

Für die abhängige nominal-dichotome Variable PC-Nutzung (ja - nein) und den unabhängigen Variablen Alter und Schulbildung ergibt sich bei der logistischen Regression[73] folgendes Ergebnis:

ABBILDUNG 66 Logistische Regression: Modellanpassung am Beispiel von PC-Nutzung (abhängig) , Alter und Bildung (unabhängig)

Informationen zur Modellanpassung

Modell	-2 Log-Likelihood	Chi-Quadrat	Freiheitsgrade	Signifikanz
Nur konstanter Term	1502,338			
Endgültig	750,683	751,655	5	,000

Der konstante Term (-2 Log-Likelihood) in obenstehender Tabelle gibt den Wert an, den die Regressionsgleichung erzielt, wenn keine unabhängigen Variablen in die Gleichung eingehen bzw. alle unabhängigen Variablen auf Null gesetzt werden. Dieser Wert (1502.338) wird mit dem Wert verglichen, in dem die unabhängigen Variablen eingesetzt wurden (endgültig: 750.683). Unterscheidet sich der Wert nicht, so haben die unabhängigen keinen Einfluss auf die abhängige Variable. Das bedeutet für das Beispiel, dass die Variable PC-Nutzung (ja - nein) alleine für sich einen Wert von 1502.338 ergibt. Hätten die Variablen Alter und Schulbildung keinerlei Einfluss, dürfte der Wert nicht (wesentlich) abweichen. Dies ist aber in diesem Beispiel der Fall und es ergibt sich ein Chi-Quadrat-Wert aus der Differenz von 1502.338 - 750.683 = 751.655. Der Chi-Quadrat-Wert zeigt in diesem Falle an, dass sich die erwartete von der tatsächlich untersuchten Häufigkeit unter der Bedingung deutlich unterscheidet, dass kein Zusammenhang besteht (er ist ungleich 0). Das Modell ist zusätzlich si-

73. Es gibt verschiedene Arten, die unabhängigen Variablen mit der abhängigen zu vergleichen. Hier wurde eine schrittweise Eingabe herangezogen, um eine bessere Vergleichbarkeit mit den Entscheidungsbäumen zu schaffen.

gnifikant (.000), die Annahme, dass PC-Nutzung abhängig von Alter und Schulbildung ist, läßt sich auf die Grundgesamtheit übertragen.[74]

An dieser Stelle gibt es gewisse Parallelen zu den Entscheidungsbäumen: die herangezogene Kennzahl (Chi-Quadrat) kommt hier ebenfalls zum Einsatz.

ABBILDUNG 67 Logistische Regression: Verschiedene Pseudo-R-Quadrat-Werte (PC-Nutzung, Alter, Bildung

Pseudo-R-Quadrat	
Cox und Snell	,308
Nagelkerke	,411
McFadden	,266

Das Pseudo-R-Quadrat gibt den Modellfit an, d. h. wie gut die PC-Nutzung durch Alter und Bildung erklärt werden kann. Die Werte liegen zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang). SPSS gibt hier drei Kennzahlen aus. Alle drei Koeffizienten folgen der Likelihood-Funktion - mit einigen Variationen (vgl. BACKHAUS (2004: 440ff.)).[75]

Während der COX und SNELL-Koeffizient aufgrund der Tatsache, dass er den Wert 1 mathematisch nicht erreichen kann, eher weniger für die Analyse herangezogen wird, präferieren viele Sozialwissenschaftler den NAGELKERKE-Koeffizienten, der eine Modifikation von COX

74. Die Angabe der Freiheitsgrade soll hier nicht irritieren und näher interpretiert werden. Ganz allgemein geben Freiheitsgrade in einer Kreuztabelle die Zellen an, die besetzt werden müssen, um auf den Rest der Tabelle zu schließen (bei gegebenen Randverteilungen). Bei einer 2 x 2-Tabelle, bei denen die Randverteilungen gegeben sind, ist der Freiheitsgrad 1, d. h. man kann eine Zelle „frei" besetzen, um anhand der Randsummen auf die restliche Tabelle zu schließen und diese mittels Subtraktion zu errechnen.

75. Die Pseudo-R-Quadrat-Statistiken der logistischen Regression sind den Verbesserungen der CART-Bäume nicht unähnlich.

und SNELL darstellt (vgl. DIAZ-BONE (o. J.: 10)). Die Maße sollten mindestens 0.2 liefern, ab 0.5 kann von einem sehr guten Modell gesprochen werden. Der ebenfalls ausgegebene MCFADDEN-Koeffizient kann als PRE-Maß verstanden werden, also als Vorhersagemaß. Ihm liegt die gleiche Logik wie z. B. dem Unsicherheitskoeffizienten in der bivariaten Statistik zugrunde: höhere Werte geben eine „*relative Verbesserung des Modells gegenüber dem Ausgangsmodell*" (DIAZ-BONE (o. J.: 9)) an. Für den MACFADDEN-Koeffizienten würde das bedeuten, dass Alter und Bildung das Ausgangsmodell (ohne Variablen) um 26.6 % - im Sinne einer Varianzaufklärung - verbessern.

Für das Beispiel bedeutet dies, dass die beiden Variablen Alter und Schulbildung die PC-Nutzung (Nagelkerke = .411) recht gut erklären - was für dieses Kurzbeispiel doch beachtlich ist.

ABBILDUNG 68 Likelihood-Quotienten-Tests für PC-Nutzung (abhängig), Alter und Bildung (unabhängig)

Likelihood-Quotienten-Tests

Effekt	-2 Log-Likelihood für reduziertes Modell	Chi-Quadrat	Freiheitsgrade	Signifikanz
Konstanter Term	750,683[a]	,000	0	.
bildung	946,092	195,409	4	,000
alter	1124,658	373,975	1	,000

Die Chi-Quadrat-Statistik stellt die Differenz der -2 Log-Likelihoods zwischen dem endgültigen Modell und einem reduziertem Modell dar. I reduzierte Modell wird berechnet, indem ein Effekt aus dem endgültige Modell weggelassen wird. Hierbei liegt die Nullhypothese zugrunde, na der alle Parameter dieses Effekts 0 betragen.

a. Dieses reduzierte Modell ist zum endgültigen Modell äquivalent, da das Weglassen des Effekts die Anzahl der Freiheitsgrade nicht erhöht.

Der Wert des konstanten Terms (750.68) entspricht dem endgültigen Wert der Modellanpassung aus Abbildung 66 auf Seite 172. Wenn nun die Variable Bildung auf Null gesetzt wird (also nur noch das Alter enthalten ist), steigt der Chi Quadrat Wert um 373.975 und es ergibt sich:

750.683 (konstanter Term) + 373.975 (Alter) = 1124.658

Analog erhält man für den Ausschluß von Alter:

750.683 (konstanter Term) + 195.459 (Bildung) = 946.092.

Alle Ergebnisse sind signifikant und somit auf die Grundgesamtheit übertragbar. Auch hier weichen die Chi-Quadrat-Werte deutlich von 0 ab.[76] Der (für sich betrachtete) Chi-Quadrat-Wert des Alters ist mit rund 374 etwas höher als der der Bildung mit rund 195, was die Wichtigkeit des Altes vor dem Bildungsabschluss wiederspiegelt (hierbei sind natürlich auch Wechselwirkungen der Bildungsexpansion zu berücksichtigen).

ABBILDUNG 69 Parameterschätzer für PC-Nichtnutzung (abhängig), Alter und Bildung (unabhängig)

Häufigkeit der PC-Nutzung						
		B	Standardfehler	Wald	Signifikanz	Exp(B)
User	Konstanter Term	4,368205684	0,249616211	306,2390619	0,00000	
	[bildung=1,00]	-2,223080592	0,186779876	141,6609825	0,00000	0,108275043
	[bildung=2,00]	-1,767091513	0,172011305	105,5368439	0,00000	0,170829121
	[bildung=3,00]	-1,276205523	0,177378971	51,7650902	0,00000	0,279094311
	[bildung=4,00]	-0,161448415	0,432604407	0,139279017	0,70900	0,850910425
	[bildung=5,00]	0	.	.	.	.
	alter	-0,06628507	0,003848779	296,6098371	0,00000	0,935864039
a	Die Referenzkategorie lautet: Non User.					
b	Dieser Parameter wird auf Null gesetzt, weil er redundant ist.					

76. Der Unterschied zu den Entscheidungsbäumen bei geringen Chi-Quadrat-Werten liegt darin, dass unabhängige Variablen nicht in den Aufbau von Entscheidungsbäumen einbezogen werden und somit nicht in einem möglichen Output auftauchen. Sie sind alleine im Fenster „Prädiktor auswählen …" sichtbar und können „manuell" zur Trennung herangezogen werden, wodurch sie flexibler einsetzbar sind als in allen anderen Verfahren.

Tabelle 69 auf Seite 175 gibt die wichtigsten Ergebnisse der sog. „Parameterschätzung“ für die Nutzer wieder. In der ersten Zeile ist der konstante Term (PC-Nutzung ohne unabhängige Variablen) wiedergegeben, die nachfolgenden Zeilen geben die Bildungskategorien (1 = Volks-, Hauptschule bis 5 = Akademiker), gefolgt vom Alter an.

Die letzte Kategorie (Akademiker) wird als Referenzkategorie betrachtet und keine Werte ausgegeben. Die oberen vier Zeilen geben das Verhältnis zu dieser Referenzkategorie an.

EXP(B) für Volks- und Hauptschulabgänger wird mit 0.1082 angegeben (Zeile bildung = 1, letzte Spalte EXP(B)). Damit liegt die Wahrscheinlichkeit der PC-Nutzung bei Volks-, Hauptschulabsolventen, im Gegensatz zu den Akademikern nur bei rund 11 %. Der EXP(B)-Wert steigt kontinuierlich bis zu den Studierenden an (Mittlere Reife/erw. Hauptschulabschluss: 17 %, (fachgeb.) Hochschulreife: 28 %, Studierende: 85 %).

Die Spalten B und EXP(B) enthalten die Regressionskoeffizienten. B ist ein unstandardisierter Koeffizient, der - wie an der Abbildung sichtbar - positive und negative Werte ergeben kann, aber schwer zu interpretieren ist, da er nicht standardisiert ist. Deshalb wird in der Regel die Werte in der Spalte EXP (B) herangezogen.

Der B-Wert ist jedoch für die Berechnung weiterer Werte aus der Tabelle wichtig. Die Wald-Statistik errechnet sich aus dem jeweiligen Koeffizienten (B) und dem Standardfehler: der Koeffizient wird durch den Standardfehler geteilt und das Ergebnis quadriert. Dadurch können keine negativen Ergebnisse entstehen.

Der Standardfehler hängt mit den Vertrauensintervallen zusammen: er gibt an, in welchem Bereich um den Mittelwert 95 % der Fälle liegen. Je größer der Standardfehler, desto weiter streuen die Werte

um den Mittelwert (das ist natürlich auch von der Skalierung abhängig: bei einer Skala von 1 - 5 ist eine geringere Streuung zu erwarten als bei einer Skala von 1 - 100).

Die Wald-Statistik gibt die Chi-Quadrat-Verteilung für die Variablen an - zum Beispiel für die Volks- und Hauptschulabsolventen:

$$\text{Wald (Volks-, Hauptschule)} = (-2.22 / 0.1868)^2 = 141.66$$

Wäre der Standardfehler deutlich höher (zum Beispiel bei 0.3) ergäbe sich ein Wald Chi Quadrat-Wert von rund 54.76 er liegt also deutlich niedriger und würde - durch die größeren Abweichungen - zu einem geringeren Unterschied beitragen.

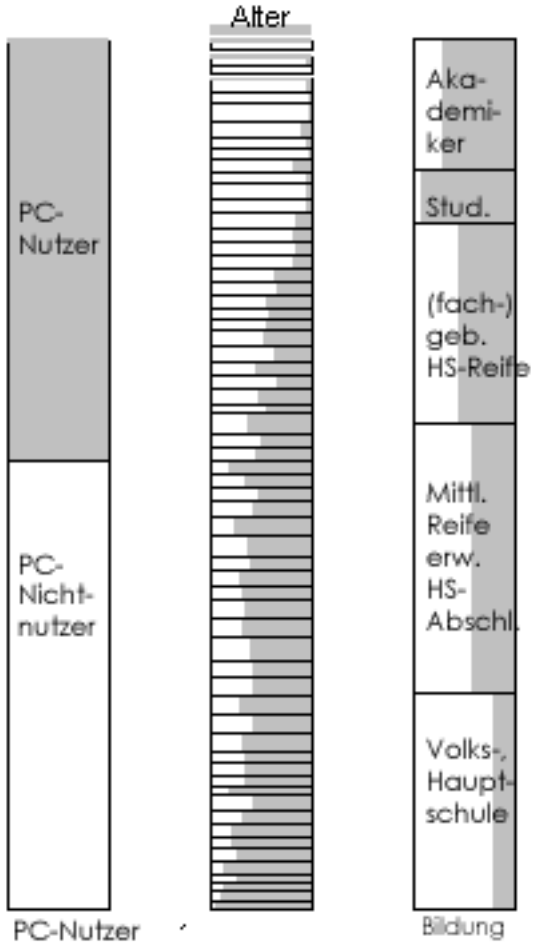
Daraus folgt: je kleiner der Standardfehler bei einer Variablen, desto größer der Chi Quadrat-Wert.

Deutlich wird, dass vor allem in geringeren Bildungsgruppen (1 - 3) die Wald- (Chi-Quadrat-)Werte am höchsten sind (141, 105, 51) - am höchsten trägt jedoch das Alter dazu bei (297). Alle Werte sind signifikant, ausser der der Kategorie (Studierende).

An diesem Beispiel offenbart sich eine grundsätzliche Schwäche der Regression: die Unanschaulichkeit der Ergebnisse bzw. die mangelnde Ausgabe in SPSS - denn auch hier wie an einigen Stellen (auch in Answertree!) müssen nachträglich mit dem Taschenrechner die Werte „nachberechnet“ werden.

Für statistische Laien und selbst Fortgeschrittene ist die Interpretation der (logistischen) Regression zumeist nicht einfach. Aus diesen Gründen werden multivariate Verfahren so gut wie nie in direkter Form mit statistischen Laien kommuniziert - was eigentlich schade ist, wenn man bedenkt, dass durch Visualisierung der Sachverhalt wesentlich deutlicher wird:

ABBILDUNG 70 Parallel Boxplot: PC-Nutzung, Alter und Bildung (N = 2038)



Die Nichtnutzergruppen (weiß) ergeben hinsichtlich der anderen Variablen eine deutliche Struktur: sie sind älter und haben eine geringere Bildung. Ohne mit logarithmierten Werten arbeiten zu müssen, liegen die Ergebnisse auf der Hand - und dies ist nicht nur für zwei, sondern für viele Variablen denkbar. Allerdings können keine mathematischen Ergebnisse, wie z. B. die Fehlklassifikation, erwartet werden. Dadurch wird bei grafischen Modellen die Vergleichbarkeit erschwert.

ABBILDUNG 71 Logistische Regression: Fehlklassifikationsmatrix (PC-Nutzung, Alter, Bildung)

Klassifikation

	Vorhergesagt		
Beobachtet	Non User	User	Prozent richtig
Non User	730	324	69,3%
User	226	758	77,0%
Prozent insgesamt	46,9%	53,1%	73,0%

Die Klassifikationsmatrix ist ebenfalls ein Validierungselement des Modells. Der große Vorteil liegt darin, dass sowohl Entscheidungsbäume als auch die Diskriminanzanalyse ebenfalls mit Klassifikationsmatritzen arbeiten, so dass die Endergebnisse direkt vergleichbar sind.

Von den PC-Nichtnutzern wurden aufgrund der unabhängigen Variablen Alter und Bildung 69 % richtig und rund 31 % falsch klassifiziert (Nutzer: 77 % vs. 23 %). Durchschnittlich wurden 73 % der Fälle den tatsächlichen Kategorien richtig zugeordnet.

Dies ist kein überragendes Ergebnis, wenn man bedenkt, dass bei zwei Kategorien der abhängigen Variablen eine reine Zufallszuordnung schon 50 % erbringen müßte. Da es sich aber nur um ein einfaches Modell handelt, ist zu erwarten, dass sich bei der Analyse mit mehreren Variablen im Folgekapitel dieser Fehleranteil weiter sinkt.

Das Ergebnis der Fehlklassifikationsmatrix wird am Ende des Kapitels mit den anderen Verfahren verglichen.

5 Diskriminanzanalyse

Die Diskriminanzanalyse untersucht Gruppen (=abhängige Variable) nach bestimmten Merkmalen (unabhängige Variablen) und faßt diese Variable in Dimensionen zusammen. Unter dem Begriff der Diskriminanz wird die Trennung von Gruppen anhand bestimmter unabhängiger Variablen verstanden.

Auch hier soll wieder das Beispiel PC-Nutzung als abhängige Variable herangezogen werden und Alter und Bildung als unabhängige Variablen. Allerdings gibt es bei dieser Methode eine Einschränkung hinsichtlich des Skalenniveaus: es dürfen nur metrische oder dichotome unabhängige Variablen verwendet werden. Auch wenn für diesen Fall eine Verletzung der Voraussetzungen vorliegt, wird der ordinalskalierte Bildungsabschluss als metrisch unterstellt. Dies ist ein grober statistischer Fehler - das Ziel ist jedoch in diesem Kapitel nicht, ein analysefähiges Modell aufzustellen, sondern die Koeffizienten und die Verfahren zu erläutern. Möglicherweise kann es dadurch zu Verzerrungen kommen.

Nachfolgendes Zitat charakterisiert die Unterschiede zwischen Diskriminanz- und Regressionsanalyse:

> „Trotz der formalen Ähnlichkeit bestehen gravierende *modelltheoretische Unterschiede* zwischen Regressionsanalyse und Diskriminanzanalyse. Die abhängige Variable des Regressionsmodells ist eine Zufallsvariable, während die unabhängigen Variablen fix sind. Im statistischen Modell der Diskriminanzanalyse ... verhält es sich genau umgekehrt, d. h. die Gruppen sind fixiert und die Merkmale variieren zufällig ...“ (BACKHAUS et al. (2000: 167))

Die Gruppen sind in diesem Beispiel PC-Nutzer bzw. Nichtnutzer, die Merkmale das Alter und der Bildungsgrad.

> „Mit Hilfe der Diskriminanzanalyse wird ein Individuum aufgrund von Merkmalen (unabhängigen Variablen) einer von zwei oder auch mehreren fest vorgegebenen Gruppen zugeordnet.“ (BÜHL und ZÖFEL (2002a: 431))

Anders ausgedrückt: wie lassen sich PC-Nutzer bzw. Nichtnutzer anhand von Alter und Bildung charakterisieren?

ABBILDUNG 72 Diskriminanzanalyse: Gleichheitstest der Gruppenmittelwerte (PC-Nutzung, Alter, Bildung)

Gleichheitstest der Gruppenmittelwerte

	Wilks-Lambda	F	df1	df2	Signifikanz
D8: Schulbildung der Befragten	,859	333,476	1	2036	,000
D 11 Alter	,753	668,837	1	2036	,000

Der Gleichheitstest der Gruppenmittelwerte zeigt an, wie gut Alter und Bildung die PC-Nutzung für sich genommen unterscheiden können. Die Werte sind höchstsignifikant, der F-Wert des Alters deutlich höher als der der Bildung - dies bedeutet, dass das (isoliert betrachtete) Alter die Kategorien der PC-Nutzung wesentlich besser trennt als der (isoliert betrachtete) Bildungsabschluss (669 : 333).[77] Die „Logik" ist eine ähnliche wie hinter der Chi-Quadrat-Statistik: höhere F-Werte haben eine höhere Erklärungskraft als niedrigere. Je höher der F-Wert, desto niedriger der (bivariate) Lambda-Wert.

Der bivariate Lambda Wert gibt in dieser Tabelle an, wie gut die Gruppen nach der unabhängigen Variablen getrennt werden können.

Da es bei der Diskriminanzanalyse jedoch um die Trennung bzw. Beschreibung von Gruppen anhand unabhängiger Variablen geht, wird nicht mit „Erklärungskraft", sondern mit „möglichst hoher Beitrag zur Trennung der Kategorien der abhängigen Variablen" gearbeitet.

77. Die F-Werte sind identisch mit den F-Werten der QUEST-Analyse (vgl. Abbildung 63 auf Seite 164).

Das Ziel ist gleich, das Prinzip ähnlich. Die Freiheitsgrade werden in den Spalten df1 und df2 angegeben.

Der multivariate WILKs Lambda-Wert in nachfolgender Abbildung gibt an, ob sich die Mittelwerte von Alter und Bildung in den beiden untersuchten (Nicht-)Nutzergruppen unterscheiden. Dieser Wert ist höchst signifikant.

ABBILDUNG 73 Diskriminanzanalyse: Eigenwerte (PC-Nutzung, Alter, Bildung)

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,463[a]	100,0	100,0	,562

a. Die ersten 1 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

Da nur zwei unabhängige Variablen in die Analyse eingehen, kann es auch nur eine Diskriminanzfunktion geben (Anzahl der Dimensionen (= unabhängigen Variablen) - 1 = 1). Der Zusammenhang zwischen Alter und Schulbildung hinsichtlich der beiden Gruppen PC-(Nicht-)Nutzer ist mit .562 ganz brauchbar, wenn auch nicht überragend gut. Allerdings unterscheiden sich die beiden Gruppen aufgrund der Diskriminanzfunktion hinsichtlich der Signifikanz deutlich:

ABBILDUNG 74 Diskriminanzanalyse: WILKs Lambda (Test der Funktion(en) (PC-Nutzung, Alter, Bildung))

Wilks' Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1	,684	773,730	2	,000

Im Beispiel kann nur eine Funktion (2 unabhängige Variablen - 1 = 1 Funktion) ermittelt werden.

WILKs LAMBDA (multivariat) zeigt an, wieviel Prozent der Information durch die zwei unabhängigen Variablen **im Gesamtmodell**, also multivariat erklärt wird (68.4 %). Die Differenz (31.6 %) kann nicht durch Alter und Bildungsgrad begründet werden - es müßten andere Variablen herangezogen werden, um mehr Varianzaufklärung zu erreichen.

ABBILDUNG 75 Diskriminanzanalyse: standardisierte kanonische Diskriminanzfunktionskoeffizienten (PC-Nutzung, Alter, Bildung)

Standardisierte kanonische Diskriminanzfunktionskoeffizienten

	Funktion
	1
D8: Schulbildung der Befragten	-,540
D 11 Alter	,806

Für den Fall mit zwei unabhängigen Variablen ergibt sich auf den ersten Blick ein etwas unanschauliches Ergebnis: Der Wert der Schulbildung ist negativ, der des Alters positiv. Mit anderen Worten bedeutet das: die „Höhe“ des Alters und der Schulbildung laufen entgegengesetzt. Je höher die Schulbildung, desto jünger sind die Befragten und umgekehrt..

Die errechneten Gruppenmittelwerte werden z-transformiert, d. h. in einen Wertebereich zwischen -3 und +3 überführt:

ABBILDUNG 76 Funktionen bei den Gruppen-Zentroiden (PC-Nutzung, Alter, Bildung)

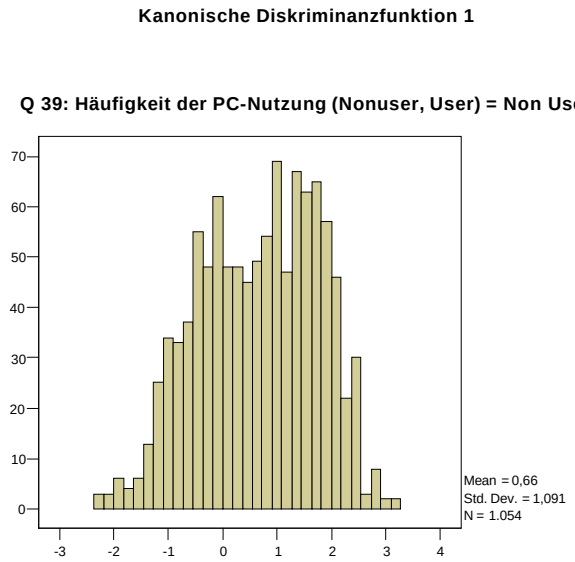
Funktionen bei den Gruppen-Zentroiden

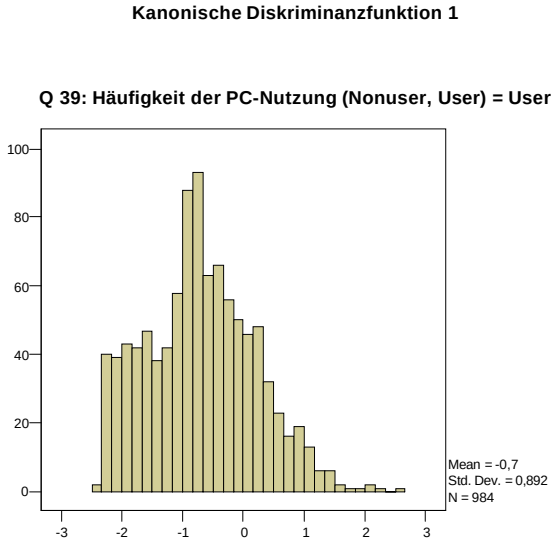
Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Funktion
	1
Non User	,657
User	-,704

Nicht-standardisierte kanonische Diskriminanzfunktionen, die bezüglich des Gruppen-Mittelwertes bewertet werden

Für zwei abhängige Merkmalsausprägungen gibt es eine Funktion (n - 1), die an den nachfolgenden Grafiken deutlich wird. In der ersten Grafik liegen die Non-User eher auf der rechten Seite, während die User ihren Schwerpunkt in der zweiten Grafik eher links haben. Die Analogie findet sich ebenfalls in der Abbildung oben (Funktionen bei den Gruppen-Zentroiden). Die Dimension reicht (standardisiert) von +3 bis -3.

ABBILDUNG 77 Kanonische Diskriminanzfunktionen für Non User und User





Die Struktur-Matrix macht deutlich, wie die unabhängigen Variablen die Funktion beeinflussen. Es handelt sich hierbei um die empirischen Variablen Alter und Schulbildung, die mit den Funktionswerten in Beziehung gesetzt werden:

ABBILDUNG 78 Struktur-Matrix (PC-Nutzung, Alter, Bildung)

Struktur-Matrix

	Funktion
	1
D 11 Alter	,843
D8: Schulbildung der Befragten	-,595

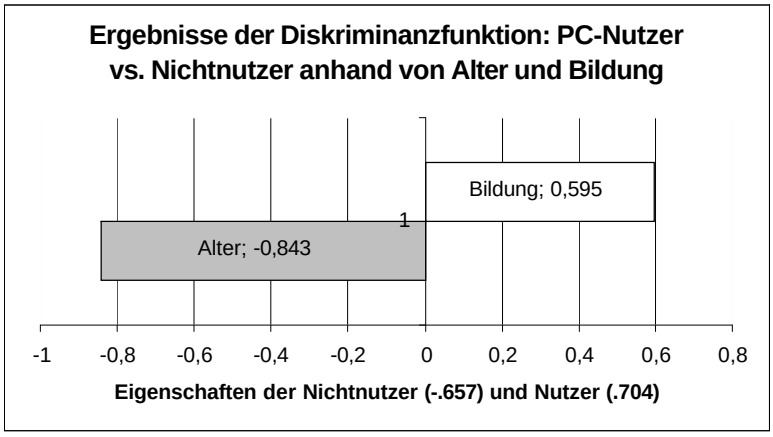
Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen
Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

Alter wird als positive, die Bildung als negative Korrelation hinsichtlich der PC-Nutzung dargestellt. Die Höhe des Alters und des Bildungsabschlusses sind also ebenfalls gegenläufig. Diese Gegenläufigkeit spie-

gelt sich auch in der Strukturmatrix wieder: PC-Nichtnutzer, die sich eher auf der rechten Seite der Grafik finden, sind älter (positiver Wert) und haben geringere Bildungsabschlüsse (negativer Wert) Bei den Nutzern, die sich im negativen Bereich finden, erfolgt ein Vorzeichenwechsel: das Alter geht negativ in die Berechnung ein (eher jünger), die Bildung (höhere Abschlüsse) positiv (- * - = +).

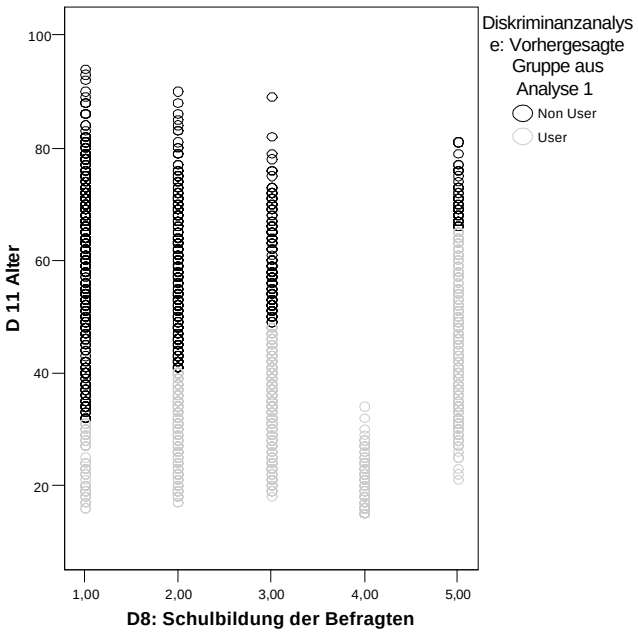
Leider stellt SPSS keine sinnvolle Grafik für zwei unabhängige Variablen zur Verfügung. Der Sachverhalt läßt sich jedoch einfach an einem Balkendiagramm illustrieren (zur besseren Lesbarkeit wurden die positiven/negativen Werte der Gruppen-Zentroiden und der Strukturmatrix vertauscht):

ABBILDUNG 79 Grafische Aufbereitung der Diskriminanzfunktion mit zwei unabhängigen Variablen (Alter, Bildung)



Während das Alter negativ auf die PC-Nutzung wirkt, ist dieser Effekt bei der Höhe des Bildungsabschlusses umgekehrt: PC-Nutzer sind also eher jünger und besser gebildet. Natürlich läßt sich dieses Ergebnis als Streudiagramm wiedergeben:

ABBILDUNG 80 Diskriminanzanalyse: zweidimensionale Grafik PC-Nutzung nach Alter und Schulbildung



Der letzte Blick richtet sich auf die Klassifikationsmatrix: wieviele Fälle wurden richtig klassifiziert?

ABBILDUNG 81 Diskriminanzanalyse: Fehlklassifikationsmatrix (PC-Nutzung, Alter, Bildung)

Klassifizierungsergebnisse[a]

		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Vorhergesagte Gruppenzugehörigkeit		
			Non User	User	Gesamt
Original	Anzahl	Non User	708	346	1054
		User	200	784	984
		Ungruppierte Fälle	3	6	9
	%	Non User	67,2	32,8	100,0
		User	20,3	79,7	100,0
		Ungruppierte Fälle	33,3	66,7	100,0

a. 73,2% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Das Ergebnis fällt hier nahezu in gleicher Höhe wie bei der Regression aus: knapp 3/4 aller Fälle wurde richtig klassifiziert. Somit ergeben sich - was die Kreuzvalidierung angeht - kaum Unterschiede.

6 Zusammenfassung und Ableitungen für die empirische Untersuchung

Ziel dieses Kapitels war es, die Methoden offenzulegen und für die vorliegende Arbeit eine Strategie auszuwählen. Die Fragestellung - Computernutzung, verknüpft mit dominanten Schichtungen - läßt sich aus o. g. Gründen eher quantitativ mit einer Sekundäranalyse untersuchen. Das Gebiet ist gut erforscht, es liegen auch umfangreiche Primärerhebungen vor, die sich auf die Fragestellung anwenden lassen.

Die Einschränkung des Datensatzes, dass viele Variablen nominal oder ordinal skaliert sind, ist nur auf den ersten Blick ein Problem. Zwar lassen sich gängige multivariate Verfahren (z. B. Faktorenanalyse) nicht einsetzen, dafür bieten Entscheidungsbäume mit ihren Visualisierungsmöglichkeiten zum einen eine gute Alternative, zum anderen auch die Möglichkeit, mit den in SPSS zur Verfügung stehenden Verfahren der Diskriminanzanalyse und der logistischen Regression zu arbeiten. Hier gilt es, zu klären, ob Entscheidungsbäume ähnlich gut/ schlecht klassifizieren wie die anderen Verfahren.

Wenn Entscheidungsbäume - so die erste methodisch-forschungsleitende Fragestellung - eine ernsthafte Alternative bzw. Ergänzung zu den bewährten Verfahren der Logistischen Regression und Diskriminanzanalyse darstellen, müssen die Gruppen nahezu identisch klassifiziert werden. Anhand der Fehlklassifikationsmatrix lassen sich diese Feststellungen treffen.

Allerdings ist es nicht Anspruch der Arbeit, das Verfahren mit der geringsten Fehlklassifikation als „bestes“ Verfahren zu bezeichnen - in der Soziologie sollten eher inhaltliche als mathematische Ergebnisse zählen. Es scheint viel interessanter, die gefundenen Segmente weiter deskriptiv zu beschreiben - auch mit dem „Nachteil“ eines vielleicht mathematisch weniger „komplizierten“ und schwer zu durchschauenden Ergebnisses, denn Menschen lassen sich nicht (bis ins letzte Detail) vermessen. Schon ein veränderter Wert (z. B. eine 2 anstatt eine 3) hat in diesem Sinne weitreichende Auswirkungen hinsichtlich der Analyse.

Somit kann auch eine quantitative Analyse nur als eine gewisse Annäherung an die Realität gesehen werden. Insofern ist der Einsatz von nominalen und ordinalen Merkmalen weniger ein Problem, weil keine Scheingenauigkeit produziert wird. Ein Mittelwert läßt sich auf - zig Stellen hinter dem Komma ausgeben und suggeriert eine Scheingenauigkeit, die in der Realität nicht existiert.

Somit geht es hier eher um ein Verständnis der gefundenen Segmente: erst mit dem Verständnis kann auch ein Segment gut charakterisiert werden. Ob nun die Fehlklassifikation bei 15 % oder bei 18 % liegt, ist einerlei - es handelt sich nur um eine Annäherung.

Der Vorteil von Entscheidungsbäumen gegenüber den bewährten Verfahren liegt in einem flexiblen Umgang mit den Skalenniveaus, da sie je nach Wertebereich unterschiedliche Maßzahlen zur Verfügung stellen. Auch statistische Laien können Entscheidungsbäume ohne Probleme interpretieren. Ein spezielles Wissen über Logarithmen, Exp(B)-Werte, Funktionen, etc., ist nicht nötig. Es müssen eher „einfachere“ statistische Maßzahlen wie Zusammenhänge verstanden werden, die durch tagtäglichen Umgang jedem präsent sind (z. B.: je

mehr man mit seinem Auto fährt, desto höher sind die Kosten für Benzin).

Die nachfolgende Tabelle gibt einen Überblick über die errechneten Maßzahlen zwischen den drei Verfahren.

TABELLE 12 VERGLEICH DER ERGEBNISSE DER FEHLKLASSIFIKATIONEN ZWISCHEN ENTSCHEIDUNGSBÄUMEN, LOGISTISCHER REGRESSION UND DISKRIMINANZANALYSE (FEHLKLASSIFIKATION, CRAMERS V, UNSICHERHEITSKOEFFIZIENT)

	Entscheidungsbäume	Logist. Regression	Diskriminanzanalyse
Cramers v	.551	.519	.471
Unsicherh.-koeffizient	.157	.217	.167
Fehlklassifikation	.26	.247	.268

Cramers v zeigt, dass sich bei den Entscheidungsbäumen mit 0.551 der höchste Zusammenhangswert, gefolgt von der logistischen Regression mit 0.519 ergibt. Die Diskrimianzanalyse liegt mit einem Wert von 0.471 leicht darunter. Der Unsicherheitskoeffizient als PRE-Maß liegt bei der logistischen Regression etwas höher (21.7 %) als die beiden anderen Verfahren (16 bzw. 17 %) - somit lassen sich bei der logistischen Regression bei Kenntnis der Kategorie der einen Variablen die andere Variable mit 21.7 % weniger Fehlern voraussagen.

Die Fehlklassifikation ist nahezu gleich hoch. Sie vergleicht die tatsächliche PC-Nutzung mit den durch die jeweiligen Verfahren gefundenen Lösungen. Rund 1/4 wird von jedem Verfahren hier falsch klassifiziert, d.h. PC-Nutzer werden irrtümlich als Nichtnutzer klassifiziert und umgekehrt.

Die Verfahren unterscheiden sich in ihren Ergebnissen nur graduell: ob nun eine Fehlklassifikation von rund .25 (logistische Regression) oder .26 (Entscheidungsbäume) bzw. .27 (Diskriminanzanalyse) gefunden wird, ist für die Güte eines Verfahrens bzw. für die Interpretation in den Sozialwissenschaften unerheblich. Es zeigt sich also, dass Entscheidungsbäume durchaus neben den etablierten Verfahren der logistischen Regression und der Diskriminanzanalyse bestehen können.

Der Informationsgehalt und auch die Anschaulichkeit der Entscheidungsbäume ist - in diesem Fall mit zwei unabhängigen Variablen - am höchsten:

- Es werden nicht nur auf die Kategorien der Zielvariablen bezogene Gruppen gebildet, sondern nach vielfältigen (unabhängigen) Variablen geclustert. So ergeben sich bei zwei unabhängigen Merkmalen mindestens vier Gruppen
- Jede gefundene Untergruppe wird anhand aller unabhängigen Variablen weiter segmentiert, so dass sich vielfältige Kombinationen (z. B. jüngere Frauen, ältere Theaterbesucher) ergeben können, die für ein klareres Schichtungsbild sorgen
- Die Gruppen werden anschaulich als Baumstruktur dargestellt
- Es müssen keine statistischen Interpretationen aufgrund von Funktionen oder vielfältigen Einzeltabellen erstellt werden, ein Blick auf den Baum und die Fehlklassifikation reichen in den meisten Fällen völlig aus
- Auch statistische Laien können Entscheidungsbäume schnell begreifen
- Entscheidungsbäume liefern ein ähnlich gutes Ergebnis
- Logistische Regression und Diskriminanzanalyse bilden aufgrund der unabhängigen Variablen Funktionen. Im Negativfalll bzw. bei nur zwei unabhängigen Variablen wird nur eine Funktion gefunden, die schwer interpretierbar ist

KAPITEL IV METHODISCHES VORGEHEN

1 Der EUROBAROMETER 56.0-Datensatz

1.1 Untersuchungssteckbrief und Beschreibung des Samples

Bei der nachfolgenden Untersuchung wurde der EUROBAROMETER 56.0-Datensatz des Zentralarchivs für empirische Sozialforschung in Köln herangezogen (vgl. http://www.gesis.org/za). Neben dem SPSS-Datensatz werden umfangreiche Informationen im pdf-Format mitgeliefert (Fragebogen, Codebuch, Untersuchungssteckbrief). Die EUROBAROMETER-Studien werden in regelmäßigen Abständen in vielen Ländern der Europäischen Union zu verschiedenen Themen erhoben. Die Schwerpunkte der vorliegenden Erhebung, die vom 22. August bis 27. September 2001 in Deutschland als standardisierte Interviews von INRA durchgeführt wurde, waren Fragen zu Informations- und Kommunikationstechnologie (incl. Computer- und Internetnutzung, Fragen 2 - 13), Finanzen (Fragen 14 - 28b), Kultur- und Freizeitaktivitäten (Fragen 34 - 50e) und soziodemografische Daten.[78] Da die Fragen zu Finanzen nicht weiter ausgewertet werden, wird auch auf eine Darstellung der Fragen bzw. Ergebnisse im folgenden verzichtet.[79]

Insgesamt wurden in Europa 16.200 Personen befragt, davon 2.047 in Deutschland, die als Grundlage dieser Untersuchung dienen. Von den 2.047 Befragten nutzen 984 den PC.

78. Die Fragen Q29 - Q33 entfielen in dieser Befragung, ebenso wie die soziodemografischen Fragen D2 bis D6, D9, D12 - D14. D20, D22 - D24 und D26 bis D28.

79. Inwieweit die Ergebnisse für diesen Teil der Befragung valide sind, ist schwierig zu beurteilen. Aus der Forschung ist jedoch bekannt, dass Fragen zu Einkommen und Finanzen häufig verweigert oder bewußt falsch beantwortet werden.

Für jedes Land wurde eine repräsentative mehrstufige Zufallsstichprobe von Personen gezogen, die 15 Jahre und älter waren. Die Bevölkerungsdichte, die z. B. in Deutschland von Bundesland zu Bundesland sehr unterschiedlich ist (z. B. Bayern ist das größte Flächenland, Nordrhein-Westfalen hat aber die größte Bevölkerungsdichte), wurden durch die Aufteilung Deutschlands in Metropolen, städtische und ländliche Gebiete berücksichtigt. Für jede dieser gefundenen Regionen wurde ein zufälliger Startpunkt ausgewählt. Mit dem Random-Route-Verfahren wurde jede n-te Adresse für die Befragung herangezogen. Das Random-Route-Verfahren legt Strategien der Zufallsauswahl fest. DIECKMANN (1995: 332) bemerkt:

> „Von diesen Startadressen ausgehend, werden sodann nach vorgegebenen Regeln die weiteren Adressen der Flächenstichprobe ermittelt. Die Regeln lauten z. B.: 'Gehen Sie von der Startadresse nach links bis zur nächsten Kreuzung. Dort biegen Sie rechts ab und bei der nächsten Querstraße wieder nach links, usw. Auf dem angegebenen Weg notieren Sie jeden sechsten Haushalt'."

Dieses Vorgehen ist nicht ganz unproblematisch, da es sehr viele Restriktionen und Vorarbeit voraussetzt (z. B. wie wird verfahren, wenn der sechste Haushalt eine Behörde, Schule oder ein Laden ist?). Werden diese Ausnahmen jedoch berücksichtigt, kann ein sehr gutes Sample gebildet werden, das die Qualität einer Primärerhebung in nahezu allen Fällen überlegen sein wird.

Ein Problem der Stichprobe ist die unproportionale (etwa gleichgewichtige) Verteilung von west- und ostdeutschen Befragten. Bei einem realen Verhältnis von ca. 1 : 3 wird deutlich, dass die Stichprobe in diesem Punkt nicht die Grundgesamtheit der bundesdeutschen Bevölkerung nach Regionen abbildet. Dies sollte aber - 15 Jahre nach der Wiedervereinigung - zumindest bei der PC-Nutzung keine Probleme aufwerfen, denn weder die Stichprobenqualität noch die Gruppe der Nichtnutzer wäre mit einer Internetbefragung erreicht worden.

1.2 Forschungsleitende Fragen

Aus den Kapiteln „Theoretischer Hintergrund“ und „Methodischer Hintergrund“ ergeben sich einige Prämissen für das Vorgehen.

Die Suche nach dominanten Schichtungen beginnt bei den herkömmlichen sozialstrukturellen Merkmalen, erweitert durch Alter und Geschlecht. Sie entscheiden grundsätzlich über die Lebenschancen und -möglichkeiten.

Kultur- und Freizeitvariablen sind der Ausdruck sozialstruktureller Möglichkeiten: sie können vielfältig sein und orientieren sich an den persönlichen Ressourcen (Bildung, Alter, Geschlecht, etc. - vor allem aber Haushaltsnettoeinkommen), denn Kultur- und Freizeitaktivitäten setzen häufig Ressourcen (z. B. finanzieller Art) voraus. Ein Arbeitslosengeld II-Empfänger wird sich kaum eine Opernpremierekarte finanziell leisten können - selbst wenn er ein großes Interesse an Opern hat. Ebenso wird er sich nicht leisten können, mehrmals wöchentlich ins Kino zu gehen, selbst wenn er ein großer Cineast wäre. Nicht die Freizeit- oder Kulturinteressen sind hier maßgebend, sondern der finanzielle Rahmen. Aus diesem Grund werden zuerst die sozialstrukturellen Variablen untersucht, die die Ressourcen ausmachen (Haushaltsnettoeinkommen, Bildung, Alter, Geschlecht, Berufliche Stellung, Ost-/West, ...). Nur sie können Ausdruck von dominanten Schichtungen sein. Die Wichtigkeit wird durch bivariate Zusammenhangsmaße mit der PC-Nutzung festgestellt: ist ein Zusammenhang hoch, kann er als dominantes Schichtkriterium herangezogen werden, Variablen mit niedrigen Zusammenhangswerten können eher gefundene Segmente näher beschreiben. Dahinter steht der Gedanke, dass multivariate Verfahren nach Zusammenhängen in einem Sample suchen - dass aber Subsamples ganz andere Zusammenhänge besitzen. So könnte es sein, dass es einen sehr hohen Zusammenhang zwischen PC-Nutzung und Alter gibt - Jüngere nutzen den PC häufiger als Ältere. Mit

einem multivariaten Verfahren ist es kein Problem, dies zu untersuchen. Nun könnte es jedoch sein, dass alle älteren Befragten mit Hochschulabschluss einen PC besitzen. Diese Information könnte bei einem multivariaten Verfahren in der Menge der Daten unentdeckt bleiben. Deshalb ist es in dieser Arbeit sehr wichtig, multivariat gefundene Zusammenhänge stets auch deskriptiv weiter zu untersuchen - was multivariate Verfahren nicht diskreditiert (ohne sie wäre es nicht möglich, Zusammenhänge zwischen unterschiedlichen unabhängigen Variablen, bezogen auf eine Zielvariable zu finden).

Durch die sozialstrukturellen dominanten Variablen können - hinsichtlich der Rechnernutzung - konkrete Gruppen gebildet werden (z. B. Hausfrauen bis 33 Jahre mit einem Haushaltsnettoeinkommen bis 2500 DM, die verheiratet sind mit einem Nutzeranteil von 25 %, weibliche Büroangestellte, ledig, bis 55 Jahre, bis 3000 DM Haushaltsnettoeinkommen mit einem PC-Nutzeranteil von 60 %). Diese Gruppen werden dann weiter nach Kultur- und Freizeitaktivitäten (z. B. Kino-, Oper- Theaterbesuch = subordinierte Schichtungen) charakterisiert, um ein besseres Schichtungsbild zu erhalten. Möglicherweise geht Gruppe 1 lieber ins Kino, hört Rock- und Popmusik, während die zweite Gruppe Volksmusik und Operetten bevorzugt und gerne ins Theater geht.

Die Wichtigkeit der sozialstrukturellen Variablen ergibt sich aus der Höhe des Zusammenhangs mit der PC-Nutzung, da bei Nominalskalenniveau keine Regel existiert, die einen Zusammenhang größer 0.2 als bedeutsam erkennt (wie bei metrischen Daten). Die Stärke des Zusammenhangs gibt die Wichtigkeit der unabhängigen Variablen, bezogen auf die PC-Nutzung an:

> „Chi-squared-based measures are difficult to interpret. Although they can be used to compare the strength of association in different tables, the strength of association being compared isn´t easily related to an intuitive concept of association.“ (NORUSIS (1998: 354))

Da Chi Quadrat-bezogene Maße Interpretationsprobleme bereiten, wird der Unsicherheitskoeffizient herangezogen. Allerdings wäre es unsinnig, Zusammenhangsmaße von 0.05 zu interpretieren. Aus diesem Grund werden in Überblickstabellen alle Zusammenhangsmaße > 0.1 angegeben - unabhängig davon, ob sie in eine weitere Analyse zu PC-Nutzung eingehen oder nicht. Ziel ist es, die Stärke der Zusammenhänge transparent zu machen, da es durchaus möglich ist, dass in der Gesamtstichprobe ein geringer Zusammenhang existiert, der aber in einer gefundenen Subgruppe deutlich höher liegt.

Nach einer ersten deskriptiven Untersuchung (mit Hilfe von Phi, CRAMERs v, Chi Quadrat und Unsicherheitskoeffizient) und einer Rangordnung nach Wichtigkeit der Koeffizienten werden die Variablen weiter multivariat mit den Entscheidungsbaumalgorithmen untersucht. Stand im Kapitel 3 die „Technik“ im Vordergrund, also wie die Generierung von Bäumen funktioniert (statistische Kennzahlen, Fehlklassifikationen, etc.), verfolgt dieser Teil der Untersuchung zweierlei: zum einen soll die Struktur und Verteilung der unabhängigen Variablen anhand der Algorithmen näher untersucht werden, um einen Überblick über die Verteilung zu erhalten, zum anderen, um zu zeigen, wo konkret die Grenzen jedes Algorithmus liegen. Fragestellung ist hier z. B.: Klassifizieren alle Algorithmen die einzelnen Variablen gleich oder unterschiedlich und welche Ableitungen ergeben sich daraus?

Entscheidungsbaumalgorithmen segmentieren hier konkrete Gruppen mit PC-Nutzeranteilen. Anhand der gefundenen Segmentierungen werden in einem zusätzlichen Schritt die Kultur- und Freizeitvariablen weiter beschrieben.

1.3 Fragen zur PC-Nutzung

Die Fragen zur Rechnernutzung mit den zugehörigen Skalenniveaus sind in der nachfolgenden Tabelle zusammengefaßt:

TABELLE 13 EUROBAROMETER 56.0: FRAGEN ZU KOMMUNIKATIONS- UND INFORMATIONSTECHNOLOGIE

Frage	Frage	Skalenniveau
Q 2	Wichtigkeit von Computern im Alltag (4stufig)	ordinal
Q 3	Ort der PC-Nutzung	nominal (MF)[a]
Q 4	Technische Mediennutzung der PC-Nutzer (z. B. Handy)	nominal-dichotom
Q 5	PC-, Internetnutzung: für welche Aufgaben?	nominal-dichotom
Q 6	Nachweise über PC-, Internetfähigkeiten	nominal-dichotom
Q 7	Computer-, PC-Schulung für Arbeit?	nominal-dichotom
Q 8	Weitere Frage zur letzten PC-Schulung	nominal-dichotom
Q 9 + 10	Telearbeit	nominal-polytom
Q 11a	Informationstechnologie am Arbeitsplatz	nominal-dichotom
Q 11b	Verbesserung/Verschlechterung durch die Einführung von Informationstechnologie am Arbeitsplatz	nominal-dichotom
Q 12	Veränderung der Arbeit durch Informationstechnologien	nominal-dichotom
Q 13	Betriebsgröße	ordinal

a. als Mehrfachantwort erfaßt

Nur zwei dieser Variablen sind ordinal, alle anderen nominal abgefragt. Wie schon weiter oben erläutert, hat dies weitreichende Auswirkungen auf die eingesetzten statistischen Verfahren.

Um die Baumalgorithmen, ihre Gemeinsamkeiten und Unterschiede herauszuarbeiten, wird in dieser Arbeit die Frage nach der grundsätz-

lichen PC-Nutzung für den nominalen, den ordinalen und den metrischen Fall untersucht - wobei die grundsätzliche PC-Nutzung (ja - nein) im Mittelpunkt steht. Möglicherweise muß in diesem Zusammenhang ein Informationsverlust in Kauf genommen werden. Für multivariate Verfahren sind jedoch größere Gruppen besser zu segmentieren als kleinere Subsamples. Deshalb wurden die Variablen „PC Nutzer ja - nein“ aus der Frage Q 39 (der Frage nach der Häufigkeit der Rechnernutzung) gebildet. Alle Ausprägungen ausser „nie“ und „weiss nicht“ wurden hierbei recodiert. Sicherlich gehen hierbei mehr Informationen verloren als bei einer ordinalen Variable mit fünf Abstufungen, es wird aber ein Problem multivariater Verfahren vermieden. In kleinen Gruppen lassen sich nur sehr schwer sinnvolle Segmente finden.

ABBILDUNG 82 Häufigkeit der PC-Nutzung - dichotom und 6stufig erfaßt (N = 2038)

Q 39: Häufigkeit: PC-Nutzung * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

Anzahl

		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)		
		Non User	User	Gesamt
Q 39: Häufigkeit: PC-Nutzung	nie	1054	0	1050
	seltener als einmal im Monat	0	40	40
	ein- bis dreimal im Monat	0	54	58
	einmal die Woche	0	108	108
	mehrmals die Woche	0	314	314
	täglich	0	468	468
Gesamt		1054	984	2038

Aus der Tabelle wird ersichtlich, dass für die PC-Nutzergruppen „seltener als einmal im Monat“ und „ein bis dreimal im Monat“ eine multi-

variate Analyse ausscheidet, da die Gruppen mit N = 40 bzw. 54 keine sinnvollen Segmente gefunden werden können. Beließe man die ordinalskalierte Variable, würden sich Fragen nach der Interpretationsfähigkeit stellen (z. B.: was unterscheidet das Segment der „seltener als einmal im Monat-User" von den „ein- bis dreimal im Monat" nutzenden Befragten?). Diese Gruppen könnten nur deskriptiv untersucht werden - und ob der Ertrag groß ist, ist zu bezweifeln, da keine generellen Ergebnisse (hinsichtlich Signifikanz) durch die geringe Fallzahl) zu erwarten sind.

Aus diesem Grund - und um die Möglichkeiten der Entscheidungsbäume darzustellen - soll neben der nominal skalierten (ja - nein) PC-Nutzung auch eine ordinale Variable gebildet werden - mit drei Ausprägungen: nie, gelegentlich und täglich. Diese Variable wird dann auch - unerlaubterweise - für den metrischen Fall herangezogen. Die Kreuztabelle sieht dann folgendermaßen aus:

ABBILDUNG 83 Häufigkeit der PC-Nutzung - dichotom und ordinal (3stufig) erfaßt (N = 2038)

Q 39: Häufigkeit: PC-Nutzung * Q 39: Häufigkeit der PC-Nutzung (Non-, Light-, Heavy) Kreuztabelle

Anzahl

		Q 39: Häufigkeit der PC-Nutzung (Non-, Light-, Heavy)			
		Non-User	Light User (bis mehrm. wöchentl.)	Heavy User (täglich)	Gesamt
Q 39: Häufigkeit: PC-Nutzung	nie	1050	0	0	1050
	seltener als einmal im Monat	0	40	0	40
	ein- bis dreimal im Monat	0	58	0	58
	einmal die Woche	0	108	0	108
	mehrmals die Woche	0	314	0	314
	täglich	0	0	468	468
Gesamt		1050	520	468	2038

Die Verteilung zeigt, dass nun alle Befragten, die den PC seltener als einmal im Monat bis mehrmals die Woche nutzen, zusammengefaßt wurden. Somit ergibt sich ein Verhältnis von 1050 Nichtnutzern : 520 Light Usern : 468 Heavy Usern (täglich).

1.4 Fragen zu Kultur- und Freizeitaktivitäten

Die Items zu den Kultur- und Freizeitaktivitäten bilden den umfangreichsten Teil der Befragung. Auch hier sind die meisten Variablen nominal skaliert, was eine multivariate Auswertung mit herkömmlichen Verfahren ausschließt. Eine Recodierung bzw. Zusammenfassung - gerade von dichotomen Variablen - ist auf den ersten Blick ebenfalls nicht möglich.

TABELLE 14 EUROBAROMETER 56.0: FRAGEN ZU KULTUR- UND FREIZEITAKTIVITÄTEN

Frage	Frage	Skalenniveau
Q 34	Fernsehen	nominal-dichotom
Q 35a	Arten von Fernsehsendungen	nominal-dichotom
Q 35b	Drei bevorzugte Arten von Fernsehsendungen	nominal (MF)
Q 36	Häufigkeit DVD/Videos sehen	ordinal
Q 37	Häufigkeit Radiohören	ordinal
Q 38	Arten von Radiosendungen	nominal-dichotom
Q 39	Häufigkeit der PC-Nutzung	ordinal
Q 40	Ort der PC-Nutzung	ordinal
Q 41a	Häufigkeit der Internetnutzung	ordinal
Q 41b	Nutzungsmuster	nominal-dichotom
Q 42a	Gelesene Bücher in den letzten 12 Monaten nach Beruf, Weiterbildung, andere Gründe	nominal (MF)
Q 42b	Anzahl der Bücher (bezogen auf Q 42a)	ordinal
Q 43	Häufigkeit Tageszeitung lesen	ordinal
Q 44	Häufigkeit Magazine/Zeitschriften lesen	ordinal
Q 45	Häufigkeit Musik hören	ordinal
Q 46a	Medien, mit denen Musik gehört wird	nominal-dichotom
Q 46b	Art von Musik	nominal-dichotom
Q 47	Kulturelle Aktivitäten	ordinal
Q 48	Konzertbesuch	nominal-dichotom
Q 49	eigene künstlerische Tätigkeiten	nominal (MF)
Q 50a	Medienbesitz (z. B. Walkman, CDs, ...)	nominal-dichotom
Q 50b	Anzahl der Fernsehgeräte im Haushalt	metrisch
Q 50c	Anzahl der Lexika in Buchform	metrisch
Q 50d	Anzahl der Lexika als CD-ROM	metrisch
Q 50e	Anzahl der Bücher	ordinal

1.5 Soziodemografische Fragen

In diesem Bereich lassen sich zahlreiche Recodierungen vornehmen. So kann das Alter, das offen erfragt wurde, sehr flexibel zusammen-

gefaßt werden. Die Familiensituation wurde sehr differenziert erfaßt, so dass sich der herkömmliche Familienstand (ledig, verheiratet, geschieden, verwitwet) und die Lebenssituation (ledig, lebe mit Partner zusammen, ledig, noch nie mit Partner zusammengelebt und ledig, früher mit Partner zusammengelebt) abbilden läßt. Die Kategorie „In zweiter Ehe lebend“ spielt im Rahmen dieser Untersuchung keine Rolle und wurde den Verheirateten zugerechnet.

Die berufliche Stellung - aus der letzten bzw. aktuellen Beschäftigung - läßt sich in die Kategorien 1 'nie erwerbstätig' 2 'sonstige (Fach)Arbeiter' 3 'sonst. Angest. Reise+Dienstl.' 4 'Meister' 5 'Ladenbesitzer, Handwerker' 6 'Landwirte, Fischer' 7 'sonstige Bürotätigkeiten' 8 'Büroangest. mit Leitungsfunktion' 9 'Freie Berufe', 10 'Grossunternehmer, Direktoren, Top Management, Angestellte mit Leitungsfunktion' und 11 'Student'. zusammenfassen[80]. Es wäre durchaus möglich, weitere Recodierungen vorzunehmen - jedoch ist es durchaus interessant, die einzelnen Berufe bzw. Berufsgruppen durch die Algorithmen zusammenfassen zu lassen. Anders als bei der Recodierung der abhängigen Variablen PC-Nutzung kann auf Seiten der unabhängigen Variablen auch Zusammenfassungen vorkommen, die so von mir nicht erwartet wurden. Wenn eine zu starke Eingrenzung von Forscherseite erfolgt, kann dies auch zu Verzerrungen führen.

Das Haushaltsnettoeinkommen - abgefragt in 250-DM-Schritten ab 1.500 DM, ab 3.000 DM in 500-DM-Schritten bis 5000 DM - läßt leider keinen Schluß auf Geringverdiener zu. Diese Variable ist aber in den meisten Fällen abhängig von Bildung und beruflicher Stellung.[81]

Aus der Postleitzahl bzw. der gebildeten Variable „Regionen“ können die Zuordnung zu West- und Ostdeutschland, auch mit Ortsgrößen

80. Die Büroangestellten mit Leitungsfunktion beziehen sich eher auf die mittlere Führungsebene (Gruppenleiter, Abteilungsleiter), während Angestellte mit Leitungsfunktion zum Top-Management gezählt werden können.

zugeordnet werden. Frage D 25 erhebt Informationen zum Stadt - Land - Verhältnis.

TABELLE 15 EUROBAROMETER 56.0: SOZIODEMOGRAFISCHE DATEN

Frage	Frage	Skalenniveau
A	Haushaltsgröße	metrisch
B	Haushaltsmitglieder über 15 Jahre	metrisch
C	Vornamen der Personen, die älter als 15 Jahre sind	nominal-polytom
Q 1	Land, in der die Befragung stattfand	nominal-polytom
D 1	Links - Rechts - Selbsteinschätzung (10stufig)	metrisch
D 7	Familiensituation (Familienstand bzw. Lebenssituation)	nominal-polytom
D 8	Alter bei Schulabgang	metrisch
D 10	Geschlecht	nominal-dicho-tom
D 11	Alter	metrisch
D 15s	eigene Berufstätigkeit	nominal-polytom
D 15b	ausgeübter Beruf	nominal-polytom
D 19	„Haushaltsvorstand“ - diejenige Person, die am meisten zum Haushaltseinkommen beiträgt	nominal-dicho-tom
D 21a	Berufstätigkeit der Person, die am meisten zum Haushalts-einkommen beiträgt?	nominal-polytom
D 21b	Beruf der Person, die am meisten zum Haushaltseinkom-men beiträgt	nominal-polytom
D 25	Wohnort (Dorf, Kleinstadt, ...)	nominal-polytom
D 29	Haushaltsnettoeinkommen (in DM)	ordinal
D 32	Telefonbesitz	nominal-dicho-tom
P 6	Ortsgröße	ordinal
P 7	Postleitzahl	nominal-polytom

81. Da die Befragung im Jahr 2001 durchgeführt wurde, sind die Angaben im Datensatz in DM angegeben. Auf eine Umrechnung in Euro wurde bewußt im Rahmen der Arbeit verzichtet, da das Einkommen ordinal in Klassen erfaßt wurde und somit eine exakte Umrechnung zu einem Kurs von 1.95 DM/Euro zu größeren Verzerrungen - vor allem bei den höheren Gehaltsgruppen und der Skala - führen würde.

2 Deskriptive Beschreibung der soziodemografischen Variablen

Von den 2.047 Befragten sind 984 PC-User, wovon 671 auch das Internet nutzen.[82] Der Anteil der Internetuser liegt bei ca. 1/3 der befragten Personen. Bei den PC-Nutzern liegt er bei knapp unter 50 %. Würde man an dieser Stelle fortfahren, dann bleibt die Hälfte des Datensatzes - nämlich die Nichtnutzer - unberücksichtigt. Diese Gruppe kann aber als „Kontrastgruppe“ wichtige Erkenntnisse liefern.

Nachfolgend werden alle soziodemografischen Variablen als Zusammenhangsmaße (Phi, Cramers v > 0.1, Unsicherheitskoeffizient) dargestellt:

82. Der Internetzugang erfolgt ausschließlich über den PC.

TABELLE 16 WICHTIGE BIVARIATE ZUSAMMENHÄNGE (> 0.1) ZWISCHEN PC-NUTZUNG UND DEN SOZIALSTRUKTURELLEN VARIABLEN (PHI, CRAMERS V (= V), ETA, UNSICHERHEITSKOEFFIZIENT (= U))[A]

unabhängige Variable	Skalenniveau	Phi / Cramers v / Eta^2 (Unsicherheitskoeff.)
Links - Rechts - Selbsteinschätzung (10stufig)	ordinal	v = .099 (u = 0.07)
Familiensituation (Lebenssituation)[b] Familiensituation (Familienstand)	nominal-polytom nominal-polytom	v = 0.375 (u = 0.116) v = 0.368 (u = 0.112)
Alter des Schulabgangs Bildungsabschluss	metrisch ordinal	eta^2 = 0.20 v = 0.411 (u = 0.134)
Geschlecht	nominal-dichotom	Phi = -0.100 (u = .007)
Alter (15 - 94 Jahre)	metrisch	eta^2 = 0.294
ausgeübter Beruf (11stufig) ausgeübter Beruf (Büro - andere)	nominal-polytom nominal-dichotom	v = 0.441 (u = 0.152) Phi = 0.402 (u = 0.120)
Ortsgröße	nominal-polytom	v = 0.093 (u = 0.006)[c]
Haushaltsnettoeinkommen (in DM)	ordinal	v = 0.355 (u = 0.094)

a. Variablen, bei denen ein Zusammenhangsmaß unsinnig ist (z. B. Vornamen) wurden nicht berechnet und nicht in die Tabelle aufgenommen.
b. Der Familienstand wird differenziert nach „ledig, verheiratet, geschieden verwitwet“, die Familiensituation differenziert weiter nach „ledig, nie mit Partner zusammengelebt“, „ledig mit Partner“, „ledig, aber früher mit Partner“, verheiratet“, „getrennt lebend“, „geschieden“, verwitwet“.
c. Sig = 0.008 bzw. 0.007.

Als höchste bivariate Zusammenhänge dominieren das Alter (eta^2 = 0.294), der Bildungsabschluss (Cramers v = 0.411), die berufliche Stellung (Cramers v: rund 0.4), die Lebenssituation bzw. der Familienstand (Cramers v: rund 0.37) und das Haushaltsnettoeinkommen (Cramers v = 0.355) - Variablen, die man auch erwarten würde. Allerdings gibt es auch einige Überraschungen: die Annahme, dass Frauen deutlich weniger den PC nutzen als die Männer, hat sich nicht

bestätigt (Phi = -0.1) - nach anderen Studien verwenden sie ihn aber anders (vgl. BÜHL (1999: 35f.)):

> „Es existieren deutliche Tendenzen in Richtung eines geschlechtsspezifischen Umgangs mit dem Computer, aber keine der beschriebenen Verhaltensweisen wird ausschließlich von Männern oder Frauen an den Tag gelegt. In allen Gruppen zeigt sich aber, dass Männer häufig einen lustbetonten Zugang zum Computer besitzen als Frauen, welche die negativen Seiten der Computertechnik deutlicher vor Augen haben."

Dieser Schluß findet sich im EUROBAROMETER-Datensatz nicht - oder nur als ganz leichter „Trend":

ABBILDUNG 84 PC-Nutzung: Kreuztabelle zwischen Geschlecht und rein beruflicher, rein privater und beruflicher/privater Nutzung (N = 1990, Spalten-%

Berufliche - private PC-Nutzung * D10 Geschlecht * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)				D10 Geschlecht männlich	D10 Geschlecht weiblich	Gesamt
Non User	Berufliche - private PC-Nutzung	keine Nuzung	Anzahl	445	609	1054
			% von D10 Geschlecht	100,0%	100,0%	100,0%
	Gesamt		Anzahl	445	609	1054
			% von D10 Geschlecht	100,0%	100,0%	100,0%
User	Berufliche - private PC-Nutzung	berufliche Nutzung	Anzahl	128	175	303
			% von D10 Geschlecht	26,2%	39,1%	32,4%
		private Nutzung	Anzahl	105	94	199
			% von D10 Geschlecht	21,5%	21,0%	21,3%
		berufl./private Nutzung	Anzahl	255	179	434
			% von D10 Geschlecht	52,3%	40,0%	46,4%
	Gesamt		Anzahl	488	448	936
			% von D10 Geschlecht	100,0%	100,0%	100,0%

Während Frauen den PC rund 1/3 häufiger beruflich nutzen (39 vs. 26 %), gibt es keine Unterschiede zwischen den Geschlechtergruppen bei der rein privaten Nutzung. Dafür nutzen Männer den PC häufiger beruflich und privat (52 % vs. 40 % Frauen). Daraus läßt sich schließen, dass diejenigen Frauen, die beruflich mit dem PC zu tun haben, häufig auf einen privaten Einsatz verzichten - die Anteile zwischen beruflicher und beruflicher und privater Nutzung sind nahezu gleich hoch.

ABBILDUNG 85 PC-Nutzung: Kreuztabelle zwischen Geschlecht und rein beruflicher, rein privater und beruflicher/privater Nutzung (N = 1990, standardisierte Residuen)

Berufliche - private PC-Nutzung * D10 Geschlecht * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

Q 39: Häufigkeit der PC-Nutzung				D10 Geschlecht männlich	D10 Geschlecht weiblich	Gesamt
Non User	Berufliche - private PC-Nutzung	keine Nuzung	Anzahl	445	609	1054
			Erwartete Anzahl	445,0	609,0	1054,0
			Standardisierte Residuen	,0	,0	
	Gesamt		Anzahl	445	609	1054
			Erwartete Anzahl	445,0	609,0	1054,0
User	Berufliche - private PC-Nutzung	berufliche Nutzung	Anzahl	128	175	303
			Erwartete Anzahl	158,0	145,0	303,0
			Standardisierte Residuen	-2,4	2,5	
		private Nutzung	Anzahl	105	94	199
			Erwartete Anzahl	103,8	95,2	199,0
			Standardisierte Residuen	,1	-,1	
		berufl./private Nutzung	Anzahl	255	179	434
			Erwartete Anzahl	226,3	207,7	434,0
			Standardisierte Residuen	1,9	-2,0	
	Gesamt		Anzahl	488	448	936
			Erwartete Anzahl	488,0	448,0	936,0

Der Anteil der Nutzer bzw. Nichtnutzer unterscheidet sich hinsichtlich des Geschlechts nicht sehr (standardisierte Residuen: 0) - d. h., ohne die empirische Untersuchung wäre ein ähnliches Ergebnis erwartet worden. Während die berufliche Nutzung bei den Frauen mit 2.5 deutlich übewiegt (eigentlich werden weniger Frauen erwartet, die den PC beruflich nutzen), ist es bei den Männern umgekehrt: es werden hier mehr erwartet. Die rein private Nutzung ist mit -/+ 0.1 recht ausgeglichen. Allerdings nutzen viele Frauen den PC nicht beruflich und privat, dafür jedoch die Männer.

Dies zeigt, dass Frauen den PC nüchterner ansehen und ihn eher als Arbeitsgerät gebrauchen, weniger für private Zwecke wie die Männer. Das läßt sich einfach über das fehlende Technikinteresse von Frauen erklären.

Der Zusammenhangswert fällt mit 0.144 (Cramers v) bzw. 0.01 (Unsicherheitskoeffizient) nicht gerade sehr stark aus. Es kann also weder von einer reinen „Männerdomäne PC" noch von einem völlig anderen Zugang (Frauen rein beruflich, Männer rein privat) ausgegangen werden.

Auch läßt sich im EUROBAROMETER-Datensatz das Ergebnis, dass PC-Autodidakten einen deutlich höheren Anteil von Männern aufweisen, nicht bestätigen.

ABBILDUNG 86 Geschlechtsspezifische PC-Qualifikation (N = 2047)

Q6: PC-Qualifikation (4stufig) * D10 Geschlecht Kreuztabelle

			D10 Geschlecht		
			männlich	weiblich	Gesamt
Q6: PC-Qualifikation (4stufig)	keine	Anzahl	432	582	1014
		% von Q6: PC-Qualifikation (4stufig)	42,6%	57,4%	100,0%
	(Weiter-)Bildung	Anzahl	70	63	133
		% von Q6: PC-Qualifikation (4stufig)	52,6%	47,4%	100,0%
	Beruf	Anzahl	162	177	339
		% von Q6: PC-Qualifikation (4stufig)	47,8%	52,2%	100,0%
	autodidaktisch	Anzahl	298	263	561
		% von Q6: PC-Qualifikation (4stufig)	53,1%	46,9%	100,0%
Gesamt		Anzahl	962	1085	2047
		% von Q6: PC-Qualifikation (4stufig)	47,0%	53,0%	100,0%

Deutlich wird, dass sich die weiblichen nicht wesentlich von den männlichen Befragten hinsichtlich der PC-Qualifikationsart unterscheiden. Dies ist vielleicht ein Ergebnis, das sich erst in den letzten Jahren herausgebildet hat. Die Zusammenhangsmaße müssen demnach mit 0.094 (Cramers v) bzw. 0.004 (Unsicherheitskoeffizient) sehr gering ausfallen. Hierbei wird deutlich, dass die generelle PC-Nutzung - keine geschlechtsspezifische Auswirkung (mehr) besitzt. Diese Variable kann folglich nicht als dominantes Kriterium für die allgemeine Neigung diese Technologie (nicht) zu nutzen, herangezogen werden. Es wird sich zeigen, ob dieses Merkmal in Subgruppen deutlicher hervortritt.

Auch hätte man erwarten können, dass bei der PC-Nutzung die Ortsgröße eine gewisse Rolle spielt und sie in Städten deutlich höher liegt. Dies hat sich ebenfalls nicht bestätigt (Cramers v = 0.09).

Somit kristallisieren sich fünf mögliche dominante Variablen für eine multivariate Analyse heraus: Alter, Bildungsabschluss, Beruf, Haushaltsnettoeinkommen und Familienstand.

Allerdings können die Zusammenhangsmaße nicht isoliert voneinander betrachtet werden - es ergibt sich eine deutliche Struktur. Alter und Geschlecht sind unabhängige Variablen. Der Bildungsabschluss ist abhängig vom Alter (Bildungsexpansion), die berufliche Stellung abhängig vom Bildungsabschluss (und indirekt vom Alter), das Haushaltsnettoeinkommen von Beruf, Bildungsabschluss und Alter - teilweise sicherlich auch vom Familienstand.

TABELLE 17 BIVARIATE ZUSAMMENHANGSMASSE (ETA^2 , R^2 FÜR ALTER, CRAMERS V BZW UNSICHERHEITSKOEFFIZIENT FÜR BILDUNG, BERUF, FAMILIENSTAND UND HAUSHALTSNETTOEINKOMMEN, N = 2038)

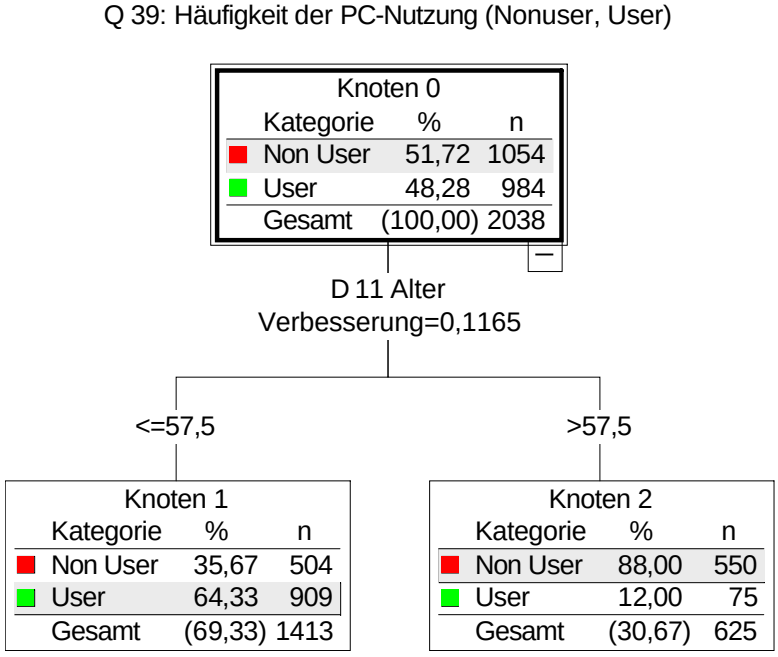
	Alter	Bildung	Beruf	Familienstand
Haushaltsnettoeinkommen	$e^2 = 0.16$ $r^2 = -0.111$	v = 0.214 u = 0.050	v = 0.163 u = 0.050	v = 0.289 u = 0.075
Alter	1.000	$e^2 = 0.10$	$e^2 = 0.13$	$e^2 = 0.36$
Bildung		1.000	v = 0.556 u = 0.169	v = 0.244 u = 0.093
Beruf			1.000	v = 0.235 u = 0.090

Bildung und Beruf weisen mit v = 0.556 die höchsten Zusammenhänge mit den restlichen Variablen auf - was auch realistisch ist. Dies wirft für die Analyse einige Probleme auf: wie soll mit den Variablen umgegangen werden, die eindeutig voneinander abhängig sind? Aus diesem Grund ist es sinnvoll, die innere Struktur der Variablen hinsichtlich PC-Nutzung näher zu betrachten. Die nominalen Zusammenhangsmaße geben hier nur Auskunft über die Stärke, aber nicht über die Richtung des Zusammenhangs.

Vor allem der Bildungsgrad steht mit dem Beruf in deutlichem Zusammenhang (über 0.5). Somit macht es wenig Sinn, diese Variablen gemeinsam multivariat zu untersuchen. Da das Alter die einzig unabhängige Variable ist, muß sie auf jeden Fall berücksichtigt werden - deshalb ist es sinnvoll, die vom Alter abhängige Variable Familienstand für die multivariate Analyse auszuschließen, da sie zu einer deutlichen Verzerrung führt. Die Entscheidung über Bildungsgrad und Beruf soll weiter unten empirisch erfolgen.

Im Kapitel 3 wurden ausführlich die Variablen Alter und Bildung als unabhängige Variablen herangezogen. Das Alter hatte eine überragende Erklärungskraft bei der PC-Nutzung - weit vor allen anderen unabhängigen Variablen. Ein nochmaliger Blick auf den CART-Ursprungsbaum, der das Alter dichotomisiert, ernüchtert allerdings etwas:[83]

ABBILDUNG 87 CART-Entscheidungsbaum: PC-Nutzung nach Alter (einstufig)



Die Trennung erfolgt - wie bekannt - bei rund 58 Jahren (was in etwa den Übergang ins Rentenalter charakterisiert). Während im Knoten 1 etwa 2/3 der Befragten den PC nutzen, sind es bei den Älteren nur 12

83. Es wäre an dieser Stelle auch möglich, den Binärbaumalgorithmus von QUEST für die Erläuterung heranzuziehen. Die Trennung erfolgt nicht bei rund 58, sondern bei rund 54 Jahren. QUEST liefert aber, wie sich zeigen wird, für dieses Beispiel die empirisch und auch theoretisch nicht ganz so effiziente Lösung wie CART - dies liegt zum einen an den bei QUEST etwas schlechteren Kennziffern, zum anderen an der theoretischen Begründbarkeit: 58 Jahre kann eindeutig als Übergang in das Rentenalter angesehen werden. Allerdings kann eine Segmentierung bei 54 Jahren, wie sie QUEST vornimmt, auf Gefährdungen und Probleme im Übergang der letzten Erwerbstätigkeitsphase in den Ruhestand gewertet werden.

%. Das wirft die Frage auf, ob die älteren Befragten die Dominanz des Alters der Rechnernutzung nicht erheblich erhöhen.

Die Bedeutung des PCs für ältere Menschen scheint eher marginal zu sein: es gibt keine beruflichen Anforderungen in diese Richtungen - und auch privat häufig keinen Anlaß, einen Rechner einzusetzen. Das bedeutet keinesfalls, dass ältere Menschen nicht in der Lage wären, mit einem Rechner umzugehen, aber neben den fehlenden Motiven und evtl. motorischen Schwierigkeiten ist es kein Thema in diesem Alterssegment.

Nimmt man die empirische Answertree-Trennung von 57.5 Jahren, so läßt sich dies theoretisch mit dem Renteneintrittsalter erläutern: die Jüngeren nutzen eher den PC, die Älteren nicht. Schließt man die über 57jährigen aus der Analyse aus, ergibt sich ein völlig neues Bild:

TABELLE 18 VERGLEICH DER ZUSAMMENHANGSWERTE (PHI, CRAMERS V (IN KLAMMERN: UNSICHERHEITSKOEFFIZIENT)) DER GESAMTSTICHPROBE (N = 2.047) MIT DER STICHPROBE DER BIS 57JÄHRIGEN (N = 1413)

unabhängige Variable	N = 2048	Rang	N = 1413	Rang
Alter	v = 0.542 u = 0.248 eta^2 = 0.29	1 1	v = 0.268 u = 0.058 eta^2 = 0.07	3 3
Berufsgruppen	v = 0.441 u = 0.152	2 2	v = 0.502 u = 0.212	1 1
Schulbildung	v = 0.411 u = 0.134	3 3	v = 0.331 u = 0.095	2 2
Familienstand	v = 0.368 u = 0.112	4 4	v = 0.188 u = 0.028	5 5
Haushaltsnettoeinkommen	v = 0.355 u = 0.094	5 5	v = 0.259 u = 0.054	4 4
Geschlecht	Phi = -0.100 u = 0.007	6 6	Phi = - 0.074 u = 0.004	6 6

Die Gesamtstichprobe unterscheidet sich in der Reihenfolge der Zusammenhänge deutlich: findet sich bei allen Befragten der höchste Zusammenhangswert beim Alter, so ist es bei den Jüngeren der Beruf - gefolgt von der Schulbildung. Dies könnte bedeuten, dass der Bildungsabschluss zwar einen gewissen Einfluß auf die Rechnernutzung hat, der Umgang mit dem PC am Arbeitsplatz jedoch wichtiger ist.

Da eta^2 ein PRE-Maß ist, wurde, der besseren Vergleichbarkeit halber, für das Alter auch Cramers v ausgewiesen.[84] Das Alter, das in der Gesamtstichprobe den höchsten Zusammenhangswert mit der PC-Nutzung aufwies, ist auf Platz 3, also hinter Bildung und beruflicher Stellung, „zurückgefallen". Auch der Familienstand hat in der Subpopulation keinen so hohen Einfluß mehr. Die Geschlechtsvariable ist nach wie vor nicht bedeutsam.

2.1 Alter

Untersuchungsgegenstand ist, ob es eher einen kontinuierlichen Effekt (steigendes Alter - abnehmende PC-Nutzung) gibt - oder ob der Effekt einen anderen Verlauf nimmt (z. B. hoher Anteil bei den jüngsten Befragten, noch höherer Anteil bei den Berufstätigen, geringer Anteil bei den Älteren). Hypothese für diese Verteilung wäre, dass der Beruf deutlich die PC-Nutzung dominiert.

Nachfolgend werden die Ergebnisse der Alterssegmentierungen anhand des EXHAUSTIVE CHAID und des QUEST-Algorithmus vorgestellt, die sich aufgrund ihrer Kennzahlen ähneln. Danach wird die CART-

84. Es gibt zwei Möglichkeiten des Vergleichs: entweder wird eta^2 mit dem Unsicherheitskoeffizienten direkt verglichen - was zu Verzerrungen führen muss, da die Voraussetzungen zwischen den beiden Kennzahlen (z. B. unterstellte Normalverteilung bei eta^2 oder die Tatsache, dass der Unsicherheitskoeffizient grundsätzlich niedriger ist als eta^2 - was hier auch zutrifft. Deshalb wurde die Rangfolge nach der Höhe der jeweiligen Unsicherheitskoeffizienten gebildet - eine mathematisch zulässige, aber nicht so exakte Messung.

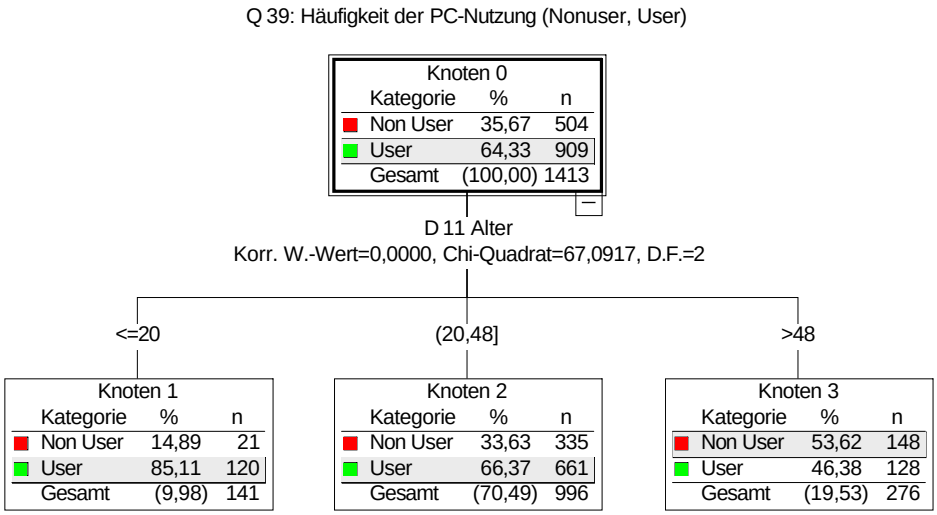
Lösung vorgestellt. Der CHAID-Algorithmus bleibt, aufgrund der gleichen bzw. häufig schlechteren Lösung mit EXHAUSTIVE CHAID, unberücksichtigt.

2.1.1 Alterssegmentierung mit EXHAUSTIVE CHAID

(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)

Die Segmentierung ergibt folgendes Bild:

ABBILDUNG 88 EXHAUSTIVE CHAID: PC-Nutzung nach Alter



Der EXHAUSTIVE-CHAID-Algorithmus segmentiert drei Gruppen, die Nutzeranteile zwischen 46 % (älter als 48 Jahre) bis 85 % (bis einschließlich 20jährige) aufweisen. Die 20- bis 48jährigen liegen mit einem Useranteil von etwa 2/3 zwischen den beiden anderen Gruppen. Somit sinkt die Nutzung kontinuierlich mit steigendem Alter. Allerdings lassen sich keine weiteren Unterknoten segmentieren - und die Fallzahlen der Altersgruppen sind sehr unterschiedlich (141 : 996 :

276). Damit werden die Möglichkeiten, aber auch die Grenzen dieses Algorithmus verdeutlicht.

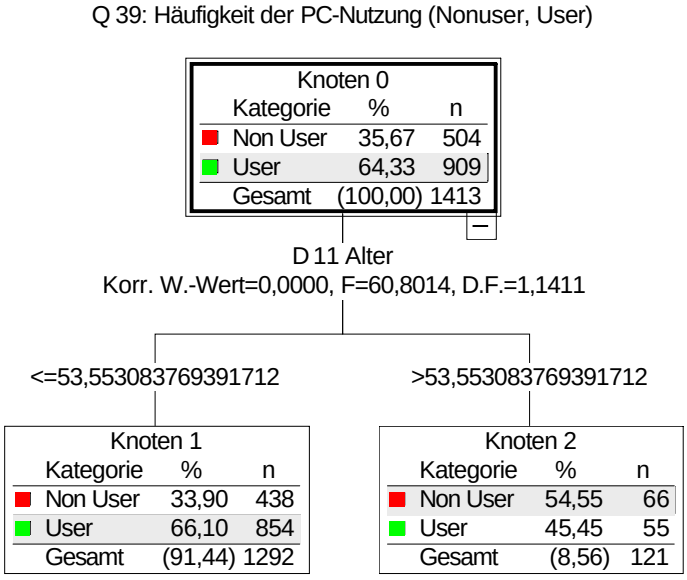
Hier wird wieder deutlich, dass in bestimmten Fällen die eher breiteren Bäume, die die CHAID-Algorithmen generieren, zu keinen weiteren Segmenten mehr beitragen können. Dafür kann es passieren, dass durch die Binärbaumalgorithmen QUEST und CART wichtige Cluster unberücksichtigt bleiben, wenn keine weiteren Stufen segmentiert werden können.

2.1.2 Alterssegmentierung mit QUEST

QUEST	• vieldiskutierter, relativ neuer Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus (unabhängige Variablen) geeignet • nur für nominale Zielvariablen geeignet, die auch dichotom sein kann • gebräuchliche statistische Kennzahlen (Chi-Quadrat, F-Test) • erlaubt Ersatzprädiktoren, Pruning und a prioris

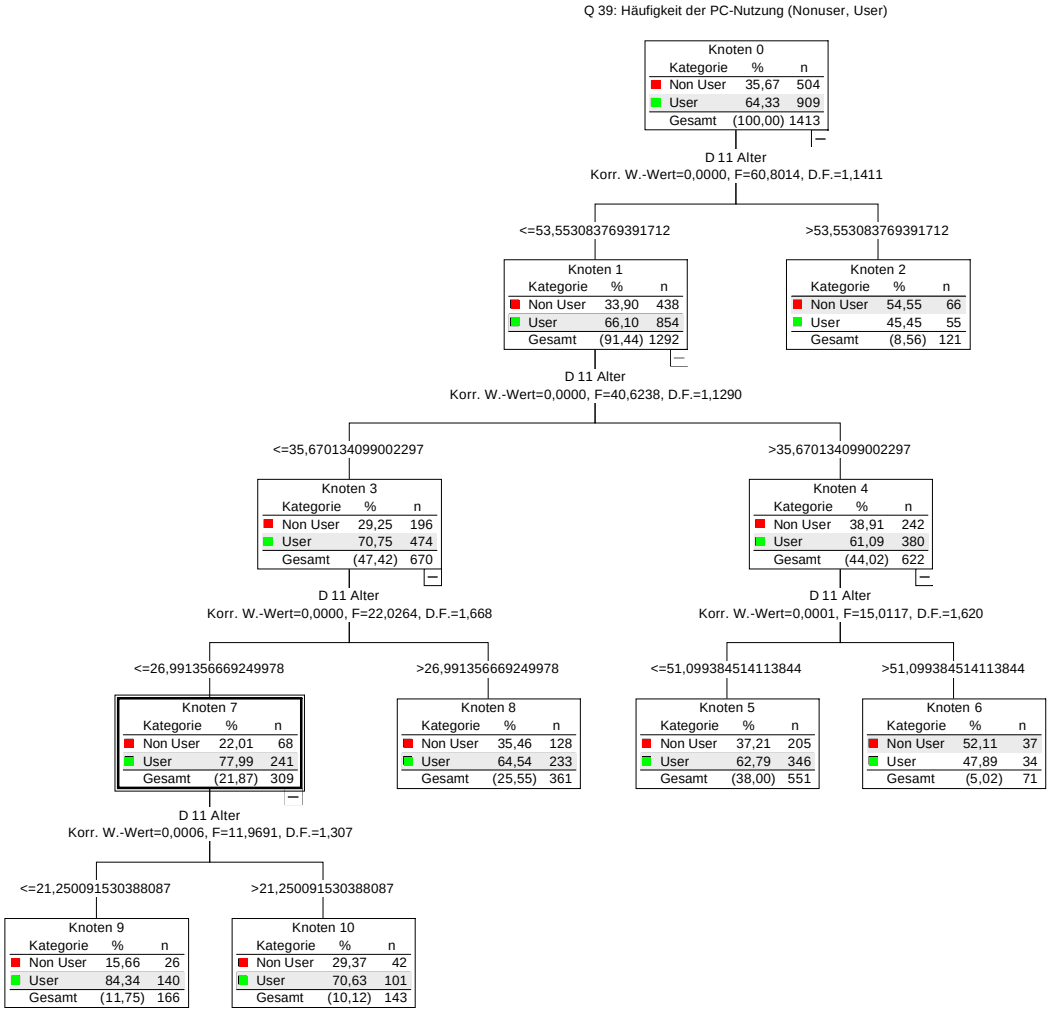
Die Ergebnisse des QUEST-Algorithmus werden stufenweise dargestellt:

ABBILDUNG 89 QUEST: Einstufige Alterssegmentierung bei den jüngeren Befragten (bis 57 Jahre, N = 1413)



Während CART (siehe nächstes Unterkapitel) die Trennung bei 49 Jahren vornimmt, erfolgt sie bei QUEST bei 54. Die Useranteile unterscheiden sich allerdings nicht wesentlich (etwa 2 - 3 Prozent).

ABBILDUNG 90 QUEST: Mehrstufige Alterssegmentierung bei den jüngeren Befragten (bis 57 Jahre, N = 1413)



Knoten 2 der über 53jährigen kann nicht weiter segmentiert werden. Die im Knoten 3 enthaltenen über 35jährigen mit einem Useranteil von 71 % verteilen sich auf die 27- bis 35jährigen mit einem Nutzeranteil von rund 64 % (Knoten 8) und den Jüngeren mit 78 % (Knoten 7),. Diese werden nochmals weiter in die unter 21jährigen Knoten 9: 84 % Nutzeranteil) und 10 (71 % User) aufgesplittet. Auch in diesem Beispiel segmentiert QUEST deutlicher die Nichtnutzer als CART.

Festzuhalten bleibt, dass die PC-Nutzung kontinuierlich mit dem Alter abnimmt: ob man nun eine Trennung bei 21 (EXHAUSTIVE CHAID, QUEST) oder bei 22 Jahren (CART) vornimmt, ist in diesem Falle nicht so entscheidend. Vielmehr zeigt sich, dass das die gleiche Trennung bei 21 Jahren möglicherweise auf den EXHAUSTIVE CHAID und QUEST beruhenden Chi-Quadrat-Wert zurückzuführen ist. Der Zusammenhang ist somit - weitestgehend - monoton.

Bei der Trennung der älteren Befragten gibt es jedoch deutliche Unterschiede: hier segmentieren EXHAUSTIVE CHAID und CART die Gruppen bei 48 Jahren, QUEST bei rund 54 Jahren - mit dem Ziel, ältere Nichtnutzer aus der weiteren Analyse auszuschließen. Anscheinend stellen jedoch die 48 - 54jährigen keine so inhomogene Gruppe dar, dass sich bessere Segmente finden lassen, so dass QUEST zu keiner sehr überzeugenden Lösung kommt.

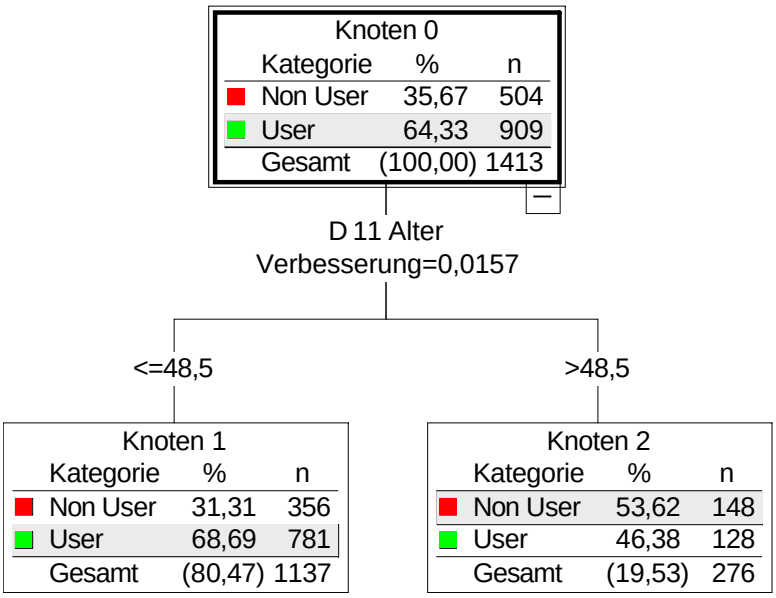
2.1.3 Alterssegmentierung mit CART

CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und a prioris

Die Trennung des Alters erfolgt hier nicht, wie in der Gesamtstichprobe bei 57.5 Jahren, sondern bei 48.5 Jahren. Das Nutzerverhältnis der Jüngeren zu den Älteren betragt rund 69 : 46.

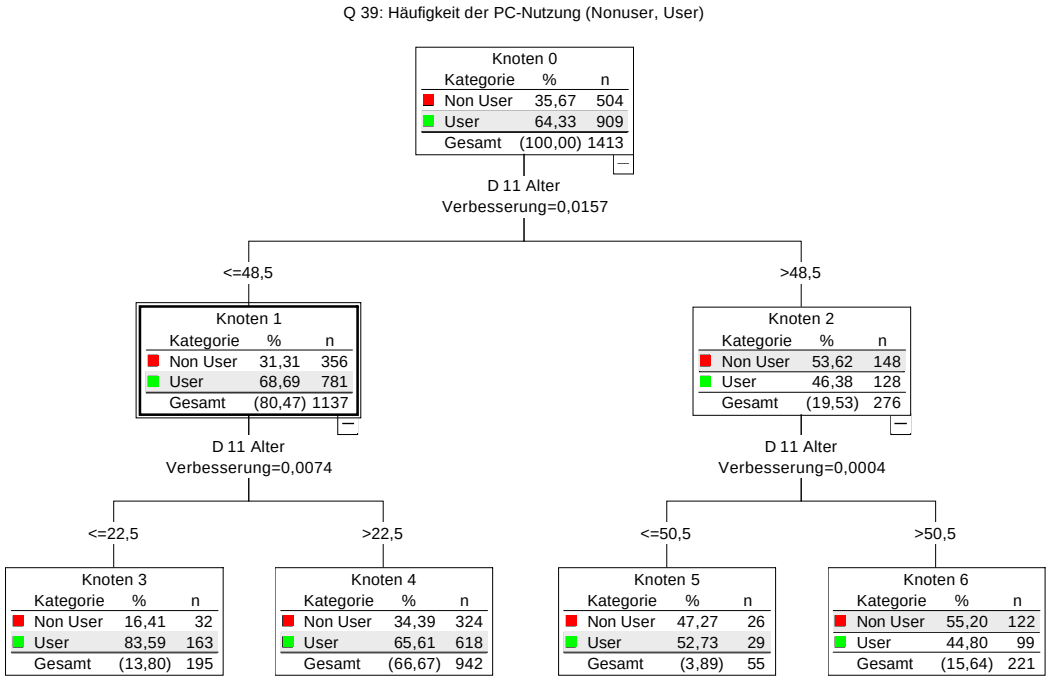
ABBILDUNG 91 CART: Einstufige Alterssegmentierung bei den jüngeren Befragten (bis 57 Jahre, N = 1413)

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)



Eine weitere Stufe des Baums ergibt folgendes Ergebnis:

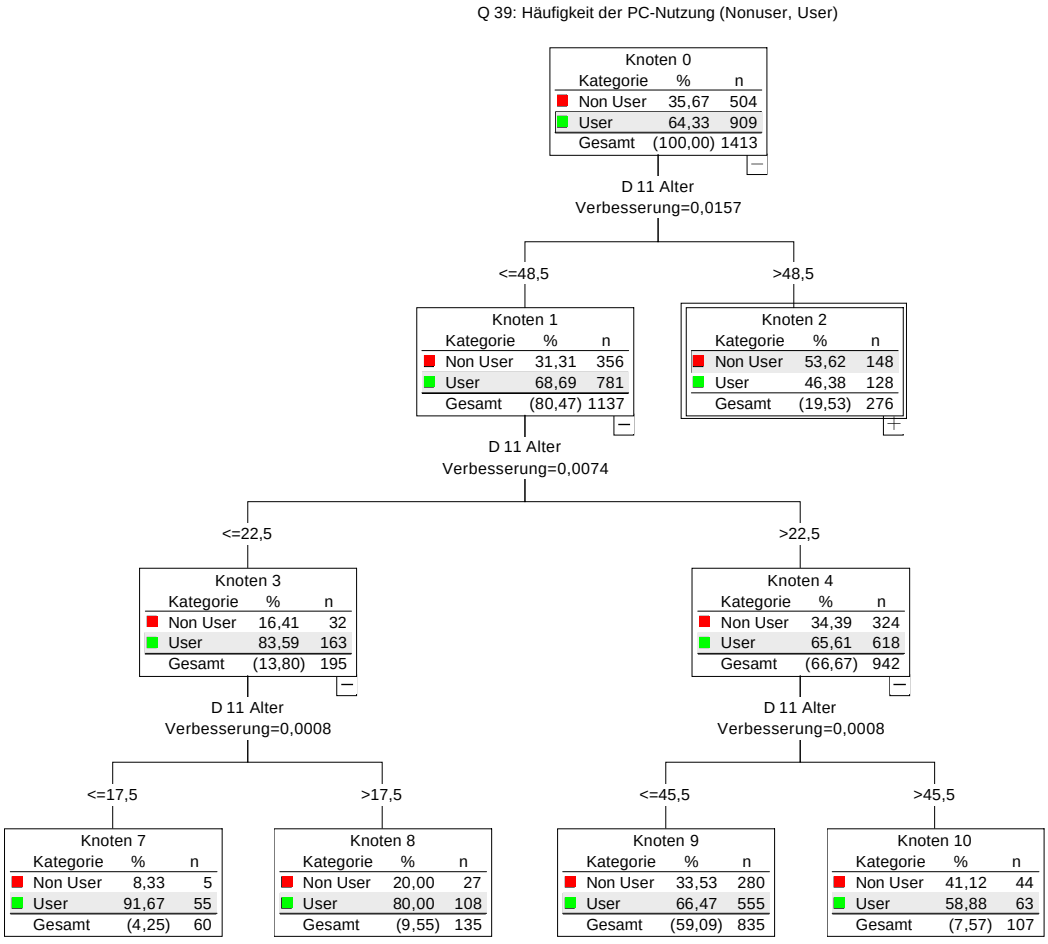
ABBILDUNG 92 Zweistufige Alterssegmentierung mit CART bei den jüngeren Befragten (bis 57 Jahre, N = 1413)



Auf der zweiten Ebene wird deutlich, dass die Anteile der PC-Nutzer kontinuierlich von Knoten 3 bis Knoten 6 abnehmen - von 84 % bis 45 %. Dabei ist die Prozentsatzdifferenz in Knoten 3 zu Knoten 4 mit knapp 20 % am größten.

Auf der dritten Stufe wird der Baum weiter segmentiert - um die Baumstruktur übersichtlicher zu halten, wurden die Knoten 5 und 6, die keine weiteren Segmente bei den Älteren generieren, gekürzt:

ABBILDUNG 93 Dreistufige Alterssegmentierung mit CART bei den jüngeren Befragten (bis 57 Jahre, N = 1413)



Auch hier ergibt sich das gleiche Bild: die noch nicht volljährigen Befragten (Knoten 7) weisen mit über 90 % den höchsten Nutzeranteil auf, gefolgt von den 18 - 22jährigen (80 %, Knoten 8). Das Segment der 23- bis 45jährigen unterscheidet sich mit 66 % Nutzung nicht wesentlich von der Gruppe der über 45jährigen (59 %).

2.2 Berufsgruppen

Die Variable der Berufsgruppen hat viele Ausprägungen und ist sehr inhomogen hinsichtlich der Größe der einzelnen Segmente. Da dies bei multivariaten Analysen bekanntermaßen zu Problemen führt,

wäre eine erste Überlegung, die Gruppen weiter zusammenzufassen. Vor allem die kleineren Gruppen der Meister, Landwirte und Fischer und der Ladenbesitzer und (selbständigen) Handwerker, die teilweise nur rund 20 Befragte aufweisen. Das könnte auf zwei Wegen geschehen: entweder aufgrund theoretischer Überlegungen oder empirischer mit nachträglicher theoretischer Erklärung. Diese Überlegungen sind bei Entscheidungsbäumen überflüssig, da (vor allem bei Binäralgorithmen) ähnliche Gruppen zusammengefaßt werden. Ein Merkmal, das kein anderes Verfahren in dieser Form aufweist.

Nachfolgende Kreuztabelle gibt einen ersten Überblick über die Berufsstruktur im Datensatz:

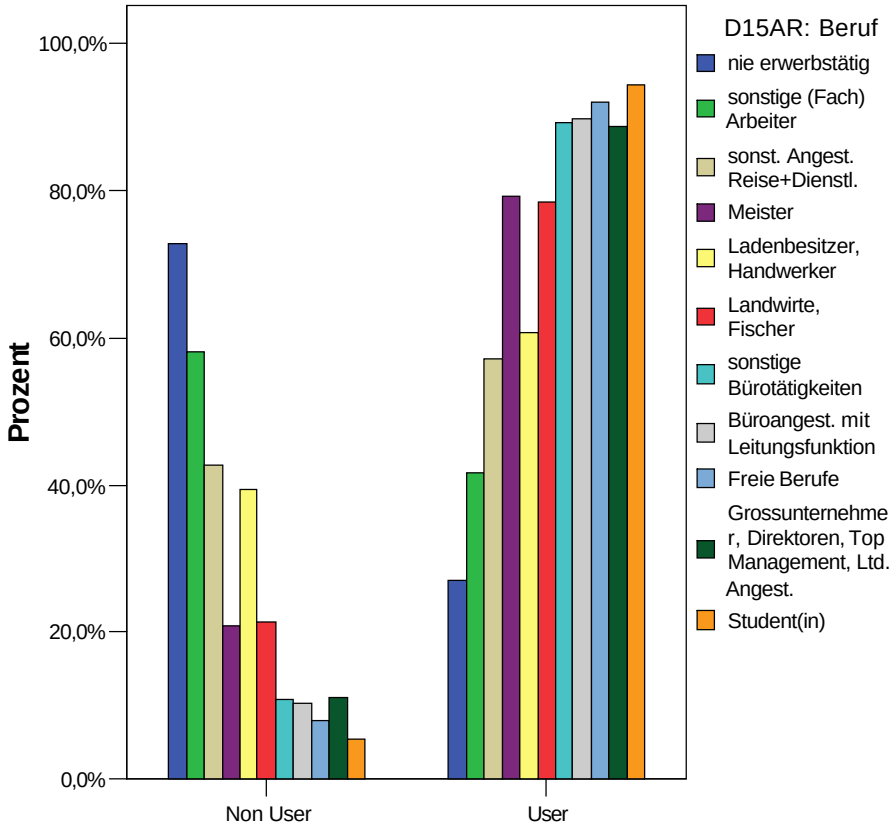
ABBILDUNG 94 Berufsgruppen nach PC-Nutzung (N = 1413)

Kreuztabelle

			Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)		
			Non User	User	Gesamt
D15AR: Berufliche Stellung	nie erwerbstätig	Anzahl	86	32	118
		% von D15AR: Berufliche Stellung	72,9%	27,1%	100,0%
	sonstige (Fach)Arbeiter	Anzahl	259	186	445
		% von D15AR: Berufliche Stellung	58,2%	41,8%	100,0%
	sonst. Angest. Reise+Dienstl.	Anzahl	86	115	201
		% von D15AR: Berufliche Stellung	42,8%	57,2%	100,0%
	Meister	Anzahl	5	19	24
		% von D15AR: Berufliche Stellung	20,8%	79,2%	100,0%
	Ladenbesitzer, Handwerker	Anzahl	11	17	28
		% von D15AR: Berufliche Stellung	39,3%	60,7%	100,0%
	Landwirte, Fischer	Anzahl	3	11	14
		% von D15AR: Berufliche Stellung	21,4%	78,6%	100,0%
	sonstige Bürotätigkeiten	Anzahl	18	149	167
		% von D15AR: Berufliche Stellung	10,8%	89,2%	100,0%
	Büroangest. mit Leitungsfunktion	Anzahl	19	168	187
		% von D15AR: Berufliche Stellung	10,2%	89,8%	100,0%
	Freie Berufe	Anzahl	3	35	38
		% von D15AR: Berufliche Stellung	7,9%	92,1%	100,0%
	Grossunternehmer, Direktoren, Top Management, Ltd. Angest.	Anzahl	7	56	63
		% von D15AR: Berufliche Stellung	11,1%	88,9%	100,0%
	Student(in)	Anzahl	7	121	128
		% von D15AR: Berufliche Stellung	5,5%	94,5%	100,0%
Gesamt		Anzahl	504	909	1413
		% von D15AR: Berufliche Stellung	35,7%	64,3%	100,0%

Deutlicher wird dies an der grafischen Darstellung:

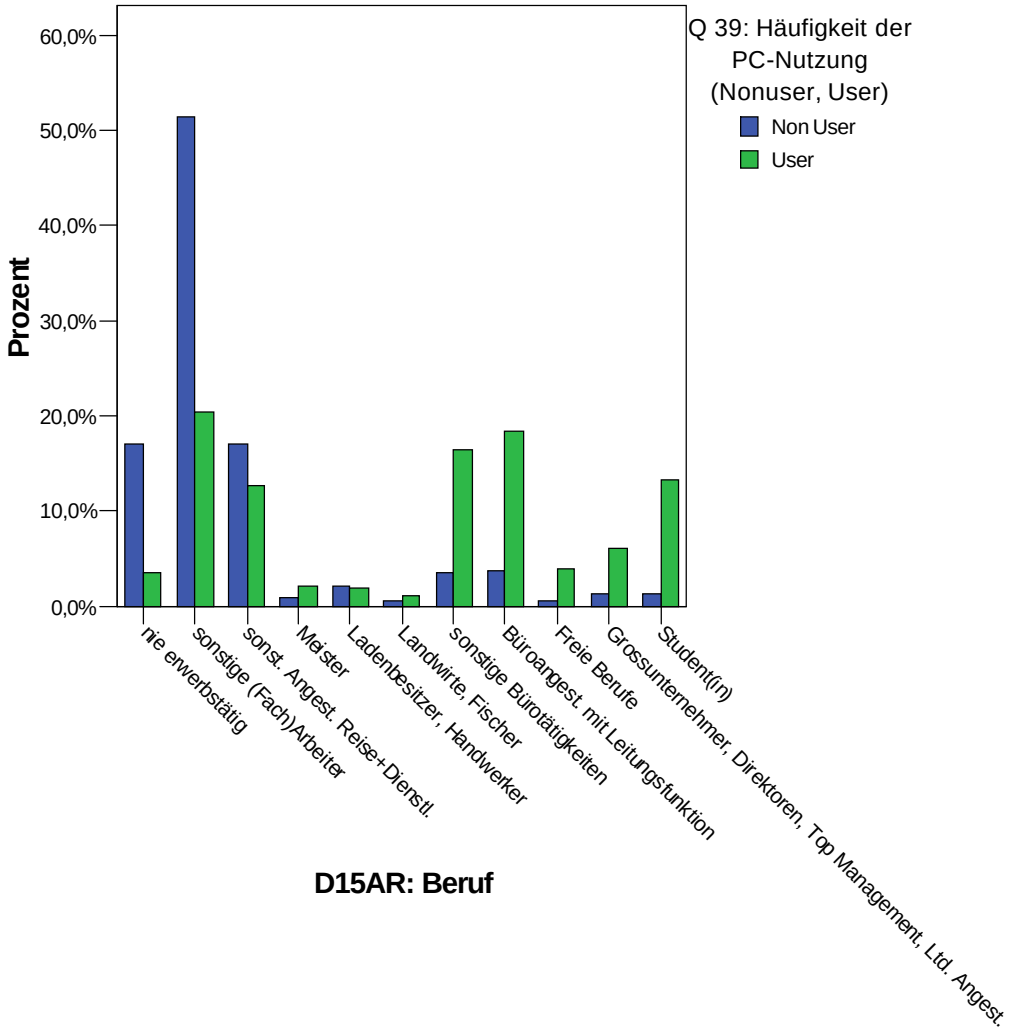
ABBILDUNG 95 Grafische Darstellung: Häufigkeit der PC-Nutzung nach Beruf (N = 1413, in %, Kategorie: PC-Nutzer



Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)

Bei vielen Kategorien bietet sich eine andere Darstellung an - der direkte Kategorienvergleich, wo auf den ersten Blick deutlich wird, welche Kategorien die höheren (Nicht-)Nutzeranteile aufweisen:

ABBILDUNG 96 Grafische Darstellung: Häufigkeit der PC-Nutzung nach Beruf (N = 1413, in %, Kategorie: Beruf)



Theoretisch denkbar wäre, Büroberufe von anderen Berufen zu trennen, die Berufe somit zu dichotomisieren. Hintergrund ist der verbreitete Einsatz von PCs in Büros. Andererseits weisen neben den Büroberufen noch andere Berufsgruppen (Landwirte, Fischer, Meister) eine recht hohe PC-Nutzung auf. Aufgrund der Kreuztabelle würde man die „nie Erwerbstätigen“ und die Arbeitergruppe, die den höchsten Anteil an Nichtnutzern aufweisen, zusammenfassen (bis etwa 40 %).

Die Ladenbesitzer, Handwerker und Reise- und Dienstleistungsberufe mit einem Anteil von rund 60 % Nutzern könnten ebenfalls zusammengefaßt werden. Büro- und Freie Berufe, Studierende, Landwirte, Fischer und Meister weisen eine sehr hohe Nutzerquote auf.

Vergleichend werden wieder die Algorithmen EXHAUSTIVE CHAID, QUEST und CART eingesetzt. Zusammengefaßt wurden nur die sehr homogenen Büroberufe.

Von den 1413 Befragten sind rund 1/3 Nichtnutzer, zwei Drittel Nutzer (in der Gesamtstichprobe lag das Verhältnis der Nutzer zu den Nichtnutzern bei 984 : 2038).

2.2.1 Berufssegmentierung mit EXHAUSTIVE CHAID

(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)

Die EXHAUSTIVE-CHAID-Lösung der Variable ist nicht sehr glücklich: es entsteht nur eine Ebene mit sieben Segmenten aus 11 Variablenausprägungen, die aufgrund der Größe der einzelnen Gruppen nicht mehr weiter aufgeteilt werden kann. Aufgrund der unübersichtlichen Baumstruktur wird das Ergebnis in einer Tabelle zusammengefaßt:

TABELLE 19 Einstufige Berufssegmentierung mit EXHAUSTIVE CHAID bei den jüngeren Befragten (bis 57 Jahre, N = 1413)

Berufssegment	PC-Useranteil in % / N
Meister, Landwirte, Fischer	76.85, N = 36
Freie Berufe	92.11, N = 38
nie erwerbstätig	27.12 (N = 118)
Ladenbesitzer, Handwerker	60.71 (N = 28)
Studierende	94.53 (N = 128)
Büroangestellte mit Leitungsfunktion	89.54, N = 187
(sonstige Fach)Arbeiter	41.80 (N = 445)
sonstige Bürotätigkeiten	89.22, N = 167
Grossunternehmer, Direktoren, Top Management	88.89 N = 63
sonstige Angestellte (Reise- und Dienstleistung)	57.21 (N = 201)

Die Lösung ist sehr ineffizient und könnte eine Schwäche des CHAID-Algorithmus offenbaren: die Schwierigkeit, Variablen mit vielen Kategorien effizient zusammenzufassen. So ist auf den ersten Blick ersichtlich, dass die sonstigen Bürotätigkeiten und die Gruppe der Großunternehmer, Direktoren, Top Management prozentual sehr nahe beieinanderliegen (89.22 vs. 88.89 % PC-Nutzung). Daneben wird deutlich, dass sich Gruppen von rund 40 Personen schwerlich weiter sinnvoll segmentieren lassen. Andererseits wird auch herausgearbeitet, wie rigoros die Gruppen bei einer Binärbetrachtung zusammengefaßt werden (bei CART zum Beispiel sonstige (Fach)Arbeiter, Reise- und Dienstleistungsberufe, nie erwerbstätig) - die PC-Useranteile liegen zwischen 27.12 % (nie erwerbstätig) über 41.8 % (Arbeiter) bis 57.21 % (sonstige Reise und Dienstleistungsberufe).

2.2.2 Berufssegmentierung mit QUEST

QUEST	• vieldiskutierter, relativ neuer Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus (unabhängige Variablen) geeignet • nur für nominale Zielvariablen geeignet, die auch dichotom sein kann • gebräuchliche statistische Kennzahlen (Chi-Quadrat, F-Test) • erlaubt Ersatzprädiktoren, Pruning und a prioris

Der QUEST-Algorithmus trennt die nie Erwerbstätigen und Arbeiter (Knoten 1) und stellt sie dem Rest der Stichprobe (Knoten 2) gegenüber[85]:

ABBILDUNG 97 QUEST: Berufssegmentierung bei den jüngeren Befragten (höhere Nutzeranteile, bis 57 Jahre, N = 1413)

Knoten 1

Kategorie	%	n
Non User	61,28	345
User	38,72	218
Gesamt	(39,84)	563

Knoten 2

Kategorie	%	n
Non User	18,71	159
User	81,29	691
Gesamt	(60,16)	850

Deutlich wird, dass die Nutzung in anderen Berufssegmenten (Knoten 2) mehr als doppelt so hoch ist wie im Knoten 1. Während Knoten 1 auf der nächsten Stufe folgerichtig in die nie Erwerbstätigen und Arbeitergruppen (Knoten 3 und 4, Anteile: 27 : 42 % Nutzer) segmentiert

85. Aufgrund der recht unglücklichen Darstellung durch das lange Label wird der Baum selbst auf einer DinA4-Seite nicht mehr lesbar. Deshalb wurde auf die Einzelknoten zurückgegriffen.

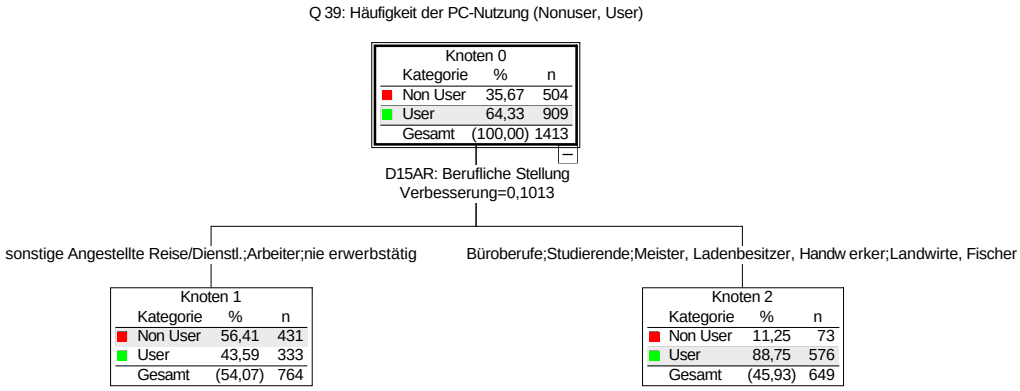
wird, werden die Reise- und Dienstleistungsberufe aus Knoten 2 herausgelöst. Es ergibt sich Knoten 3 mit einem Nutzeranteil von 57 %, die restlichen Gruppen (Büroberufe, Landwirte, Fischer, Ladenbesitzer, Meister, Studierende) weisen 88 % Nutzer auf.

QUEST erkennt in diesem Fall also keinen Unterschied zwischen den Büroberufen und Studierenden mit höchsten Nutzeranteilen einerseits und den Berufen, die als „alter Mittelstand“ bezeichnet wurden.

2.2.3 Berufssegmentierung mit CART

CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und a prioris

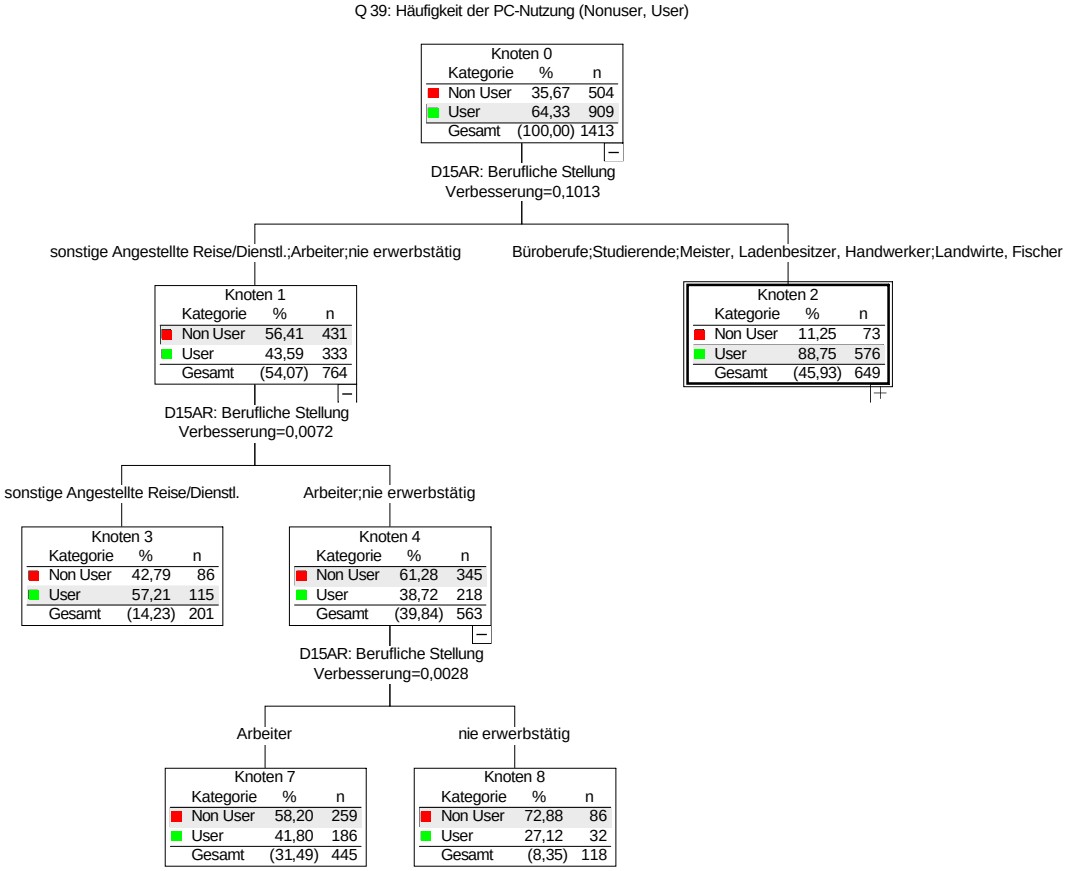
ABBILDUNG 98 Einstufige Berufssegmentierung mit CART bei den jüngeren Befragten (bis 57 Jahre, N = 1413)



Während Knoten 1 mit den Berufsgruppen 'nie erwerbstätig', '(sonstige Fach)Arbeiter' und 'sonst. Angest. Reise+Dienstl.' recht inhomogen ist (Non-User : User 56 : 44 %) fällt Knoten 2 (Büroberufe, Studierende, Meister, Landwirte, Fischer, Handwerker und Ladenbesitzer) mit dem Verhältnis 11 : 89 recht homogen aus. CART kommt somit statistisch zu

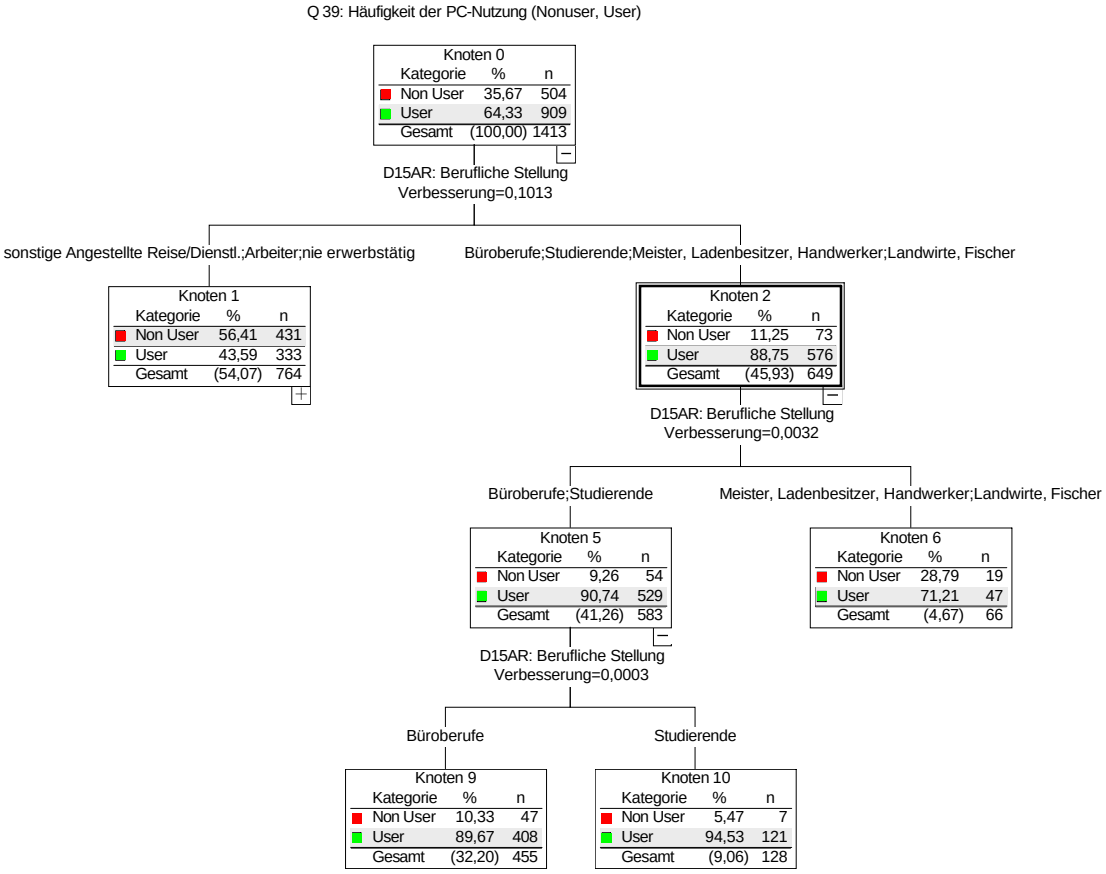
einem anderen Ergebnis als durch die oben aufgestellte Hypothese. Auf den weiteren Baumstufen werden die Gruppen differenzierter.

ABBILDUNG 99 Berufssegmentierung mit CART bei den jüngeren Befragten (geringere Nutzeranteile, bis 57 Jahre, N = 1413)



Die geringsten Nutzeranteile weisen die nie Erwerbstätigen und die Arbeitergruppen auf, wobei die zweite Gruppe den PC deutlich häufiger nutzt (42 %) als die erste (27 %). Bei den sonstigen Angestellten der Reise- und Dienstleistungsberufe nutzt gut jeder zweite (57 %) den Rechner.

ABBILDUNG 100 CART: Berufssegmentierung bei den jüngeren Befragten (höhere Nutzeranteile, bis 57 Jahre, N = 1413)



Neun von zehn Befragten, die studieren oder einen Büroberuf ausüben, nutzen den PC, wobei der Studierendenanteil mit knapp 95 % am höchsten ist (Büroberufe: 90 %). Bei den Meistern, Ladenbesitzern, Handwerkern, Landwirten und Fischern sind das gut drei von vier Befragten. Es darf deshalb nie vergessen werden, dass die Useranteile „Durchschnittswerte“ darstellen. Nicht alle Berufsgruppen nutzen zu 90 % den Computer, sondern die Anteile liegen (hier) zwischen 85 % und 95 %, was einen kleinen Unterschied ausmacht.

Es ergibt sich also ein deutliches „Gefälle“ zwischen Studierenden/ Büroberufen und den Reise- und Dienstleistungsberufen, den Arbeitern und nie Erwerbstätigen. Ein Grund für diese Diskrepanz ist die

deutlich höhere berufliche Nutzung in Büro und Studium. Aber auch die Gruppen des „alten Mittelstandes“ kommen nicht um den Einsatz des PCs herum - dann schon eher die Arbeiter und nie Erwerbstätigen.

CART liefert in diesem Beispiel das überzeugendste Segmentierungsergebnis.

Isolierte Betrachtungen von einer unabhängigen Variablen (z. B. berufliche Stellung) machen multivariat wenig Sinn, da weitere unabhängige Variablen diese beeinflussen. Als grafisches Merkmal sind sie aber ideal: zwar lassen sich die Informationen im wesentlichen auch aus einer Kreuztabelle ablesen, die multivariate und grafische Darstellung liefert allerdings dieses Verfahren nicht. Aussagen wie: „Nie Erwerbstätige und Arbeiter sind sich in der PC-Nutzung ähnlich“ sind so nicht direkt ersichtlich. Deshalb können Entscheidungsbaumverfahren auch deskriptiv eingesetzt werden, um sich über die Struktur (d. h. die Variablenausprägungen) einer Variablen Klarheit zu verschaffen.

Somit lassen sich also drei Berufssegmente identifizieren: die Gruppe mit den höchsten Useranteilen (um 90 %) besteht aus Büroberufen und Studierenden. Landwirte, Fischer, Meister, Ladenbesitzer und Handwerker nutzen zu rund 70 % den PC. Die niemals Erwerbstätigen Arbeiter und Reise- und Dienstleistungsangestellten bilden die Gruppe mit den geringsten Nutzeranteilen, wobei es hier nochmals von Segment zu Segment größere Schwankungen gibt.

2.3 Bildung

Die Erwartung, dass die Nutzergruppen eher bei den höheren Schulabschlüssen zu finden sind, wurde bereits im letzten Kapitel ausführlich dargestellt. Der Kennwert von Cramers v war ebenfalls deutlich

hinsichtlich der PC-Nutzung. Allerdings zeigt sich hier kein „monotoner" Verlauf von der Hauptschule bis hin zu Hochschulabsolventen: die prozentualen Anteile der Studierenden liegen höher als die der Hochschulabsolventen. Ansonsten steigt der Anteil der PC-Nutzer parallel mit höheren Bildungszertifikaten an.

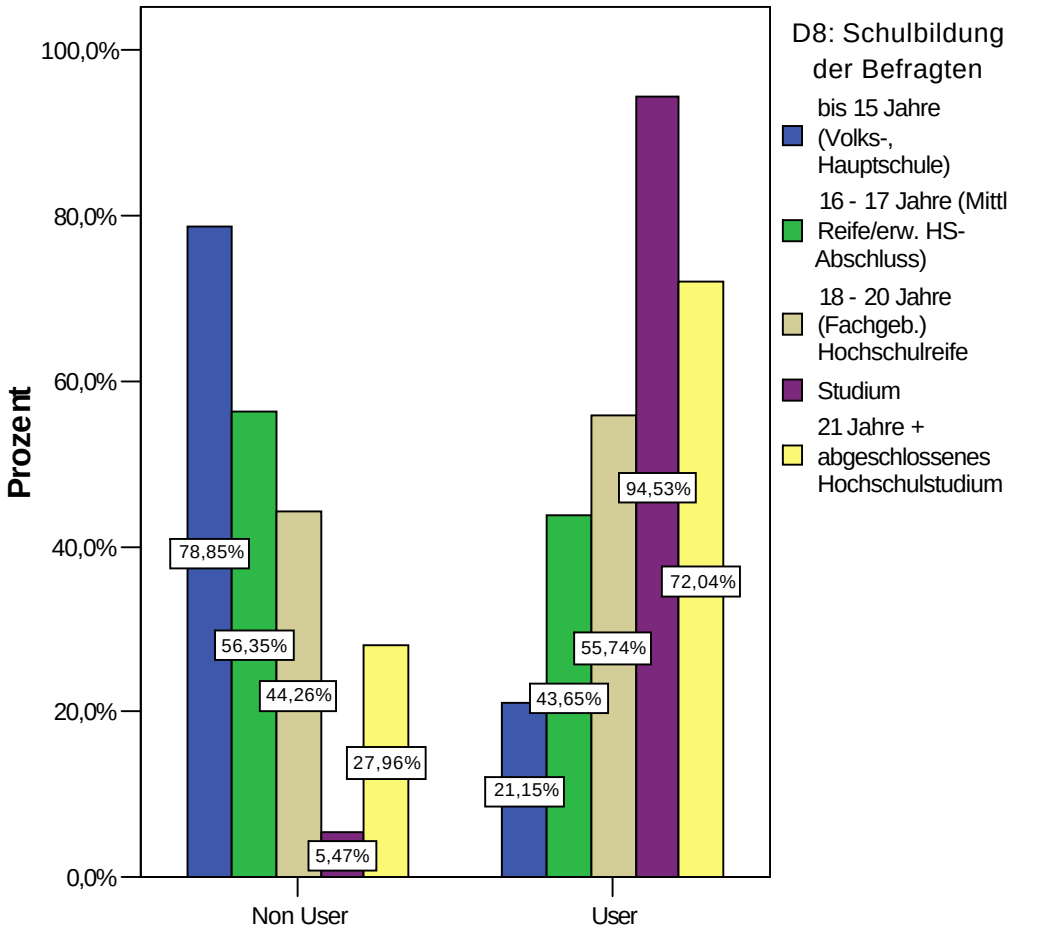
ABBILDUNG 101 PC-Nutzung nach Schulbildung (N, Zeilen-%, N = 1413

D8: Schulbildung der Befragten * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

			Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)		
			Non User	User	Gesamt
D8: Schulbildung der Befragten	bis 15 Jahre (Volks-, Hauptschule)	Anzahl	120	95	215
		% von D8: Schulbildung der Befragten	55,8%	44,2%	100,0%
	16 - 17 Jahre (Mittl Reife/erw. HS-Abschluss)	Anzahl	217	258	475
		% von D8: Schulbildung der Befragten	45,7%	54,3%	100,0%
	18 - 20 Jahre (Fachgeb.) Hochschulreife	Anzahl	131	252	383
		% von D8: Schulbildung der Befragten	34,2%	65,8%	100,0%
	Studium	Anzahl	7	121	128
		% von D8: Schulbildung der Befragten	5,5%	94,5%	100,0%
	21 Jahre + abgeschlossenes Hochschulstudium	Anzahl	29	183	212
		% von D8: Schulbildung der Befragten	13,7%	86,3%	100,0%
Gesamt		Anzahl	504	909	1413
		% von D8: Schulbildung der Befragten	35,7%	64,3%	100,0%

Diese Kreuztabelle läßt sich übersichtlicher als Grafik darstellen:

ABBILDUNG 102 Grafische Darstellung: Häufigkeit der PC-Nutzung nach Schulbildung (in %, N = 1413)



Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)

Im Gegensatz zur Kreuztabelle sieht man auf einen Blick, dass sich die Prozentanteile der jeweiligen Kategorien hinsichtlich des Schulabschlusses relativ monoton entwickeln: je höher der Bildungsabschluss, desto höher die PC-Nutzeranteile und umgekehrt. Der Anteil der Studierenden bei den Nichtnutzern ist deshalb mit rund 5.5 % niedriger als bei den Akademikern, weil hier das (junge) Alter eine Wirkung zeigt: es gibt heute kaum Studierende ohne PC. Hier sind wahrscheinlich keine großen Überraschungen bei den Algorithmen zu erwarten.

(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)

Der EXHAUSTIVE-CHAID-Algorithmus erbringt kein befriedigendes Ergebnis: durch die Prozentsatzdifferenzen von mindestens 10 % werden hier so gut wie keine Trennungen gefunden. Es werden fünf Knoten segmentiert, was keine sinnvolle Trennung darstellt.

TABELLE 20 Einstufige Bildungssegmentierung mit EXHAUSTIVE CHAID bei den jüngeren Befragten (bis 57 Jahre, N = 1413)

Bildungskategorien	gefundene Segmente
Volks-, Hauptschule	Knoten 1 Kategorie % n Non User 55,81 120 User 44,19 95 Gesamt (15,22) 215
Mittlere Reife, erweiterter Hauptschulabschluss	Knoten 2 Kategorie % n Non User 45,68 217 User 54,32 258 Gesamt (33,62) 475
(fachgebundene) Hochschulreife	Knoten 3 Kategorie % n Non User 34,20 131 User 65,80 252 Gesamt (27,11) 383
Studierende	Knoten 4 Kategorie % n Non User 5,47 7 User 94,53 121 Gesamt (9,06) 128
Akademiker	Knoten 5 Kategorie % n Non User 13,68 29 User 86,32 183 Gesamt (15,00) 212

Der Vergleich mit der Kreuztabelle PC-Nutzung und Bildung zeigt keine Unterschiede - die Kreuztabelle ist mit den von EXHAUSTIVE CHAID

gefundenen Segmenten identisch und liefert keine zusätzlichen Informationen.[86]

2.3.2 Bildungssegmentierung mit QUEST

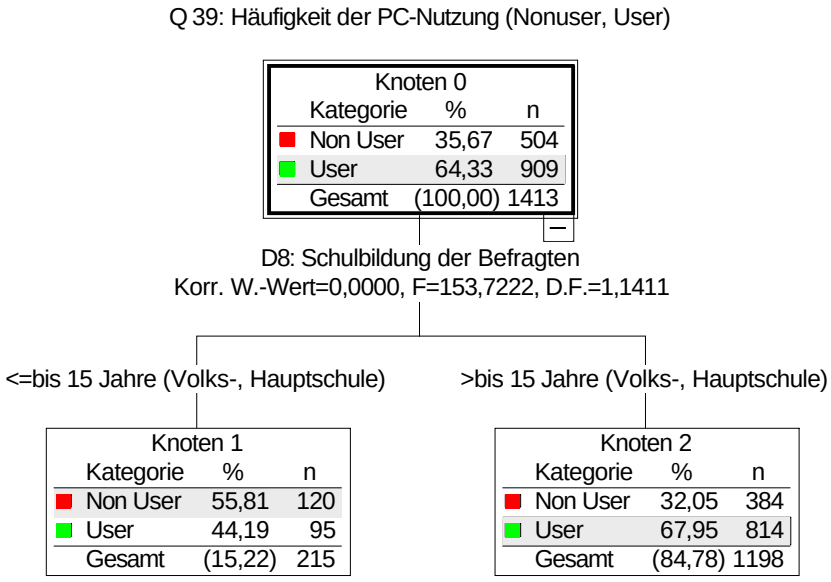
QUEST	• vieldiskutierter, relativ neuer Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus (unabhängige Variablen) geeignet • nur für nominale Zielvariablen geeignet, die auch dichotom sein kann • gebräuchliche statistische Kennzahlen (Chi-Quadrat, F-Test) • erlaubt Ersatzprädiktoren, Pruning und a prioris

An diesem Punkt, wo bei den auf CHAID beruhenden Algorithmen keine sinnvollen Segmente gefunden werden, bieten binäre Algorithmen eine Alternative: sie müssen die fünf Kategorien zu jeweils zwei Knoten pro Baumebene zusammenfassen.

QUEST segmentiert im ersten Schritt wird den höchsten Anteil der Nichtnutzer heraus (Knoten 1):

86. Allerdings sind multivariate Verfahren wie die der Entscheidungsbäume nicht primär für das Ziel entwickelt worden, eine Variable zu segmentieren (das Ergebnis liefert häufig auch eine Kreuztabelle), sondern ein Modell aus unterschiedlichsten unabhängigen Variablen zu untersuchen.

ABBILDUNG 103 QUEST. Einstufige Bildungsabschlussegmentierung bei den jüngeren Befragten (höhere Nutzeranteile, bis 57 Jahre, N = 1413)



Die Trennung erfolgt hier nicht bei höheren Schulabschlüssen, sondern diejenigen mit Hauptschulabschluss, die die geringste Nutzung aufweisen, werden aus der Stichprobe „herausgelöst". Der Nutzeranteil liegt bei 44 % (Knoten 1) und 68 % (Knoten 2).

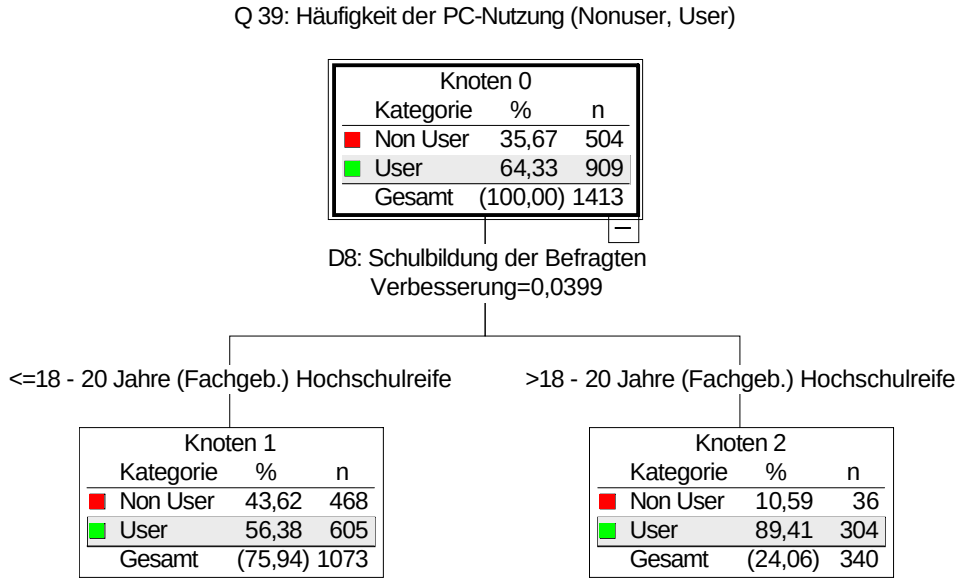
Der Nachteil bei diesem Vorgehen liegt darin, dass Knoten 1 nicht weiter aufgesplittet werden kann, da er nur eine Variablenausprägung (Volks-, Hauptschule) enthält. Dafür lassen sich die restlichen Kategorien in Knoten 2 differenzierter untersuchen.

2.3.3 Bildungssegmentierung mit CART

CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und a prioris

Im Gegensatz zu QUEST versucht CART eher, Trennungen zu finden, die sich auf beiden Seiten besser segmentieren lassen - mit dem Nachteil, dass dadurch das Ergebnis möglicherweise nicht so deutlich wird.

ABBILDUNG 104 CART: Einstufige Bildungsabschlussegmentierung bei den jüngeren Befragten (höhere Nutzeranteile, bis 57 Jahre, N = 1413)



Mit der Erzwingung von zwei Unterknoten weist Knoten 1 alle Abschlüsse bis einschließlich der Hochschulreife (Hauptschule, Mittlere Reife, (fachgebundene) Hochschulreife) auf. Knoten 2 enthält die Gruppe der Studierenden und Hochschulabsolventen. Die weitere Segmentierung ist klar: aus Knoten 2 resultieren die Studierenden und Hochschulabsolventen, aus Knoten 1 die geringeren Bildungszertifikate.

Die beiden Strategien von CART und QUEST zeigen die unterschiedlichen Möglichkeiten auf, an ein Klassifizierungsproblem heranzugehen: entweder wird (unter Informationsverlust) ein Grossteil der einen Gruppe heraussegmentiert (QUEST) - mit dem Nachteil, eine der Gruppen evtl. nicht mehr weiter segmentieren zu können. CART geht

hier behutsamer vor, hat aber auf der ersten Ebene größere „Mischgruppen“, was dazu führen kann, nicht mehr so effiziente Strukturen zu finden - dafür werden aber alle Gruppen besser charakterisiert.

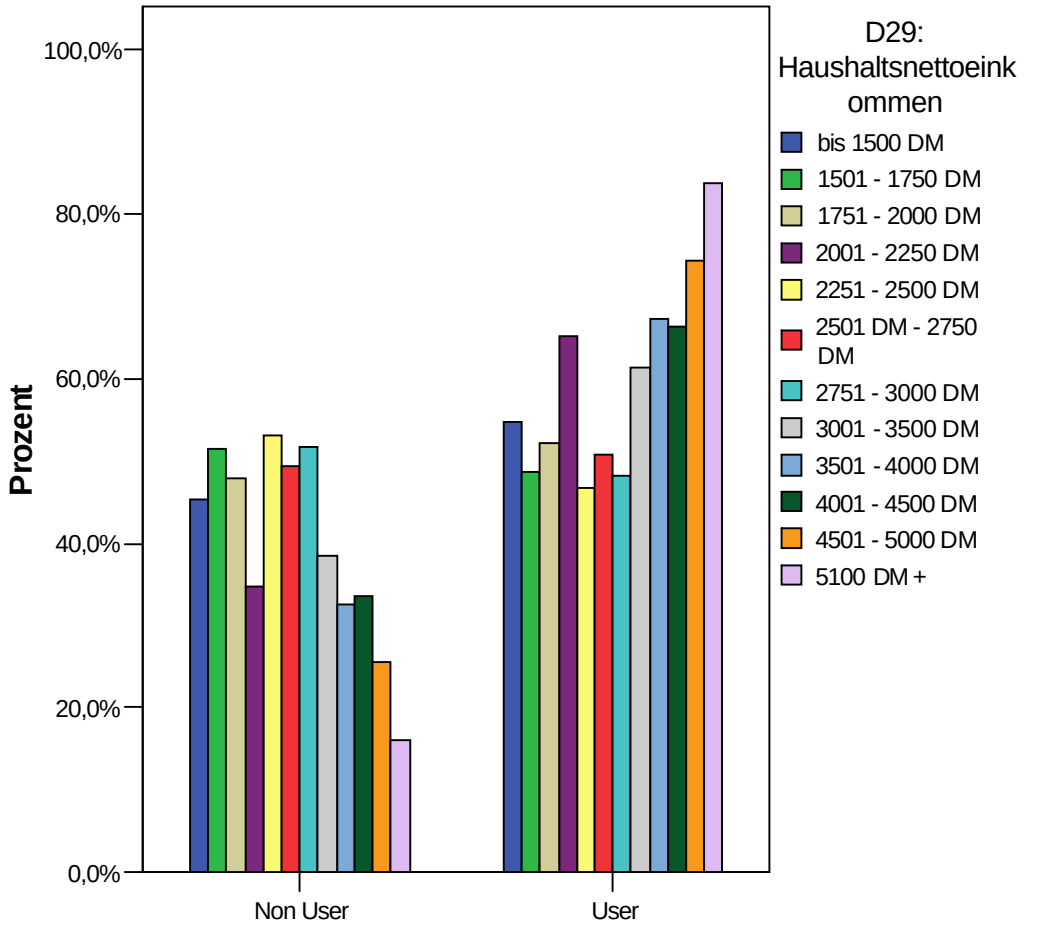
2.4 Haushaltsnettoeinkommen

These für diese Variable, bezogen auf PC-Nutzung wäre, dass mit zunehmendem Einkommen der Anteil der PC-Nutzung steigt. Cramers v liegt mit 0.259 auf Rang 4 der untersuchten Zusammenhänge.

Nun könnte vermutet werden, dass der mit Rang 4 nicht so hohe Zusammenhangswert dadurch zustandekommt, dass es Nutzergruppen mit geringem Einkommen gibt, die den PC nutzen (z. B. Studierende), Nutzergruppen mit etwas höherem Einkommen, jedoch weniger und gutverdienende Gruppen, die eine hohe Nutzung aufweisen, wieder mehr.

Die Kreuztabelle ist durch die zwölf Kategorien sehr umfangreich und nicht sehr anschaulich. Das Ergebnis wird deshalb als Balkendiagramm wiedergegeben:

ABBILDUNG 105 Haushaltsnettoeinkommen nach PC-Nutzung (in %, N = 1413)



Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)

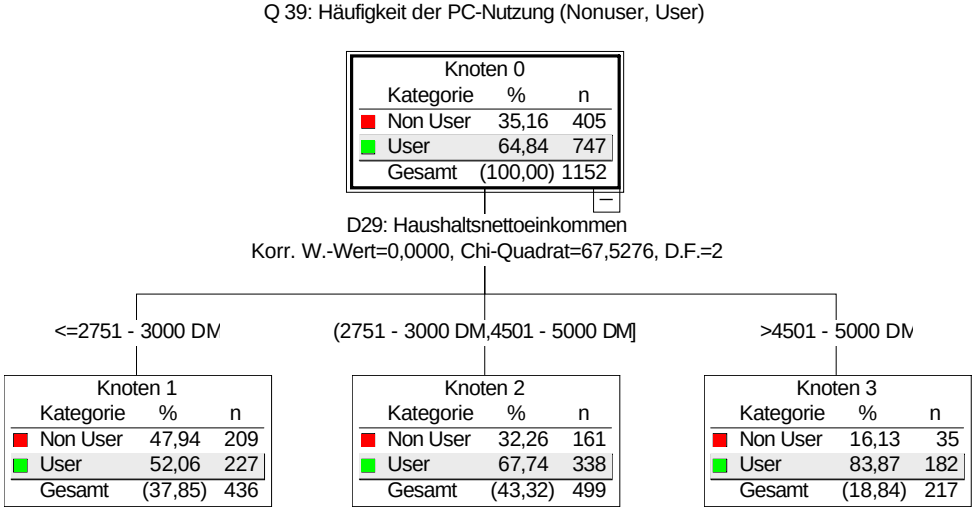
Deutlich wird, dass die User- bzw. Non-User-Anteile in den unteren Einkommensgruppen ähnlich hoch sind (Ausnahme: 2001 bis 2250 DM). Erst in den Gruppen ab 3000 DM nimmt die PC-Nutzung deutlich zu.

Das Ergebnis läßt sich auch durch Entscheidungsbäume validieren. Hier könnten QUEST und C&RT durch ihre Dichotomisierung zu einer Informationsreduktion beitragen (Gruppe der Einkommensbezieher von 2000 bis 2250 DM).

2.4.1 Haushaltsnetto-Einkommenssegmentierung mit EXHAUSTIVE CHAID

(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)

ABBILDUNG 106 Einstufige Haushaltsnettosegmentierung mit EXHAUSTIVE CHAID bei den jüngeren Befragten (bis 57 Jahre, N = 1152)



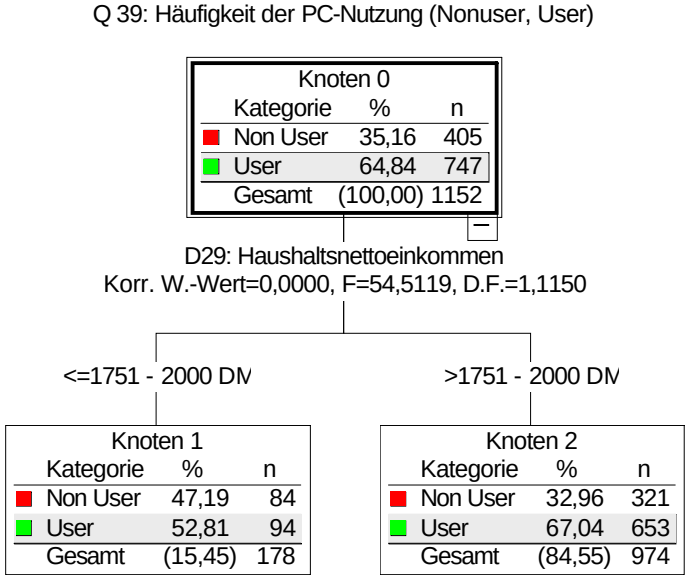
Die Nutzeranteile steigen kontinuierlich von Knoten 1 (52 %) über Knoten 2 (rund 2/3) auf knapp 84 % in Knoten 3 an. Die Überschneidung der Einkommenskategorien in den beiden linken Ästen (2751 bis 3000 DM) ist darauf zurückzuführen, dass es Nutzer gibt, die statistisch hinsichtlich des Chi-Quadrat-Werts besser in das mittlere Segment passen. Es wird aber deutlich, dass sich der rechte Knoten 3 mit 84 % Anteil deutlicher hinsichtlich der Prozentsatzdifferenz unterscheidet als die beiden anderen Segmente. Die im Balkendiagramm sichtbaren Unterschiede in der Gruppe 2000 bis 2250 DM gehen hier unter - eine Gefahr, die leider allen multivariaten Verfahren zueigen ist und

verdeutlicht, wie wichtig es ist, der multivariaten eine bivariate Analyse vor- und auch nachzuschalten.

2.4.2 Haushaltsnetto-Einkommenssegmentierung mit QUEST

QUEST	• vieldiskutierter, relativ neuer Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus (unabhängige Variablen) geeignet • nur für nominale Zielvariablen geeignet, die auch dichotom sein kann • gebräuchliche statistische Kennzahlen (Chi-Quadrat, F-Test) • erlaubt Ersatzprädiktoren, Pruning und a prioris

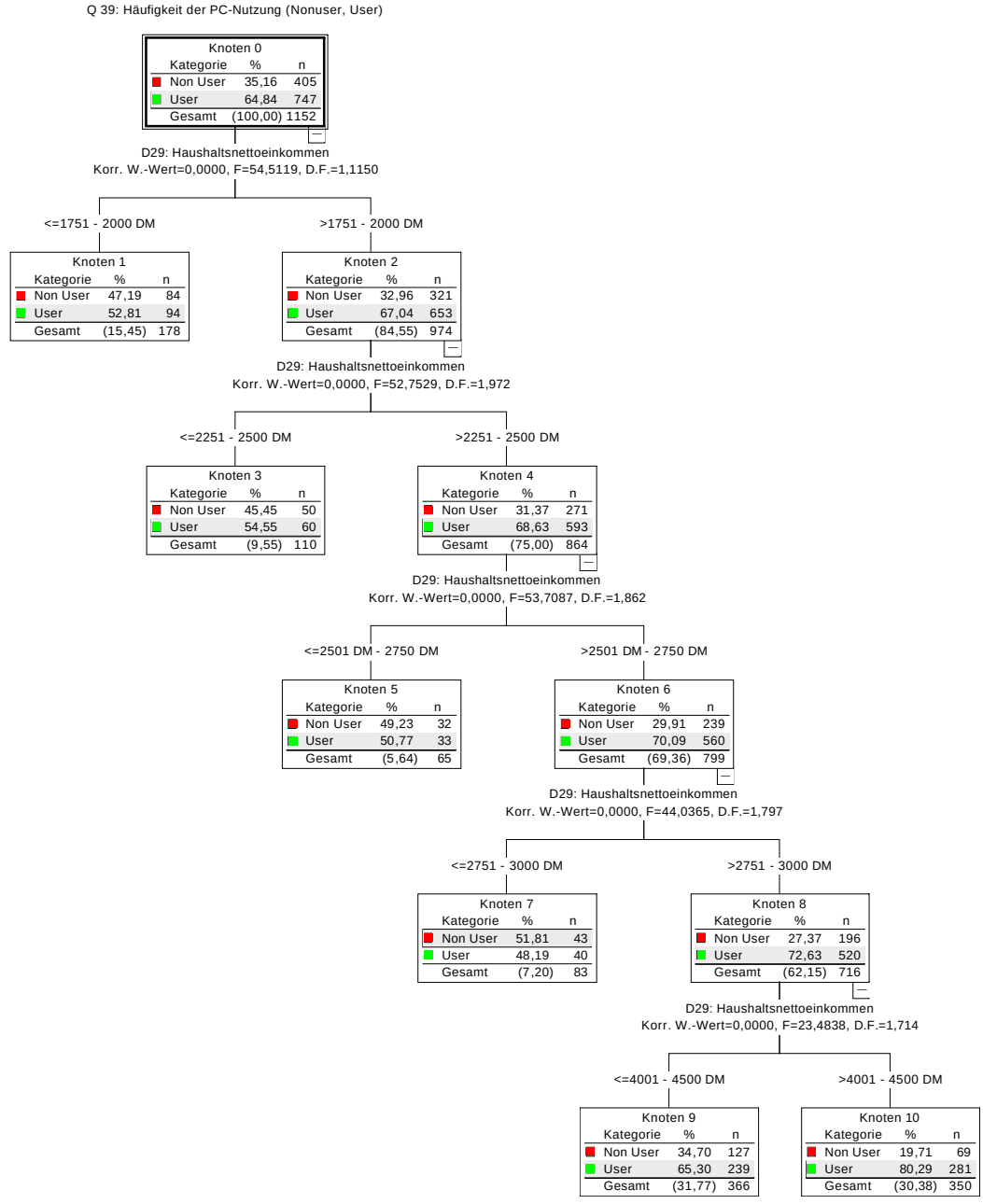
ABBILDUNG 107 QUEST: Einstufige Haushaltsnettosegmentierung bei den jüngeren Befragten (bis 57 Jahre, N = 1152



Auch in diesem Beispiel wird deutlich, wie sich die Ergebnisse von QUEST und CART unterscheiden. Die Segmentierungen sehen auf der ersten Stufe zwar ähnlich aus, allerdings kann Knoten 1 aufgrund der geringen Fallzahlen nicht weiter segmentiert werden. Es zeigt sich auch hier die Neigung von QUEST, diejenigen „auszusortieren“, die

ein weiteres Ergebnis stören. Dafür wird der rechte Teil des Astes (Knoten 2) sehr differenziert aufgesplittet. Dies ist jedoch nicht immer von Vorteil, da beim Herauslösen von nur einer Kategorie die restlichen Gruppen inhomogen bleiben:

ABBILDUNG 108 Haushaltsnettosegmentierung mit QUEST bei den jüngeren Befragten (bis 57 Jahre, N = 1152



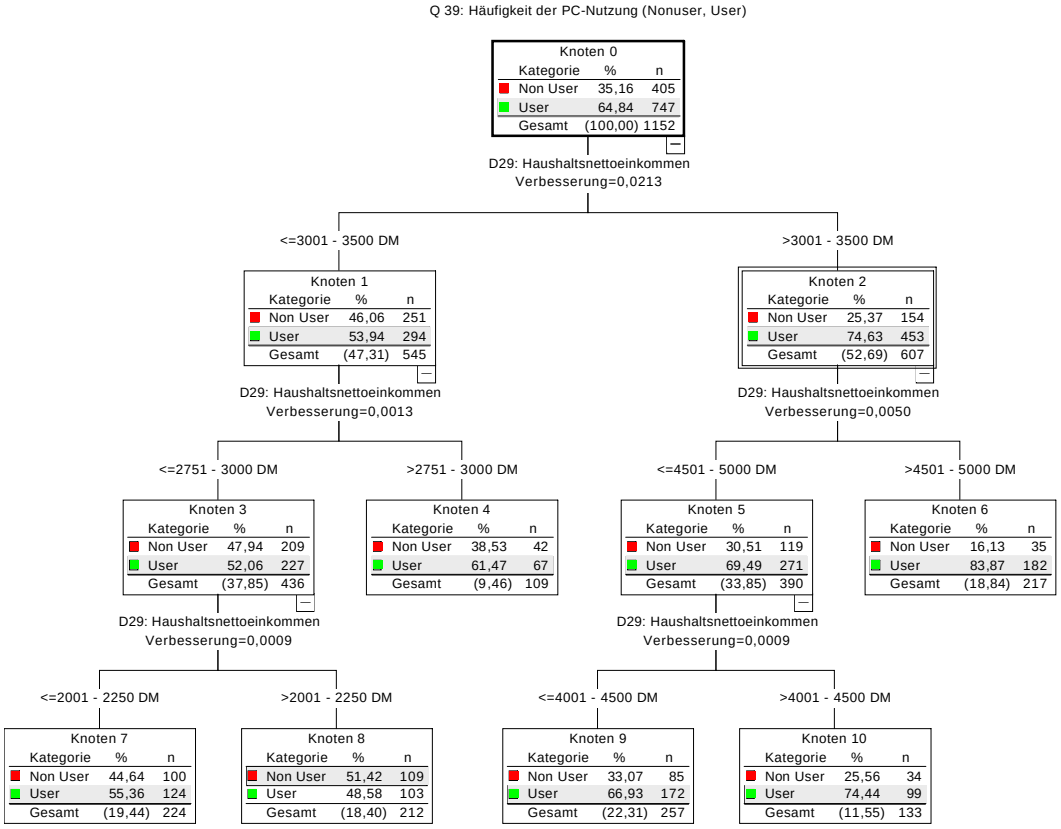
Der erhöhte Anteil der Nutzer in der Kategorie 2000 - 2250 DM wird nicht direkt erkannt - er ergibt sich nur im Vergleich der Knoten 3 und 5 - und wenn man weiß, dass es diesen „Ausreißer“ gibt, denn eigentlich müßte Knoten 3 (mit geringerem Haushaltsnettoeinkommen) auch einen kleineren Anteil an PC-Nutzern aufweisen - was im Vergleich zu Knoten 5 nicht der Fall ist.

2.4.3 Haushaltsnetto-Einkommenssegmentierung mit CART

CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und a prioris

Zum Vergleich der CART-Baum:

ABBILDUNG 109 CART: Haushaltsnettosegmentierung bei den jüngeren Befragten (bis 57 Jahre, N = 1152)



Die Lösung ist hier - bedingt durch die etwa gleich hoch besetzten Ausprägungen - nicht sehr effizient. Hier kommt der EXHAUSTIVE-CHAID-Baum zu einem etwas erhellenderem Ergebnis - allerdings können keine weiteren Stufen mehr extrahiert werden.

In den Knoten 1 (54 % User) bzw. 2 (75 % User) werden die Einkommenskategorien bei 3000 DM getrennt. Die Kategorie 3000 bis 3500 DM überschneidet sich in beiden Segmenten. Während die Rechnernutzung aus dem Knoten 2 resultierenden Einkommensgruppen mit zunehmenden Einkommen ansteigt, gilt dies nicht durchgängig für die aus Knoten 1 resultierenden Befragten, die ein geringeres Haushaltsnettoeinkommen beziehen: das Segment der 2000 bis 2250 DM verdienenden Personen weist mit 55 : 49 % einen höheren Useranteil auf als jene Gruppe, die zwischen 2251 und 3000 DM verdienen.

Festzuhalten bleibt, dass das Haushaltsnettoeinkommen erst in den Gehaltsgruppen ab 3000 DM deutlich über 50 % liegt. In den unteren Gehaltsgruppen sieht es so aus, als würde der „Zufall" über die Nutzung entscheiden (ca. 50 % Nutzung - Nichtnutzung).

2.5 Zusammenfassung

Beruf ist - im deskriptiven Fall - die wichtigste Variable zur Segmentierung von PC-Nutzern, gefolgt von Bildung und Alter. Sie gehören damit eindeutig zu den dominanten Schichtungen - ebenso wie das Haushaltsnettoeinkommen. Familienstand und Geschlecht weisen zu geringe Zusammenhangswerte auf. Sie können nur als untergeordnet angesehen werden.

Betrachtet man die Struktur der Variablen, ergibt sich eigentlich ein umgekehrter Effekt: die Variable Alter ist unabhängig von Bildung, Haushaltsnettoeinkommen und Beruf, der Bildungsgrad unabhängig von Beruf, aber abhängig vom Alter. Der Beruf ist abhängig vom Bildungsgrad, evtl. vom Alter - und das Haushaltsnettoeinkommen vom Beruf, wahrscheinlich auch vom Bildungsgrad. Die deskriptiven Zusammenhänge sind in nachfolgender Tabelle zusammengefaßt:

TABELLE 21 BIVARIATE ZUSAMMENHÄNGE (E = ETA, V = CRAMERS V, U = UNSICHERHEITSKOEFFIZIENT, R_S = SPEARMANS RANGKORRELATIONSKOEFFIZIENT) DER DOMINANTEN SCHICHTUNGSVARIABLEN

Variable	Alter	Berufs-gruppe	Bildungsgrad	HH-Nettoeink.
Alter (eta^2)	1	$e^2 = 0.145$	$e^2 = 0.057$	$e^2 = 0.131$
Berufsgruppe (v, u)		1	v = 0.557 u = 0.201	v = 0.182 u = 0.066
Bildungsgrad (v, u, tau b, tau c, s)			1	$r_s = 0.131$ tau b = 0.101 tau c = 0.104

Interessanterweise sind die Zusammenhänge nicht so hoch wie erwartet - ausser beim Bildungsabschluss und den Berufsgruppen. Trotzdem sind die gegenseitigen Beeinflussungen der Variablen durchaus gegeben - was allerdings bei jeder sozialstrukturellen Analyse der Fall ist, da stets Abhängigkeiten (z. B. Alter und Bildung, erklärbar durch die Bildungsexpansion) existieren. Das ist durchaus ein Problem, das sich aber als unvermeidbar bei der Untersuchung sozialstruktureller Merkmale darstellt. Es auszuklammern hieße, keine sozialstrukturellen Analysen durchzuführen.

Für die Zusammenhangsmaße hinsichtlich der PC-Nutzung ergibt sich:

TABELLE 22 BIVARIATE ZUSAMMENHÄNGE (ETA^2, CRAMERS V (UNSICHERHEITSKOEFFIZIENT)) DER DOMINANTEN SCHICHTUNGSVARIABLEN MIT DER PC-NUTZUNG

Variable	PC-Nutzung	Rang
Alter	eta^2 = 0.072	3
Berufsgruppe	v = 0.502 u = 0.212	1
Bildungsgrad	v = 0.331 u = 0.095	2
Haushaltsnettoeinkommen	v = 0.259 u = 0.054	4

Wichtigste Segmentierung ist der Beruf (Rang 1), gefolgt von Bildungsgrad, Alter und Haushaltsnettoeinkommen.

Da der Beruf deutlichere Zusammenhänge als der Bildungsabschluss mit der PC-Nutzung liefert, sollen als dominante Schichtkriterien Alter, Haushaltsnettoeinkommen und Beruf herangezogen werden - Aus diesen drei Variablen werden Segmente gebildet. Jedoch ist das bei weitem nicht ausreichend, PC-Nutzung zu beschreiben.

Die teilweise unterschiedlichen Segmentierungsergebnisse stellen auch einen Vorteil dar: so könnten Gruppen mit ähnlich hohen Nutzeranteilen über verschiedene Algorithmen identifiziert und dann zusammengefaßt werden. Beispielsweise könnten, neben den Büroberufen und Studierenden mit 90 % Nutzeranteil, ein kleines Segment von Reise- und Dienstleistungsangestellten mit hohem Einkommen und geringem Alter ebenfalls 90 % Nutzeranteil aufweisen. Diese Gruppen lassen sich zusammenfassen. Sind die „Grobgruppen" gebildet, können sie sowohl mit weiteren sozialstrukturellen (Bildung, Familienstand, Geschlecht) Merkmalen als auch mit den Kultur- und Freizeitvariablen untersucht werden. Es ergibt sich dadurch ein sehr differenziertes Schichtungsbild.

3 Multivariate Analyse I: Dominante und subordinierte Variablen

3.1 Dominante Schichtungen der Entscheidungsbäume

3.1.1 EXHAUSTIVE CHAID

(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)

Der EXHAUSTIVE CHAID-Algorithmus liefert mit einer Fehlklassifikation von 24.5 % das beste Ergebnis. Allerdings sind die gefundenen Gruppen recht unspezifisch, da kaum Berufe zusammenfaßt werden. Da die Baumstruktur sehr unübersichtlich und der Informationsgrad recht gering ist, wird an dieser Stelle ausnahmsweise auf die grafische Darstellung verzichtet).

Der Baum hat zwei Ebenen, auf der ersten Ebene wird nach den Berufen segmentiert, auf der zweiten nach Alter - allerdings nur für zwei Berufssegmente:

TABELLE 23 EXHAUSTIVE CHAID-SEGMENTIERUNG: PC-NUTZUNG NACH ALTER, BERUF UND HAUSHALTSNETTOEINKOMMEN (NUR REISE-, DIENSTLEISTUNGSANGESTELLTE, LADENBESITZER, HANDWERKER, ARBEITER)

Berufsgruppe	Alterssegmente
sonstige Angestellte Reise & Dienstleistung, Ladenbesitzer, Handwerker	bis 25 Jahre: 80 % 26 - 48 Jahre: 60 % 49 Jahre +: 35 %
Arbeiter	bis 20 Jahre: 72 % 20 - 48 Jahre: 45 % 49 Jahre +: 19 %

Die Trennung bei 48 Jahren liegt zwischen den Binärsplits von CART (45.5 Jahre) und QUEST (53 Jahre). Interessant ist weiter, dass beide Berufsgruppen exakt nach den gleichen Alterskategorien unterteilt werden. Dadurch wird deutlich, dass die bei CART und QUEST gefundenen Ergebnisse der geringeren Nutzung in diesen Gruppen einen starken Altersbezug haben. Hier wird auch die Neigung von QUEST deutlich, zuerst nach dem Alter zu segmentieren, da der Effekt - zumindest in den Gruppen der weniger rechneraffinen Berufe - altersabhängig ist - in den PC-orientierteren Berufen aber nicht. Hier werden 90 % Nutzung erreicht - unabhängig von einer weiteren Differenzierung.

Wenn ein Beruf ausgeübt wird, der in den meisten Fällen auch einen Einsatz des PCs erfordert, ist der PC-Nutzeranteil logischerweise sehr hoch (ca. 90 %). Mit zunehmendem Alter sinkt der Anteil leicht. Wird ein anderer Beruf (z. B. Arbeiter, Reise- und Dienstleistungsangestell-

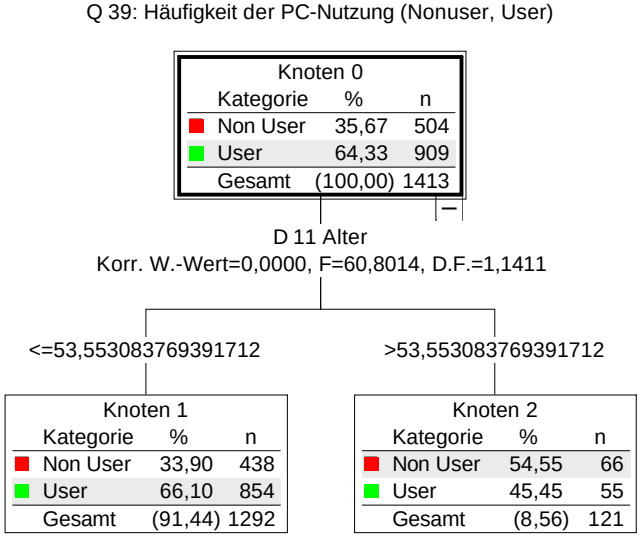
te) ausgeübt, entscheidet das Alter über die Nutzung (z. B. sinken die altersbezogenen PC-Useranteile nach EXHAUSTIVE CHAID bei den Berufen mit geringer PC-Affinität von 80 % auf 35 % bzw. 72 % auf 19 %). Auch ein höheres Einkommen kann den PC-Nutzeranteil deutlich erhöhen (CART: weniger PC-bezogene Berufe, Einkommen <> 4500 DM).

3.1.2 QUEST

QUEST	• vieldiskutierter, relativ neuer Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus (unabhängige Variablen) geeignet • nur für nominale Zielvariablen geeignet, die auch dichotom sein kann • gebräuchliche statistische Kennzahlen (Chi-Quadrat, F-Test) • erlaubt Ersatzprädiktoren, Pruning und a prioris

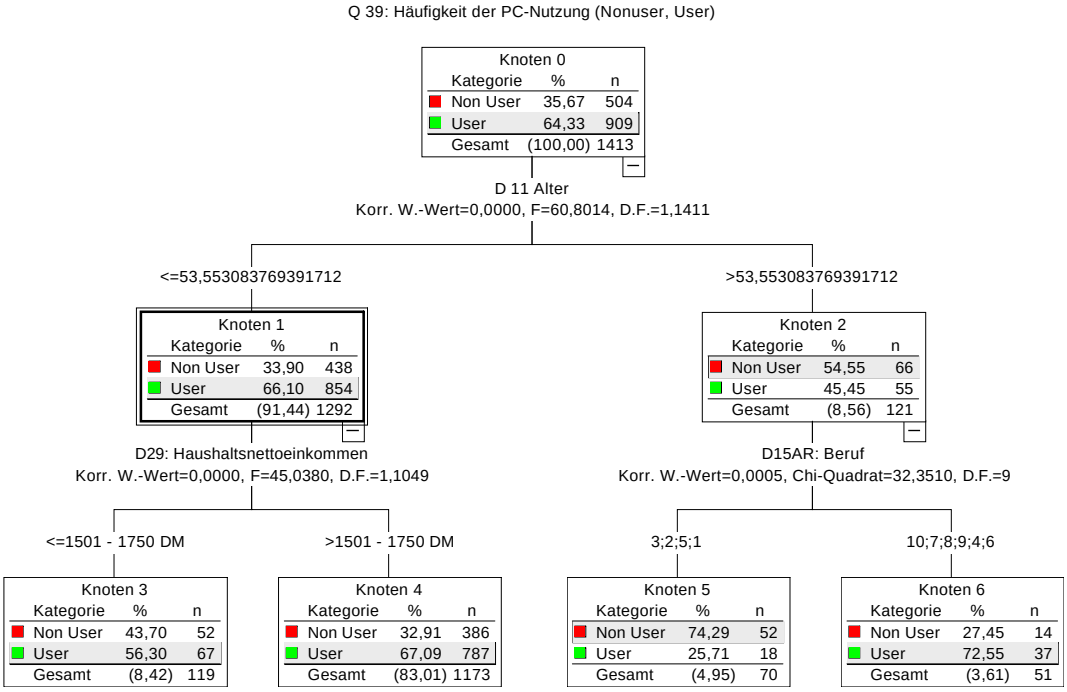
Wie bereits weiter oben dargestellt, zieht QUEST als wichtigste Prädiktorvariable das Alter heran und kommt zu folgender Aufsplittung:

ABBILDUNG 110 QUEST: PC-Nutzung nach Alter, Beruf und Haushaltsnettoeinkommen (einstufig, bis 57 Jahre, N = 1413)



Beim CART-Baum werden auf der ersten Ebene die Büroberufe mit 90 % Nutzeranteil den anderen Berufen mit 44 % Usern gegenübergestellt. Hier wird das Lebensalter als ausschlaggebendste Prädiktorvariable herangezogen. Die Trennung (bis 53/über 53 Jahre) teilt hier die User im Verhältnis 66 : 45 auf. Wie auch schon beim bivariaten Teil angedeutet, versucht hier QUEST über das Alter eine kleine Gruppe (N = 121) mit geringem Nutzeranteil herauszulösen.

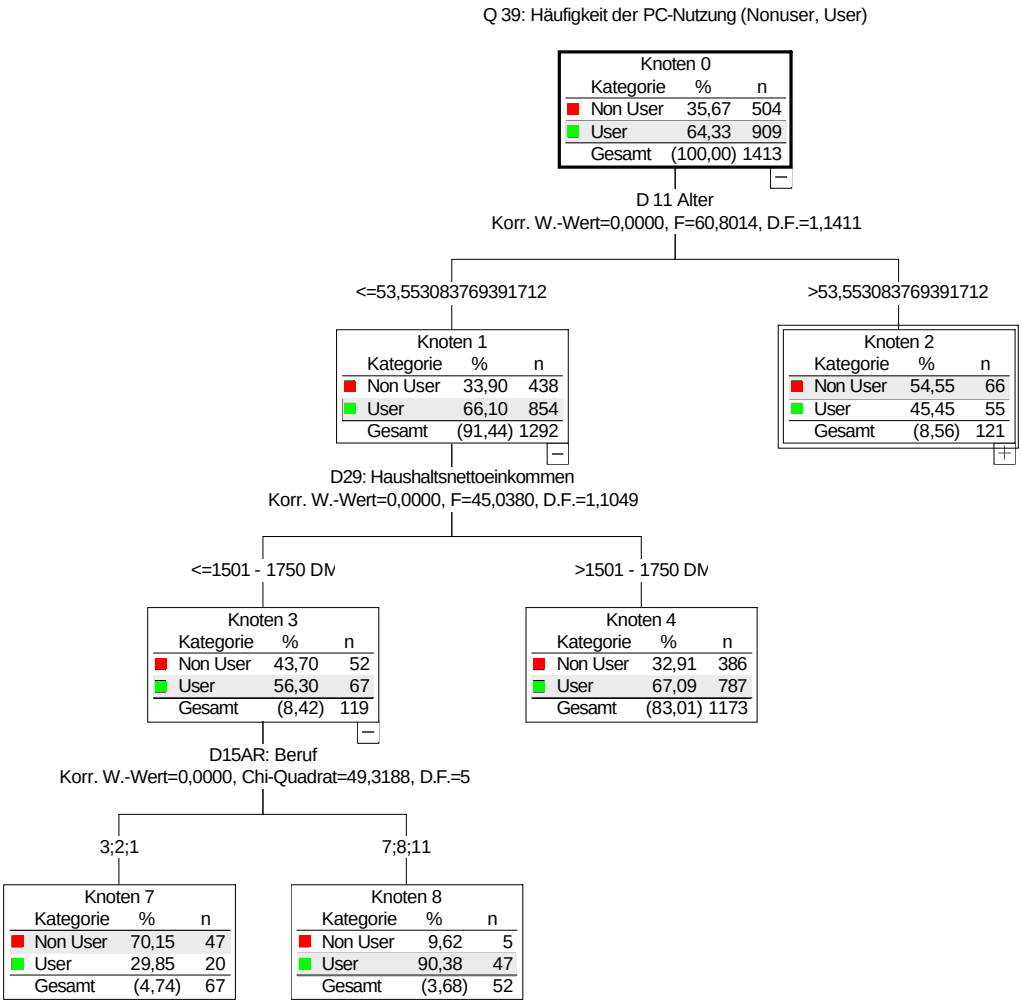
ABBILDUNG 111 QUEST: PC-Nutzung nach Alter, Beruf und Haushaltsnettoeinkommen (zweistufig, bis 57 Jahre, N = 1413)



Hier zeigt sich auch ein weiterer Vorteil der Entscheidungsbäume. Subgruppen werden nach unterschiedlichen Variablen weiter segmentiert. Während sich bei den Jüngeren eher das Haushaltsnettoeinkommen für die PC-Nutzung als ausschlaggebend erweist (Knoten 1, 3, 4), ist es bei den Älteren der Beruf (Knoten 2, 5, 6). Knoten 5 enthält die Gruppen „nie erwerbstätig“, „Arbeiter“, „Reise- und Dienstleistungsangestellte“ sowie „Ladenbesitzer und Handwerker“.

Knoten 6 alle Büroberufe, Studierende, Meister, Landwirte und Fischer sowie Freie Berufe und Unternehmer.

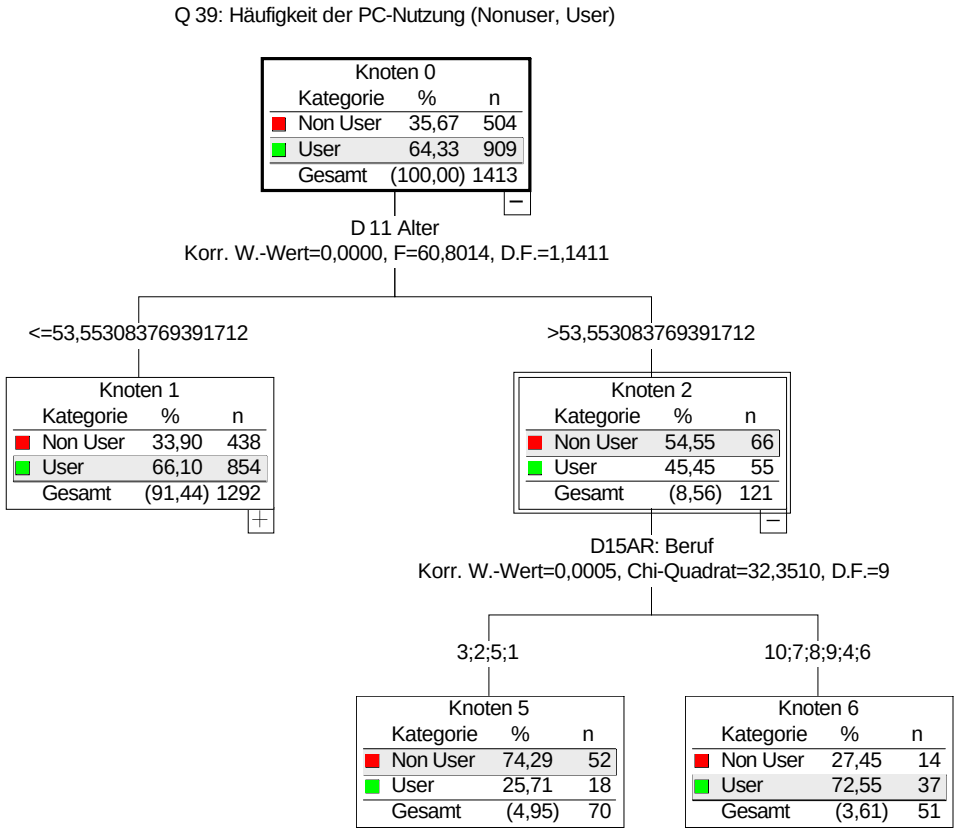
ABBILDUNG 112 QUEST: PC-Nutzung nach Alter, Beruf und Haushaltsnettoeinkommen (dreistufig, jüngere Nutzer, bis 53 Jahre, N = 1413)



Im Knoten 7 und 8 werden die geringeren Einkommensgruppen (bis 1750 DM) nochmals weiter nach dem Beruf aufgesplittet: während die nie Erwerbstätigen, Arbeiter und Reise- und Dienstleistungsangestellten einen Useranteil von knapp 30 % aufweisen (Knoten 7), liegt er bei den Büroberufen und Studierenden dreimal so hoch (Knoten 8).

Die im Knoten 4 enthaltenen jüngeren Befragten mit einem Haushaltsnettoeinkommen von mehr als 1750 DM lassen sich nur noch unbefriedigend weiter hinsichtlich dieses Merkmals segmentieren (Verhältnis: 61 : 68). Diese Aufteilung soll aufgrund ihres geringen Unterschieds nicht weiter berücksichtigt werden.

ABBILDUNG 113 QUEST: PC-Nutzung nach Alter, Beruf und Haushaltsnettoeinkommen (ältere Nutzer, 54 bis 58 Jahre, N = 1413)



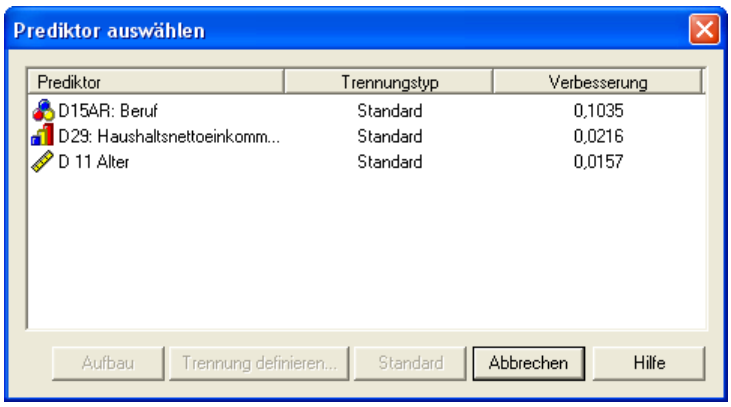
Bei den Befragten, die 54 bis 58 Jahre alt sind, entscheidet der Beruf über die PC-Nutzung. Auch hier liegt der Anteil der Büroberufe, Handwerker und Ladenbesitzer bei etwa 1/3 der restlichen Berufe (25 : 73 %) - ein deutliches Indiz dafür, dass eine gewisse Altersabhängigkeit gegeben ist (Knoten 5 vs. 6).

3.1.3 CART

CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und a prioris

Um den Umgang mit hoch korrelierten Daten zu erläutern, wird ein Vergleich zwischen zwei Modellen durchgeführt: die PC-Nutzung wird zum einen mit den unabhängigen Variablen Alter, Haushaltsnettoeinkommen und Beruf untersucht, zum anderen wird zusätzlich der Bildungsgrad eingesetzt, um zu sehen, wie sich die Kennzahlen ohne bzw. unter Berücksichtigung dieser Variable verändern. Für CART ergibt sich:

ABBILDUNG 114 CART: Prädiktoren bei abhängiger Variable PC-Nutzung, unabhängige Variablen Berufsstellung, Haushaltsnettoeinkommen, Alter und Bildung (nur zweite Darstellung)



Beruf, Haushaltsnettoeinkommen und Alter gehen im ersten Fall, Beruf, Schulbildung, Haushaltsnettoeinkommen und Alter im zweiten Fall in die Analyse ein. Das entspricht in etwa der Reihenfolge im bivariaten Fall.

Der Vergleich der Verbesserungswerte zeigt, dass die berufliche Stellung mit 0.1035 deutlich vor den anderen unabhängigen Variablen liegt und einen sehr starken Einfluss auf die Generierung des Baums hat. Wird die Schulbildung herangezogen, liegt sie an der zweiten Stelle der Verbesserungswerte.

Weiterhin wird deutlich, dass die Verbesserungswerte sich nicht verändern - auch wenn neue Variablen hinzukommen. Alle unabhängigen Variablen werden also isoliert von ihrem Beitrag, bezogen auf die Zielvariable, untersucht, was durchaus sinnvoll ist.

Würde man bei diesem Punkt stehenbleiben - und Entscheidungsbäume genauso interpretieren wie z. B. die Regression - ware das Ergebnis: Die PC-Nutzung ist in ganz entscheidender Weise von dem beruflichen Umgang mit dem Rechner abhängig, andere Variablen tragen kaum zur Verbesserung bei. Durch die zunehmende Selbstverständlichkeit, mit der sich die PC-Nutzung vollzieht, wird sie - vor allem bei den Jüngeren - Teil des Alltagshandelns und in das Leben inte-

griert. Die Fehlklassifikation liegt in beiden Fällen (mit bzw. ohne Schulbildung) bei 25.6 %, die Verbesserung des Gesamtbaums bei rund 0.14, wobei der Hauptanteil durch die berufliche Stellung (0.10) beigetragen wird (zur näheren Interpretation des Baums weiter unten).

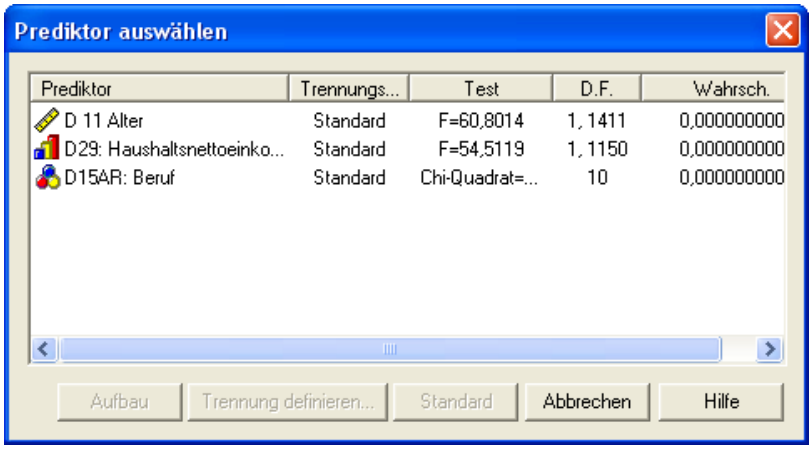
Der EXHAUSTIVE CHAID-Algorithmus generiert ähnliche Prädiktoren:

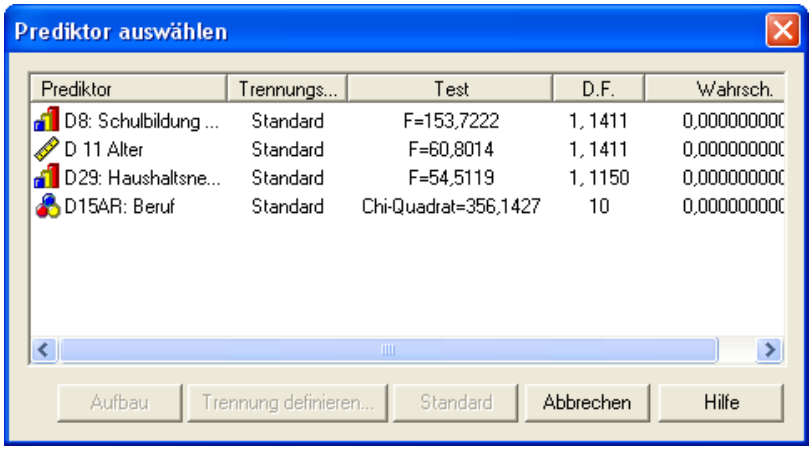
ABBILDUNG 115 EXHAUSTIVE CHAID-Algorithmus: Prädiktoren bei abhängiger Variable PC-Nutzung, unabhängige Variablen Berufsstellung, Haushaltsnettoeinkommen, Alter und Bildung (nur zweite Darstellung

Die Wichtigkeit der Variablen, ausgedrückt in Chi-Quadrat-Werten, unterscheidet sich in der Reihenfolge nicht von CART. Die Fehlklassifikation ist allerdings mit 0.245 (mit Schulbildung: 0.236) etwas besser als CART. Hier trägt die Schulbildung auch zur Reduktion der Fehlklassifikation bei - nicht so bei CART.

Bei der Heranziehung des QUEST-Algorithmus stellt sich allerdings das Ergebnis „genau auf den Kopf“:

ABBILDUNG 116 QUEST-Algorithmus: Prädiktoren bei abhängiger Variable PC-Nutzung, unabhängige Variablen Berufsstellung, Haushaltsnettoeinkommen, Alter und Bildung (nur zweite Darstellung)





In diesem Fall wäre - in der Logik anderer Verfahren - das Alter bzw. die Schulbildung am wichtigsten, gefolgt vom Haushaltsnettoeinkommen. Der Beruf, auf Rang 3 bzw. 4, spielt keine so große Rolle. Die Fehlklassifikation ist mit 0.314 höher als bei CART. Wird die Schulbildung herangezogen, gibt es kaum Unterschiede hinsichtlich der Fehlklassifikation: CART bleibt mit 25.6 % unverändert, QUEST verbessert sich auf 0.263. Aus diesem Grund nur den CART-Algorithmus vorzuziehen, der 0.7 % Fehlklassifikation weniger aufweist, wäre - zumindest soziologisch betrachtet - unsinnig.

Wie läßt sich das erklären? - Die bei CART und QUEST dargestellten Werte unterscheiden sich erheblich: bei CART werden Verbesserungswerte, bei QUEST (und den CHAID-Algorithmen) statistische Kennwerte (Chi-Quadrat, F-Wert) ausgegeben. Weiterhin gibt es folgende Unterschiede (vgl. SPSS (o. J.: 4)):

> „It is well-known that exaustive search methods such as C&RT tend to select variables with more discrete values, which can afford more splits in the tree growing process. This introduces bias into the model, which reduces the generalizability of results. Another limitation of C&RT is the computational investment in searching for splits. QUEST method is designed to address these problems. QUEST was demonstrated to be much better than exhaustive search methods in terms of variable selection bias and computational cost. In terms of classification accuracy, variability of split points and tree size, however, there is still no clear winner when univariate splits were used."

Hauptunterschied der beiden Verfahren ist der Anspruch von QUEST, bessere Trennungen zu finden und damit realitätsgerechtere Bäume zu generieren. QUEST trägt dem Rechnung, indem Variablenauswahl und Splittung der Variablen nicht gleichzeitig wie bei CART, sondern nacheinander erfolgt. Dies trifft - für dieses Beispiel - jedoch nicht zu: CART segmentiert „inhaltlich" besser - was die Ergebnisse der anderen Algorithmen nicht diskreditiert. In diesem Fall schlägt SPSS vor, CART dann einzusetzen, wenn es - im Vergleich zu den anderen Verfahren - den besten Entscheidungsbaum liefert (vgl. SPSS (o. J.: 4)).

Eine Wissenschaft wie die Soziologie sollte sich nicht von reinen Zahlen (vor allem nicht hinter dem Komma), sondern von realitätsgerechten Ergebnissen leiten lassen. Bei ähnlichen Fehlklassifikationen ist es auf jeden Fall sinnvoll, die Ergebnisse gründlich zu vergleichen und die gefundenen Resultate zu sammeln: auch wenn zwei Verfahren unterschiedliche Ergebnisse liefern, muss nicht unbedingt ein Ergebnis falsch und eines richtig sein, es kann durchaus sein, dass unterschiedliche Modelle zur Erhellung eines Zusammenhangs beitragen. Zusammenfassend die Ergebnisse der bivariaten Auswertung:

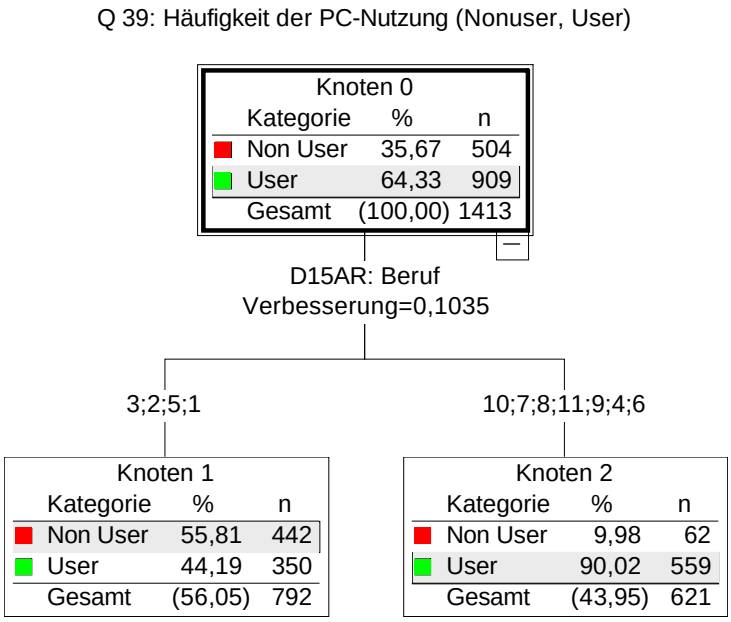
- Der Beruf liefert mit einem Zusammenhangswert von 0.5 (Unsicherheitskoeffizient: 0.21) mit der PC-Nutzung den höchsten Zusammenhangswert, gefolgt von Bildung, Alter und Haushaltsnettoeinkommen.
- Andere sozialstrukturellen Variablen (Geschlecht, Familienstand) liefern keine hohen Zusammenhangswerte mit PC-Nutzung
- Der Beruf ist in hohem Maße abhängig vom erreichten Bildungszertifikat.

Der wichtigste Punkt bei Entscheidungsbäumen ist, mathematisch gesehen, die Frage, nach der Fehklassifikation, der Verbesserung (bei CART) sowie evtl. die Gewinne. Soziologisch ist interessanter, welche Segmente die Algorithmen finden und ob sich die gefundenen Gruppen ähneln oder nicht.

Ausgangspunkt ist die Lösung mit drei unabhängigen Variablen: Beruf, Haushaltsnettoeinkommen und Alter. Der CART-Baum präsentiert folgende Segmente:[87]

87. Zur besseren Übersichtlichkeit wurden die Werte der einzelnen Variablenausprägungen angegeben, da Answertree bei langen Labels keine Umbrüche vornimmt und somit die Bäume schlecht dargestellt werden können. Die Werte werden im Text erläutert.

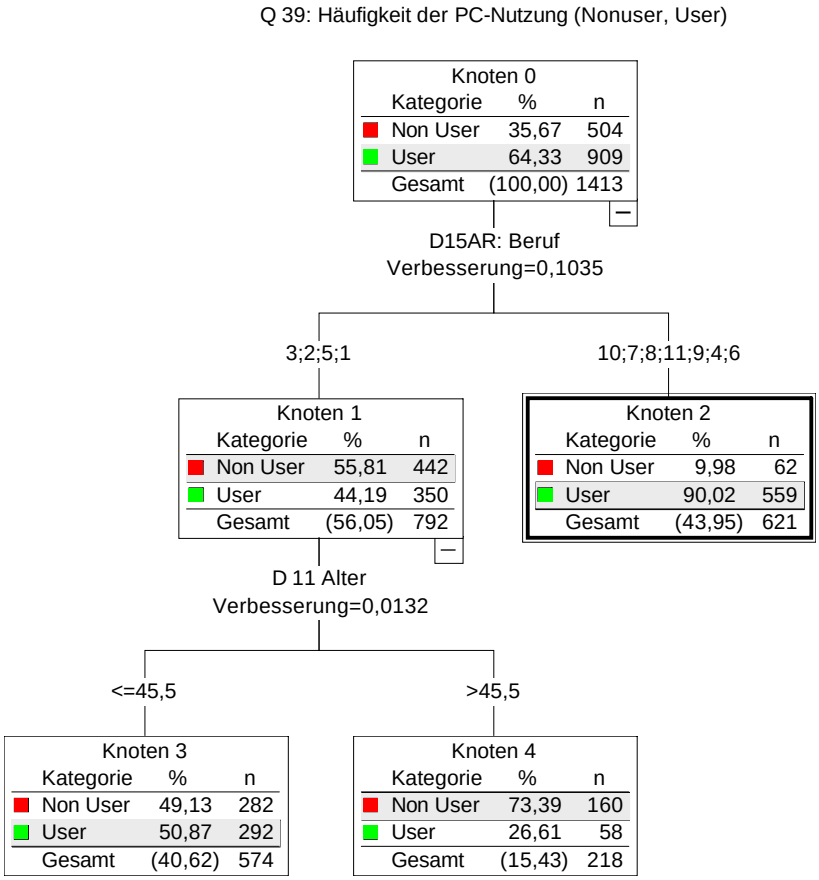
ABBILDUNG 117 CART: PC-Nutzung nach Alter, Beruf und Haushaltsnettoeinkommen (einstufig, bis 57 Jahre, N = 1413)



Knoten 1 enthält die Gruppen der nie Erwerbstätigen, Arbeiter, Ladenbesitzer, Handwerker und Reise- und Dienstleistungsangestellten. Die Nutzung ist mit 44 % nicht besonders hoch - im Gegensatz zu Knoten 2 (90 %). In diesem Segment sind alle Büroberufe, Freien Berufe, Studierenden, Meister, Landwirte und Fischer enthalten. Die Prozentsatzdifferenz zwischen den Knoten liegt bei 46 %. Der rechte Ast kann aufgrund der geringen Fallzahlen der Non-User nicht weiter segmentiert werden.

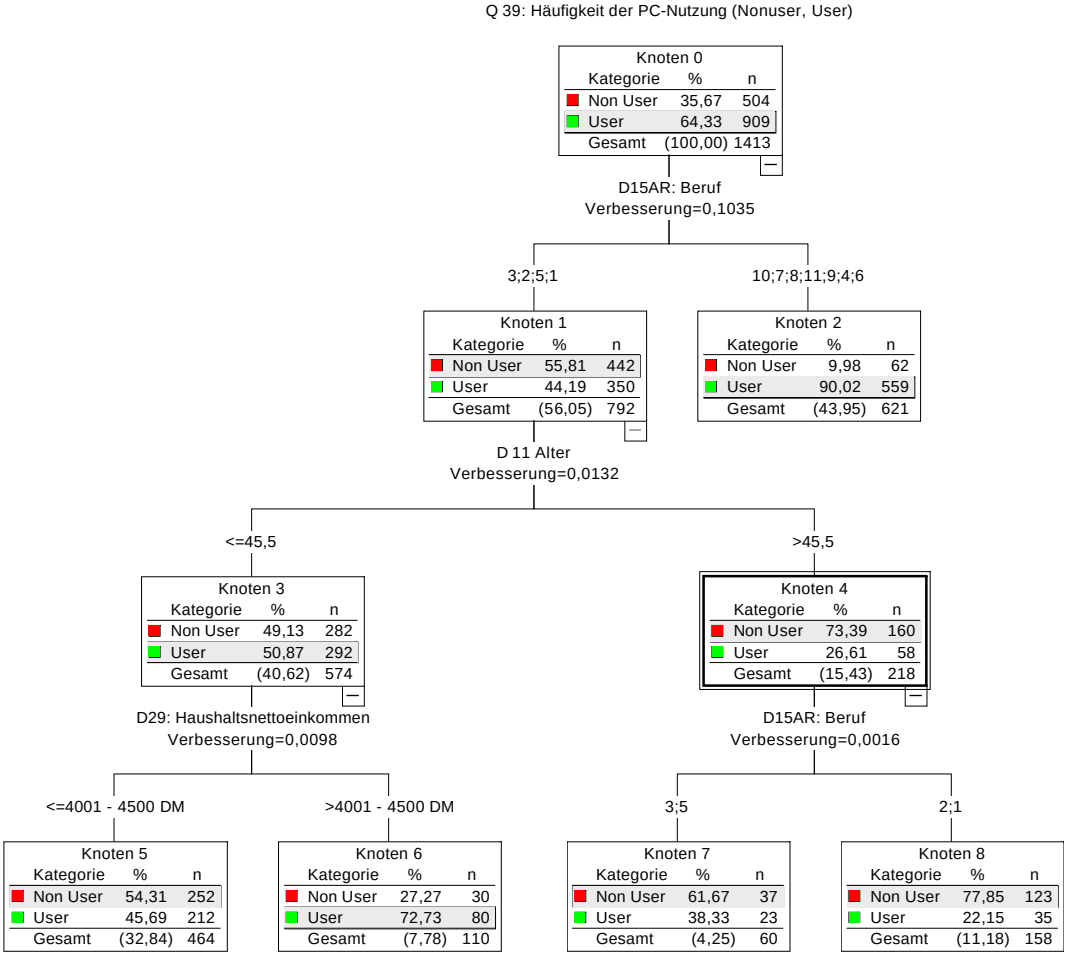
Knoten 1 weist mit einem Verhältnis von 44 zu 55 auf eine noch recht unbefriedigende Trennung hin, die auf der zweiten Stufe durch das Alter konkretisiert wird. Von den Vorhersagewerten für den Gesamtbaum trägt zwar das Haushaltsnettoeinkommen mehr bei, für dieses Segment gilt dies aber nicht.

ABBILDUNG 118 CART-Baum: PC-Nutzung nach Alter, Beruf und Haushaltsnettoeinkommen (zweistufig, bis 57 Jahre, N = 1413)



Die - bereits aus dem deskriptiven Fall - bekannte Trennung bei 45.5 Jahren ändert das Bild: bei den Jüngeren liegt der Anteil der Nutzer bei 51, bei den Älteren bei 27 %.

ABBILDUNG 119 CART-Baum: PC-Nutzung nach Alter, Beruf und Haushaltsnettoeinkommen (bis 57 Jahre, N = 1413)



Die bis 45jährigen werden besser durch das Haushaltsnettoeinkommen, die über 45jährigen durch den Beruf segmentiert (Knoten 5 bis 8). Ist die Höhe des Haushaltsnettoeinkommens über 4000 DM (Knoten 6) für einen Nutzeranteil von knapp 73 % verantwortlich, verwenden nicht einmal die Hälfte (Knoten 5, 46 %) den Rechner. Fast jeder vierte aus der Berufsgruppe der Angestellten der Reise- und Dienstleistungsgewerbe und Ladenbesitzer bzw. Handwerker sind Nutzer, die Arbeiter und nie Erwerbstätigen gehören mit 22 % zur Gruppe mit dem höchsten Nichtnutzeranteil.

Deutlicher wird das Ergebnis, wenn konkrete Gruppen gebildet werden. Dies kann auf unterschiedliche Weise (von oben nach unten oder von unten nach oben) geschehen. Für diese Arbeit wird von den Endknoten 2 und 5 - 8 begonnen, die Gruppen zusammenzufassen. Hier zeigt sich letztendlich, ob eine Segmentierung gelungen ist und aussagekräftige Ergebnisse liefert - oder nicht:

TABELLE 24 CART-SEGMENTE: PC-NUTZUNG NACH ALTER, BERUF UND HAUSHALTSNETTOEINKOMMEN

Gruppe	Nutzeranteil	Beschreibung
Knoten 2	90 %	Büroberufe, Freie Berufe, Studierende, Meister, Landwirte und Fischer ohne weitere Merkmale
Knoten 5	46 %	Einkommen unter 4500 DM, jünger als 46 Jahre, nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker
Knoten 6	73 %	Einkommen über 4500 DM, jünger als 46 Jahre, nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker
Knoten 7	38 %	Reise- und Dienstleistungsberufe, Ladenbesitzer, Handwerker, älter als 45 Jahre
Knoten 8	22 %	Arbeiter, nie Erwerbstätige, älter als 45 Jahre

Die Nutzung liegt also innerhalb der Gruppen zwischen 22 % (Knoten 8: „ältere Arbeiter oder nie Erwerbstätige“) und 90 % (Knoten 2 Büroberufe, Freie Berufe, Studierende, Meister, Landwirte, Fischer). Bei den jüngeren Befragten erhöht sich dieser Wert bis auf 73 % (Knoten 6) - wenn das Einkommen relativ hoch ist. Die Nutzung des PCs scheint also in nicht-PC-affinen Berufen mit der Höhe des Haushaltsnettoeinkommens zusammenzuhängen. Die Differenz zur weniger verdienenden, aber sonst identischen Gruppe in Knoten 5 beträgt immerhin 27 %.

Die jüngeren Befragten aus Knoten 5 unterscheiden sich allerdings nicht wesentlich von den Älteren mit Reise- bzw. Dienstleistungsberu-

fen, Ladenbesitzer und Handwerker (Userverhältnis: 46 : 38 %). Das bedeutet, dass das Alter in diesem Bereich weniger eine Rolle spielt, vielmehr der Beruf oder das geringere Haushaltsnettoeinkommen.

Die Gruppe der Arbeiter bzw. niemals Erwerbstätigen mit einem Nutzeranteil von 22 % stellt die kleinste Userkategorie dar. Hier könnten fehlende berufliche bzw. fehlende finanzielle Anreize verantwortlich sein.

Auf den ersten Blick wird deutlich, dass sich die Endsegmente zwar etwas, aber nicht wesentlich unterscheiden - obwohl CART zuerst den Beruf, QUEST zuerst das Alter als Segmentierungsvariable herangezogen hat. Insofern ist nicht der Blick auf die Prädiktoren die entscheidendste Frage, sondern auf die gefundenen Endsegmente. QUEST liefert folgende Ergebnisse:

TABELLE 25 QUEST-SEGMENTE: PC-NUTZUNG NACH ALTER, BERUF UND HAUSHALTSNETTOEINKOMMEN

Gruppe	Nutzeranteil	Beschreibung
Knoten 4	67 %	Einkommen > 1750 DM, Befragte bis 54 Jahre
Knoten 5	26 %	nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsberufe, Handwerker, Ladenbesitzer, 54 Jahre und älter
Knoten 6	73 %	Meister, Landwirte, Fischer, Büroberufe, 54 Jahre und älter
Knoten 7	30 %	nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsberufe, jüngere Befragte bis 54 Jahre, Einkommen bis 1750 DM
Knoten 8	90 %	Büroberufe, Studierende, Einkommen bis 1750 DM, bis 54 Jahre

QUEST ermittelt ebenfalls vier relevante Gruppen. Die aus Knoten 4 resultierende weitergehende Segmentierung nach dem Haushalts-

nettoeinkommen ergibt mit einem Verhältnis von 61 : 68 keine befriedigende Trennung und soll hier unberücksichtigt bleiben.

Der Vergleich der beiden Segmentierungsalgorithmen ergibt:

TABELLE 26 VERGLEICH DER CART- UND QUEST-SEGMENTE: PC-NUTZUNG NACH ALTER, BERUF UND HAUSHALTSNETTOEINKOMMEN (IN PROZENT)

Anteil	CART-Gruppen	Anteil	QUEST-Gruppe
90 %	Büroberufe, Freie Berufe, Studierende, Meister, Landwirte und Fischer ohne weitere Merkmale	90 %	Büroberufe, Freie Berufe, Studierende, Meister, Landwirte und Fischer Einkommen bis 1750 DM und bis 54 Jahre
73 %	Einkommen über 4500 DM, jünger als 46 Jahre, nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker	73 %	Meister, Landwirte, Fischer, Büroberufe, 54 Jahre und älter
46 %	Einkommen unter 4500 DM, jünger als 46 Jahre, nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker	67 %	Einkommen > 1750 DM, Befragte bis 54 Jahre
38 %	Reise- und Dienstleistungsberufe, Ladenbesitzer, Handwerker, älter als 45 Jahre	30 %	nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsberufe, jüngere Befragte bis 54 Jahre, Einkommen bis 1750 DM
22 %	Arbeiter, nie Erwerbstätige, älter als 45 Jahre	26 %	nie erwerbstätig, Arbeiter, Reise- und Dienstleistungsberufe, Handwerker, Ladenbesitzer, 54 Jahre und älter

Durch die jeweils wichtigste Prädiktorvariable werden die Gruppen stark auf dieses Merkmal bezogen: während es bei CART die Berufe sind, ist es bei QUEST das Alter. Beide Algorithmen segmentieren fünf Gruppen, die sich von den Prozentsätzen teilweise gleichen oder ähneln.

CART erkennt, dass die Freien und Büroberufe, Studierende, Meister, Landwirte und Fischer einen Nutzeranteil von 90 % aufweist. Diese Gruppe wird von der Restgruppe segmentiert. Der Informationsgehalt könnte durch weitere Segmentierungen erhöht werden, allerdings nur, wenn sich die Gruppen weiter segmentieren lassen, was durch eine zu geringe Fallzahl oder auch durch eine hohe Konzentration, wie in diesem Fall, nicht realisiert werden kann.

Auch QUEST identifiziert eine Gruppe mit einem Nutzeranteil von 90 % - allerdings mit einem anderen Informationsgehalt: hier finden sich Befragte bis 1750 DM Einkommen, die jünger als 54 Jahre alt sind. Die segmentierten Berufsgruppen (nie Erwerbstätige, Arbeiter, Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker vs. Büroberufe, Studierende, Meister, Landwirte, Fischer) werden von CART und QUEST gleich segmentiert.

Das zweite Segment, sowohl bei CART als auch bei QUEST mit jeweils 73 % Nutzeranteil identifiziert, unterscheidet sich deutlich: bei CART finden sich hier die Jüngeren (unter 46 Jahre) mit einem Einkommen von 4500 DM und mehr - und gerade hier finden sich nicht die PC-affinen Berufsgruppen. Alter und Haushaltsnettoeinkommen „gleichen" hier den nicht rechnerbezogenen Beruf aus. QUEST findet hier die Berufsgruppen mit deutlich hohem Nutzeranteil über 54 Jahre. Das Alter senkt hier den Anteil der Nutzer auf 73 % und ist inhaltlich - im Vergleich mit CART - nicht sehr hilfreich. Allerdings macht dieses Ergebnis deutlich, dass es durchaus Unterschiede zwischen diesen Berufsgruppen hinsichtlich des Alters existieren: dass Ältere z. B. mit Büroarbeitsplätzen doch nicht unbedingt den PC nutzen. 73 % Nutzeranteil ergeben sich entweder aus einem höheren Lebensalter und Berufen, die häufig den PC-Einsatz erfordern - oder eher dem Alter (jünger als 46 und einem höheren Haushaltsnettoeinkommen > 4500 DM).

27 % Differenz ergibt sich zwischen den nicht auf Rechner bezogenen Berufen, die ein höheres Nettoeinkommen erzielen, jüngeren Befragten und der gleichen Gruppe mit niedrigerem Einkommen. An diesem Punkt entscheidet das Haushaltsnettoeinkommen deutlich über die Nutzung.

Bei QUEST ist dies genau entgegengesetzt: die 73 % Nutzeranteil generieren sich aus rechnerbezogenen Berufen, aber durch ein höheres Lebensalter. Dem schließen sich im dritten Segment mit einem Nutzeranteil von 2/3 die Jüngeren mit einem Einkommen > 1750 DM an. Der Nutzeranteil liegt in beiden Segmenten recht nahe zusammen.

Mit einem Prozentanteil von 38 : 22 unterscheiden sich die zwei letzten CART-Segmente deutlicher als die QUEST-Gruppen mit 30 : 26. Die Gruppen enthalten nicht PC-affine Berufe, einmal jünger mit geringerem Einkommen, einmal älter ohne weitere Differenzierung. Diese Gruppen sind inhaltlich kaum zu trennen.

Um die Bäume zu beurteilen, kann die Fehlklassifikation herangezogen werden. CART liegt mit 26 % Fehklassifikation vor QUEST mit 31 %. Mathematisch ist das Ergebnis eindeutig: CART ist in diesem Fall der „bessere“ Algorithmus. Als weitere Kriterien können die Prozentsatzdifferenzen zwischen den Gruppen herangezogen werden, die bei CART ebenfalls deutlicher ausfallen. Nicht zuletzt sollte allerdings die realitätsgerechte Überprüfung der gefundenen Gruppen berücksichtigt werden.

Trotzdem liefert auch QUEST einige Zusatzinformationen. Erst im Zusammenhang ergeben sich tiefere Einblicke in den Datensatz. Zusammenfassend nochmal die Ergebnisse:

Beide Algorithmen liefern identische Berufssegmentierungen - also können rechnerbezogene Berufe anhand des Datensatzes identifiziert werden.

QUEST trägt folgenden Informationen bei:

- Auch mit geringem Einkommen (bis 1750 DM) und geringerem Alter liegt bei rechnerbezogenen Berufen der Nutzeranteil bei 90 %
- Bei diesen Berufen ist der Anteil der älteren Nutzer mit 73 % hoch - höher als bei den jüngeren Befragten mit höherem Einkommen (67 %). Hier entscheidet klar der Beruf gegenüber den beiden anderen Variablen
- Bei geringerem Einkommen oder höherem Lebensalter ohne rechnerbezogenen Beruf ist die Nutzung mit 22 % bzw. 30 % eher gering

CART-Ergebnisse:

- Der Beruf liefert beim CART-Algorithmus eine überzeugende Trennung, die Strukturierung ist klar definiert: Auf einer ersten Stufe werden die PC-affinen Berufe von den anderen Berufen getrennt und zu einem Knoten zusammengefaßt. Die anderen Berufe werden weiter segmentiert
- Höheres Einkommen und geringeres Lebensalter gleichen zu einem gewissen Maß den nicht PC-bezogenen Beruf „aus"
- Geringeres Einkommen und höheres Lebensalter lassen den Anteil der PC-Nutzung sinken - vor allem bei den nie Erwerbstätigen und den Arbeitern.

Auch wenn es zunächst so aussah, als würden sich die Ergebnisse der Algorithmen deutlich unterscheiden, finden sich doch viele Gemeinsamkeiten, die sich ergänzen.

Der Beruf ist bei beiden Algorithmen das deutlichste Unterscheidungsmerkmal - auch wenn es bei QUEST nicht als wichtigste Prädiktorvariable herangezogen wird, ist diese Variable in der Gruppenbildung sehr präsent - vor allem, da die beiden Algorithmen zu den gleichen Berufssegmenten kommen.

3.1.4 Zusammenfassung

Die eingesetzten Algorithmen liefern Ergebnisse, die sich durchaus ergänzen können. Auch wenn es auf den ersten Blick so aussieht, dass sich die Bäume deutlich hinsichtlich der Ergebnisse unterscheiden, ergeben sich doch viele Gemeinsamkeiten - aber auch deutliche Un-

terschiede. Der Informationsgehalt steigt - zumindest in diesem Fall - mit der Zahl der eingesetzten Verfahren. An den Ergebnissen werden nochmals die Stärken und Schwächen deutlich: während die Schwächen von EXHAUSTIVE CHAID darin liegen, dass die Berufe nicht effizient zusammengefaßt werden und damit viele Berufsknoten nicht weiter segmentiert werden können, fassen CART und QUEST manchmal auch recht „unsanft“ Berufsgruppen zu Binärsplits zusammen. Dies wird deutlich am EXHAUSTIVE-CHAID-Ergebnis, dass sich vor allem die weniger nutzenden Berufsgruppen sich deutlich hinsichtlich des Alters unterscheiden - ein Ergebnis, das so nicht mehr von den dichotomisierten Bäumen herausgearbeitet werden kann - dafür werden wesentlich mehr Segmente gefunden.

Ausgehend von den drei unabhängigen Variablen Alter, Haushaltsnettoeinkommen und Beruf bleibt festzuhalten, dass bei den Jüngeren (bis etwa 25 Jahren) vielleicht der Beruf oder das Haushaltsnettoeinkommen die Nutzeranteile (geringfügig) erhöhen können, jedoch der PC zum Alltag gehört (ca. 70 - 90 %) - aus Gründen der Übersichtlichkeit wurde für die nachfolgende Tabelle das Alter in Gruppen zusammengefaßt.

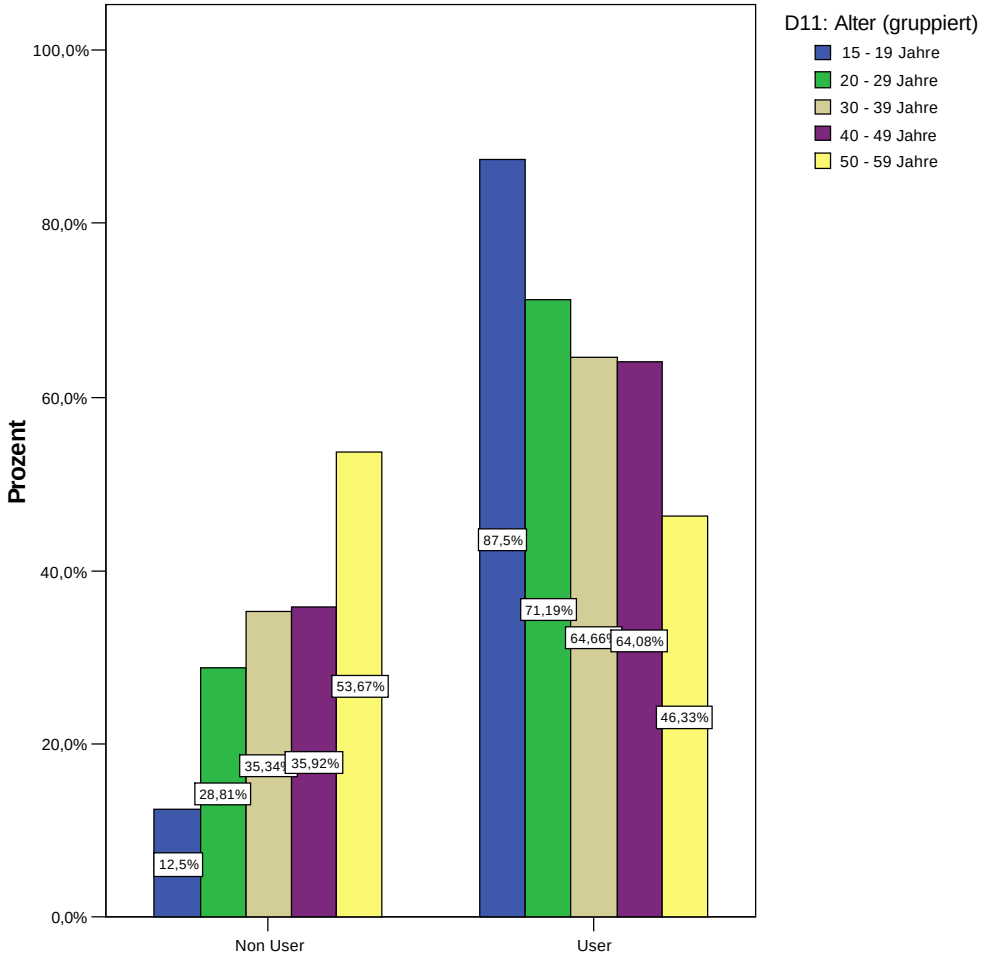
ABBILDUNG 120 Altersspezifische PC-Nutzung (N = 1413, Zeilen-%, gruppierte Altersvariable)

D11: Alter (gruppiert) * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

			Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)		
			Non User	User	Gesamt
D11: Alter (gruppiert)	15 - 19 Jahre	Anzahl	14	98	112
		% von D11: Alter (gruppiert)	12,5%	87,5%	100,0%
	20 - 29 Jahre	Anzahl	85	210	295
		% von D11: Alter (gruppiert)	28,8%	71,2%	100,0%
	30 - 39 Jahre	Anzahl	141	258	399
		% von D11: Alter (gruppiert)	35,3%	64,7%	100,0%
	40 - 49 Jahre	Anzahl	125	223	348
		% von D11: Alter (gruppiert)	35,9%	64,1%	100,0%
	50 - 59 Jahre	Anzahl	139	120	259
		% von D11: Alter (gruppiert)	53,7%	46,3%	100,0%
Gesamt		Anzahl	504	909	1413
		% von D11: Alter (gruppiert)	35,7%	64,3%	100,0%

Dieses Ergebnis unterstreicht nochmals die Deutlichkeit, mit der der PC gerade in den jüngsten Alterssegmenten genutzt wird (knapp 90 % vs. rund 70 % bei den 20 - 29jährigen). Immerhin fast 2/3 der Befragten zwischen 30 bis 49 Jahren und knapp die Hälfte der 50 - 58jährigen nutzt diese Technologie. Deutlicher wird dies noch an einem Balkendiagramm:

ABBILDUNG 121 Grafische Darstellung: PC-Nutzung nach Alter (gruppiert, in %, N = 1413)



Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)

Deutlich wird, dass die Nutzeranteile in de jüngsten betrachteten Gruppe am höchsten ist und sich zwischen 20 - 49 Jahre relativ konstant bei einem Nutzeranteil von rund 64 % - 72 % bewegt. In der letzten Gruppe - diejenigen, die kurz vor dem Rentenübergang stehen - sinkt der Anteil auf rund 46 % ab.

Das legt die Vermutung nahe, dass der PC - zumindest in einigen Jahren - für alle ein Alltagsmedium werden wird - und er ist es heute schon für viele, vor allem jüngere Menschen. Nicht ein Lebensstil (Konsumstil) läßt die Menschen den PC nutzen, sondern Alltagshan-

deln (bei den jüngsten Befragten bis 25, wo der PC zum Leben dazugehört), konkrete Anwendungsfragestellungen (deutlich berufliche Anforderungen bei den bis 45- bzw. 58jährigen). Die Älteren, die weder mit dem PC aufgewachsen sind, noch irgendeine Notwendigkeit sehen, den Rechner zu nutzen, haben wenig Interesse an diesem Medium.

Die Variablen Alter und Beruf für sich genommen - ergeben alltägliche Ergebnisse: die PC-Nutzung ist vom Alter oder vom PC-Einsatz am Arbeitsplatz abhängig. im Zusammenspiel ergeben sich jedoch interessante Konstellationen - vor allem, wenn man die Unabhängigkeit des Geschlechts heranzieht.

Bleibt zum Abschluß die Frage nach der Schulbildung. Nachdem es eine hohe Korrelation zwischen Schulbildung und Beruf gibt, ist es nicht ratsam, die beiden Variablen zusammen in eine Analyse aufzunehmen. Wenn anstatt des Berufs nun der Bildungsabschluss herangezogen wird, ergibt sich eine etwas schlechtere Fehlklassifikation von 0.312 (EXHAUSTIVE CHAID) bzw. 0.316 (CART). QUEST liegt bei 32.7 %. Das läßt sich leicht dadurch erklären, dass der Beruf sehr konkret den PC-Einsatz erfordert - oder den Zugang zu diesem Medium erleichtert. Der Bildungsabschluss selbst ist nicht unbedingt Ausdruck der PC-Nutzung - auch wenn die Schulbildung bei allen drei Algorithmen als Prädiktorvariable an erster Stelle steht.

Die letztendliche Frage - welcher Algorithmus nun „der beste“ sei - läßt sich nicht beantworten. Grundsätzlich clustern alle Verfahren etwas unterschiedlich - häufig mit ähnlichen Ergebnissen (z. B. Fehlklassifikation). Aus diesem Grund ist es, um ein differenziertes Bild zu erhalten, sinnvoll, sich alle Segmentierungen anzusehen. Nachfolgend jedoch gebe ich einige Hinweise zur Auswahl des sinnvollsten Algorithmus - nicht nach Vor- oder Nachteilen, sondern nach Eigen-

schaften: Die Forschungsfrage und der Umgang mit dem Datenmaterial sollte im Vordergrund stehen:

- Wie ist die Zielvariable skaliert? - Bei nicht nominaler Zielvariable kann der QUEST-Algorithmus nicht herangezogen werden.
- Wie sind die zu untersuchenden Daten skaliert? - Besitzen sie viele oder wenige Ausprägungen? - Möglicherweise kommen CHAID und EXHAUSTIVE CHAID nicht zu einer befriedigenden Zusammenfassung vieler Variablenausprägungen. In diesem Fall wären die binären Algorithmen vorzuziehen - immer jedoch mit dem Wissen, dass zwei Kategorien nicht immer die Realität wiederspiegeln.
- Weisen die Variablen viele fehlende Werte auf? - In diesem Fall wären CART und QUEST von Vorteil da sie mit Ersatzprädiktoren arbeiten.
- Möchte ich neben der Fehlklassifikation ein Maß für den „Modellfit", der vergleichbar ist - In diesem Fall ist CART vorzuziehen, da Inhomogenitätsmaße ausgegeben werden.

Dies ist sicherlich keine abschließende Darstellung verschiedener Vorgehensweisen. Sich auf einen Algorithmus zu beschränken, wäre sicherlich häufig nicht sehr hilfreich - Forschung sollte auch mit Neugier einhergehen und neue Bäume mit anderen Algorithmen sind nur mit ein paar Mausklicks (also mit minimalem zeitlichen Aufwand) erstellt.

Das Ergebnis sollte überzeugen - wenn sich zum Beispiel herausstellt, dass die von CART und/oder QUEST zusammengefaßten Kategorien keinen Sinn ergeben, sollte man auf (EXHAUSTIVE) CHAID zurückgreifen.

3.1.5 Exkurs: Befragte über 57 Jahre

Die Gruppe der älteren Personen umfaßt 625 Befragte, wovon 75 (= 12 %) den Rechner nutzen. Ein multivariates Verfahren anzuwenden, ist problematisch, da diese Subpopulation sehr ungleichgewichtig (88 : 12) ist und mit N = 75 für die Nutzer daher die Grenzen multivariater Statistik erreicht sind. Deshalb soll an dieser Stelle eine kurze deskriptive Beschreibung erfolgen:

TABELLE 27 STATISTISCHE KENNWERTE (ETA, CRAMERS V, UNSICHERHEITSKOEFFIZIENT) FÜR DIE GRUPPE DER ÄLTEREN BEFRAGTEN (AB 58 JAHRE, N = 625

unabhängige Variable	Phi/Cramers v (Unsicherheitskoeffizient)
Schulbildung	v = 0.361, u = 0.144
Beruf	v = 0.349, u = 0.145
Haushaltsnettoeinkommen	v = 0.361, u = 0.164
Geschlecht	phi = - 0.182, u = 0.046
Alter	Eta^2 = 0.1116

Haushaltsnettoeinkommen und Schulbildung spielen hier mit einem Zusammenhangswert von jeweils 0.361 (Cramers v) die bedeutsamste Rolle. Der Unsicherheitskoeffizient ist beim Haushaltsnettoeinkommen höher als bei der Schulbildung, was darauf hindeutet, dass vor allem die finanziellen Mittel vorhanden sein müssen, um sich im (Renten-)Alter mit dem Rechner zu befassen. Im Zusammenhang mit dem (früher) ausgeübten Beruf ist erkennbar, dass es sich hierbei auch um leitende, vor allem auch Büroberufe handelt.

Da es keine konkreten beruflichen oder lebensstilspezifischen Anforderungen an die Nutzung dieser Technologie gibt, wird auf eine Anschaffung häufig verzichtet. Ähnlich ist es wahrscheinlich in jüngeren Jahren, wo z. B. auf medizinische Hilfsmittel kein Wert gelegt wird, da sie für den Alltag nicht notwendig sind. Deutlich wird dies z. B. auch bei der Anschaffung von Walk-, Disc- bzw. mp3-Playern. Eta^2 liegt hier hinsichtlich des Alters bei 0.1275, da vor allem Jüngere diese Technik nutzen. Weitergehende Schlüsse sind an dieser Stelle allerdings nicht ratsam. Die häufig verbreitete Ansicht, dass Ältere mit der Technik nicht mehr Schritt halten können, kann so nicht behauptet werden, da der Besitz anderer technischer Geräte (z. B. Fotoapparat, Videokamera) nicht so starke altersspezifische Züge aufweisen. Ob -

und in welchem Umfang - diese Geräte genutzt werden, ist nicht Fragestellung dieser Arbeit. Eine endgültige empirische Erfassung ist schwierig. So könnte es sein, dass der Umgang mit einer Videokamera weniger vom Alter, sonder vielmehr von der Urlaubshäufigkeit abhängt. Das müßte weiteren Untersuchungen vorbehalten bleiben.

3.1.6 Inhaltliches Fazit

Die Analyse hat gezeigt, dass nicht das Alter, wie in der Gesamtstichprobe, für die PC-Nutzung ausschlaggebend ist, sondern sich ein komplexeres Geflecht aus den drei Variablen ergibt: bei den jüngsten Befragten ist das Alter ausschlaggebend, bei den ca. 25 - 45jährigen der Beruf und bei den bis 58jährigen der Beruf und das Einkommen, wobei letztere Variable die PC-Nutzung nochmals erhöht, auch bei nicht PC-bezogenen Berufen.

Auch wenn sich einige unerwünschte Korrelationen ergeben (vor allem zwischen Bildung und Beruf) - sozialstrukturelle Analysen haben stets mit diesem Manko zu kämpfen, da der ausgeübte Beruf in gewissen Sinne von Alter und Bildungsgrad abhängig ist, der Bildungsgrad vom Alter, etc.

Das Ergebnis selbst ist sicherlich nicht überraschend. Durch die multivariate Analyse zeigt sich jedoch, dass gerade in den Berufen, wo der Nutzeranteil sehr niedrig liegt, das Alter die Gruppen deutlich diversifiziert. Vor allem bei den Arbeitern und Reise- und Dienstleistungsangestellten gibt es in dieser Hinsicht deutliche Unterschiede.

Es kann erwartet werden, dass der PC für nahezu alle gesellschaftlichen Gruppen in ein bis zwei Jahrzehnten zum Alltagsmedium wird. Der Beruf oder die Bildung wird dann keine oder nur noch eine marginale Rolle spielen - mit Ausnahme der nie Erwerbstätigen, die in allen Segmenten einen sehr geringen Nutzeranteil aufweisen. Ob es

die befürchtete digitale Spaltung auf breiter gesellschaftlicher Ebene geben wird, bleibt somit abzuwarten - Bildungszertifikate und Berufe scheinen jedoch die Voraussetzungen für die digitale Welt zu sein. Ohne Beruf bzw. Bildungsabschluss scheint es schwierig zu werden, mitzuhalten. Diese Gruppen könnten hierbei ausgeschlossen werden.

Vor allem die Möglichkeiten des Internets und die zunehmende Konzentration als Vertriebsweg oder Informationsquelle (z. B. Vorbestellung von Karten für die Fußballweltmeisterschaft, Informationssuche, Preisvergleiche, Stellensuche) stellt eine große Gefahr für die ausgeschlossenen Gruppen dar. Viele Firmen schalten ihre Stellenanzeigen nicht mehr in der lokalen Tageszeitung, da es ihnen zu teuer ist und setzen sie lieber ins Internet - auf ihre Homepage und/oder in Jobsuchmaschinen. Für ältere Menschen, die heute weder einen Beruf ausüben noch vielleicht zur Fußballweltmeisterschaft möchten, ist der PC und auch das Internet somit entbehrlich - aber nicht für Jobsuchende, die vielleicht auch schon älter sind und somit auch eine „Problemgruppe“ für die Vermittlung darstellen

Dieser Teil war eine Grobsegmentierung nach dominanten Variablen. Im nachfolgenden Kapitel werden die Ergebnisse der Entscheidungsbäume mit denen der Diskriminanz- und der logistischen Regression verglichen. Danach werden weitere Differenzierungen nach Kultur- und Freizeitvariablen anhand der segmentierten Gruppen vorgenommen.

3.2 Ergebnisse der Logistischen Regression und der Diskriminanzanalyse

3.2.1 Ergebnisse der logistischen Regression

Der Vergleich mit den Entscheidungsbäumen und den anderen Verfahren erfolgt über die Fehlklassifikationsmatrix, die hier ein Ergebnis

von 22.7 liefert und somit geringfügig besser ist als das Ergebnis der Entscheidungsbäume:

ABBILDUNG 122 Fehlklassifikationsergebnis der Logistischen Regression (N = 1152)

Klassifikation

	Vorhergesagt		
Beobachtet	Non User	User	Prozent richtig
Non User	267	138	65,9%
User	123	624	83,5%
Prozent insgesamt	33,9%	66,1%	77,3%

Allerdings gehen nur 1152 Fälle in die Analyse ein - ein Umstand, der den fehlenden Angaben beim Haushaltsnettoeinkommen zuzurechnen ist. Leider ist es sowohl bei der logistischen Regression als auch bei der Diskriminanzanalyse nicht - wie bei einigen Entscheidungsbaumalgorithmen - möglich, mit Ersatzprädiktoren zu arbeiten.

Die Modellanpassung ist hoch signifikant und liefert einen hohen Chi-Quadrat-Wert, was auf die Bedeutsamkeit der unabhängigen Variablen hindeutet, die einen hohen, signifikanten Einfluß auf die PC-Nutzung bestzen:

ABBILDUNG 123 Logistische Regression: Modellanpassung (N = 1152)

Informationen zur Modellanpassung

Modell	-2 Log-Likelihood	Chi-Quadrat	Freiheitsgrade	Signifikanz
Nur konstanter Term	1360,351			
Endgültig	875,661	484,690	63	,000

Die Pseudo R-Quadrat-Statistiken ergeben:

ABBILDUNG 124 Pseudo R-Quadrat-Statistiken (N = 1152)

Pseudo-R-Quadrat	
Cox und Snell	,343
Nagelkerke	,473
McFadden	,324

Die Ergebnisse liegen zwischen 0.32 und 0.47 - es kann also von einer akzeptablen bis recht guten Modellanpassung gesprochen werden (bis zu einem Modellfit von 0.20 würde man kein gutes Modell unterstellen, ab 0.2 kann man von einem ganz akzeptablen Modell ausgehen, ab 0.4 von einem recht guten).

Die Güte der Anpassung fällt hinsichtlich des Signifikanzwertes sehr gut aus: die Signifikanz liegt deutlich unter 0.001.

ABBILDUNG 125 Logistische Regression: Likelihood-Quotienten-Tests (N = 1152)

Likelihood-Quotienten-Tests

Effekt	-2 Log-Likelihood für reduziertes Modell	Chi-Quadrat	Freiheitsgrade	Signifikanz
Konstanter Term	875,661[a]	,000	0	.
hhnetto	932,671[b]	57,010	11	,000
alter	989,842	114,181	42	,000
beruf	1151,581[b]	275,920	10	,000

Die Chi-Quadrat-Statistik stellt die Differenz der -2 Log-Likelihoods zwischen dem endgültigen Modell und einem reduziertem Modell dar. Das reduzierte Modell wird berechnet, indem ein Effekt aus dem endgültigen Modell weggelassen wird. Hierbei liegt die Nullhypothese zugrunde, nach der alle Parameter dieses Effekts 0 betragen.

a. Dieses reduzierte Modell ist zum endgültigen Modell äquivalent, da das Weglassen des Effekts die Anzahl der Freiheitsgrade nicht erhöht.

b. Bei den Daten liegt möglicherweise eine quasi-vollständige Trennung vor. Entweder existieren die Maximum-Likelihood-Schätzungen nicht, oder einige Parameterschätzungen sind unendlich.

Die Chi-Quadrat-Werte verdeutlichen die Wichtigkeit der unabhängigen Variablen - bezogen auf die PC-Nutzung. Auch hier wirkt der Beruf deutlicher vor dem Alter und dem Haushaltsnettoeinkommen - ähnlich wie bei den Entscheidungsbäumen.

Die Parameterschätzer werden schrittweise vorgestellt, da die Tabelle zu unübersichtlich ist, um sie sinnvoll darzustellen. Auf die Spalten Konfidenzintervalle und Freiheitsgrade wird verzichtet, da sie für die Fragestellung nicht bedeutsam sind - die wichtigsten Ergebnisse werden nach den unabhängigen Variablen zusammengefaßt:

ABBILDUNG 126 Logistische Regression: Parameterschätzer des Haushaltsnettoeinkommens (N = 1152)

	B	Standardfehler	Wald	Signifikanz	Exp(B)
Konstanter Term	-2,92230942	0,989168198	8,727947074	0,003134	
[hhnetto=1,00]	2,04501534	0,408236112	25,09402364	0,000001	7,72927714
[hhnetto=2,00]	1,87665165	0,529366493	12,56766027	0,000392	6,53159811
[hhnetto=3,00]	1,36951939	0,425914751	10,33929687	0,001302	3,93345977
[hhnetto=4,00]	1,00777458	0,460150893	4,796519968	0,028517	2,73949768
[hhnetto=5,00]	1,64820655	0,388570476	17,99217533	0,000022	5,19764975
[hhnetto=6,00]	1,54076816	0,381906333	16,27649434	0,000055	4,66817481
[hhnetto=7,00]	1,21686091	0,363209583	11,22450717	0,000807	3,37657171
[hhnetto=8,00]	1,10494538	0,335530948	10,84467272	0,000991	3,01905955
[hhnetto=9,00]	0,67134368	0,313026467	4,599680803	0,031978	1,95686496
[hhnetto=10,00]	0,74743902	0,335209198	4,971869027	0,025763	2,11158536
[hhnetto=11,00]	0,1452091	0,329587094	0,194109702	0,659518	1,15628132
[hhnetto=12,00]	0	.	.	.	.

Die Einkommensklassen sind von der niedrigsten (hhnetto = 1) bis zur höchsten mit den Signifikanzen dargestellt. Als Referenzkategorie dient die höchste Einkommensklasse. Die grau schraffierten Zellen zeigen die interessantesten Ergebnisse.

An den Spalten B bzw. Exp(B)[88] sind die (un)standardisierten Regressionskoeffizienten für das Haushaltsnettoeinkommen ersichtlich.

Der Wert für B bzw. EXP(B) sinkt kontinuierlich von der Gruppe 1 zur Gruppe 4 bzw. von der Gruppe 5 bis zur Gruppe 9. Gruppe 10 ist vom Wert leicht erhöht gegenüber Gruppe 9, Einkommensgruppe 11 liefert den niedrigsten Regressionskoeffizienten. Genau an diesen Grenzen gibt es keine signifikanten Ergebnisse.

Die Wald-Statistik - in der Tabelle ist der höchste und der geringste Wert gekennzeichnet - zeigt ein ähnliches Bild: geringere Einkommensgruppen liefern höhere Chi-Quadrat-Werte, und umgekehrt.

Ähnliches gilt auch für den Beruf:

88. Bei den EXP(B)-Werten handelt es sich um einheitlich positive Werte, die besser vergleichbar sind. So ist die Wahrscheinlichkeit der PC-Nichtnutzung bei der Einkommensgruppe 11 (EXP(B) = 1.16) siebenmal geringer als bei Gruppe 1 (EXP(B) = 7.72)

ABBILDUNG 127 Logistische Regression: Parameterschätzer des Berufs (N = 1152)

	B	Standardfehler	Wald	Signifikanz	Exp(B)
nie erwerbstätig	4,80780797	0,85497716	31,6216989	0,000000	122,462881
sonst. (Fach)Arbeiter	3,81481403	0,81981081	21,6530689	0,000003	45,3683183
Sonst. Ang. Reise u. D.	3,2708237	0,83621407	15,2995736	0,000092	26,3330209
Meister	2,45121325	1,00427426	5,95741054	0,014656	11,6024148
Ladenbesitzer, Handw.	2,34893737	0,98478345	5,68933354	0,017068	10,4744333
Landw., Fischer	2,38156391	1,12496657	4,48172544	0,034259	10,821814
sonst. Bürotätigk.	1,19328143	0,86871801	1,88680975	0,169562	3,29788526
Bürotätigk. mit Leitung	1,11192404	0,87275512	1,62317531	0,202650	3,04020225
Freie Berufe	1,38034998	1,12832144	1,49662413	0,221192	3,976293
Grossunternehmer	1,33425067	0,96917979	1,89524848	0,168611	3,79714956
Studierende	0	.	.	.	.

Die Parameterschätzung sind in den letzten vier Kategorien (vor der Referenzkategorie) nicht signifikant. Da alle B-Koeffizienten positiv sind, sind auch die EXP(B)-Koeffizienten > 1 (ansonsten wären sie kleiner 1). Eine weitere Erläuterung bei fast der Hälfte der nichtsignifikanten Kategorien scheint problematisch,[89]

Auf die Darstellung der 79 Alterskategorien soll hier bewußt verzichtet werden - die standardisierten Koeffizienten liegen in nahezu allen Jahrgängen unter 1. Auch hier sind die Ergebnisse häufig nicht signifikant.

Nach der Formel

```
PC-Nutzung = Konstante + hhnettoeinkommen + alter + beruf
```

ergibt sich für jede (unstandardisierte) Konstellation ein „PC-Nutzungswert“, zum Beispiel für 18jährige Angestellte eines Reise- und Dienstleistungsberufs mit einem Einkommen der Gruppe 4:

```
PC-Nutzung = -2.922 konstante -3.85 alter + 1.01 hhnetto + 3.27 beruf (Gruppe 3) = -2.49
```

89. Auch hier der Verweis auf die Entscheidungsbäume: die Höhe des maximalen Signifikanzniveaus ist voreinstellbar (in den Beispielen der Arbeit wurde mit einem Signifikanzniveau von 0.00 gearbeitet) - somit werden nur Ergebnisse errechnet, die unter diesem Signifikanzniveau bleiben und müssen nicht im Nachhinein zusätzlich interpretiert werden.

Für 18jährige Büroangestellte (Gruppe 7) mit einem Haushaltsnettoeinkommen der Gruppe 6 sieht die Gleichung folgendermaßen aus:

```
PC-Nutzung = -2.922 konstante -3.85 alter + 1.54 hhnetto + 1.19 be-
                        ruf (Gruppe 7) = -4.04
```

Da die Werte jedoch unstandardisiert sind (positive und negative Effekte können sich hier „aufheben“ und das kann Gruppen verzerren), werden die Werte aus der Spalte Exp(B) verwendet - diese Werte sind einheitlich positiv.

An dieser Stelle ergibt sich jedoch das Problem, dass der Wert für die Konstante nicht signifikant ist - und auch kein EXP(B)-Wert geliefert wird. Somit ist die Aussagekraft des Modells stark eingeschränkt bzw. nicht verwendbar.

Gerade bei den Parameterschätzern wird die Komplexität der logistischen Regression deutlich: die Werte sind nicht einfach zu interpretieren bzw. wenn einer der vielen Werte falsch interpretiert wird, kann dadurch die Aussagekraft deutlich leiden. In der Regel sind auch nicht alle gefundenen Konstellationen signifikant oder liefern besonders herausgehobene Kennwerte. Für viele Sozialwissenschaftler ist es auch ein Problem, Menschen in Form von Funktionswerten zu vermessen und zu standardisieren. Selbst standardisierte Koeffizienten sind schwierig hinsichtlich Nähe und Distanz zu interpretieren und nur für diese eine Analyse maßgeblich.

Das Ergebnis ist den Entscheidungsbäumen ähnlich, liefert aber etwas andere Schwerpunkte: die Fehlklassifikation ist etwas besser, dafür ist der Informationsgehalt deutlich geringer: es können nur Aussagen über die Wichtigkeit der Einflußvariablen vorgenommen werden, nicht jedoch über die Verteilungen innerhalb unterschiedlicher Segmente. Es ist zwar möglich, die Variablen nach Gruppen in SPSS zu speichern, aber eine differenziertere Analyse, die alle drei Va-

riablen berücksichtigt (z. B. 18 - 25jährige Arbeiter mit einem Haushaltsnettoeinkommen bis zu 2500 DM).

Ein deutlicher Unterschied zwischen Entscheidungsbäumen und den anderen beiden Verfahren liegt darin, dass die errechneten Bäume immer signifikant sind und sich somit die Frage, ob es ein signifikantes Modell ist oder nicht überhaupt nicht stellt. Das gefundene Ergebnis muss nicht auf diffizile Koeffizienten überprüft werden und kann „as is" interpretiert werden. Sind Ergebnisse nicht signifikant, werden sie von Answertree auch nicht ausgegeben.

3.2.2 Ergebnisse der Diskriminanzanalyse

Auch hier gehen von den 1413 Fällen nur 1152 in die Analyse ein, da mindestens eine Diskriminanzvariable (Haushaltsnettoeinkommen) fehlt.

Eine Untersuchung der Variablen wie bei der logistischen Regression und den Entscheidungsbäumen ist nicht so ohne weiteres möglich, da die Voraussetzungen für die Diskriminanzanalyse ordinales bzw. metrisches Skalenniveau bei den unabhängigen Variablen voraussetzt. Nur dichotome nominale Variablen können untersucht werden, nicht jedoch polytome.[90]

Daraus folgt, dass die Variablen Beruf und Haushaltsnettoeinkommen entweder nicht in die Analyse einbezogen werden - oder dass ein Weg gefunden wird, sie zu dichotomisieren. Hierbei gibt es - wie immer - zwei Möglichkeiten: entweder eine theoretische oder eine empirische Lösung.

90. Dem liegt die Überlegung zugrunde, eine nominale Variable wie eine metrische zu behandeln: es werden nicht zwei Ausprägungen (0 und 1) unterstellt, sondern eine Art „metrisches Kontinuum" zwischen diesen Werten.

Eine theoretische Lösung ware zum Beispiel, die Büroberufe einerseits, die restlichen Berufe andererseits zu untersuchen - mit der Annahme, dass Büroberufe in besonderem Maße mit der Bedienung des PCs in Verbindung stehen. Andererseits haben die binären Baumalgorithmen empirisch gezeigt, dass es auch andere Berufsgruppen gibt, die den PC vordringlich nutzen - zum Beispiel Studierende oder Freie Berufe.

Ein weiteres Vorgehen wäre, zwei Berufssegmente ab einem bestimmten PC-Nutzeranteil zu splitten - was allerdings methodisch und theoretisch schwierig begründbar ist.

Deshalb scheint eine empirische Lösung - wie sie durch Entscheidungsbäume gefunden wird - eine bessere Lösung - natürlich muss sie auch theoretisch fundiert sein.

Für diese Trennung kommen zwei Baumalgorithmen in Frage: CART und QUEST. Diese liefern folgende Ergebnisse:

TABELLE 28 BINÄRE BERUFSSEGMENTIERUNG MIT CART UND QUEST (N = 1413)

	CART	QUEST
gefundene Gruppen Nichtnutzer	• nie erwerbstätig • sonstige Angest. R & D • sonstige (Fach)Arbeiter • Ladenbesitzer, Handwerker	• -sonstige (Fach-)Arbeiter • nie erwerbstätig
gefundene Gruppen Nutzer	• Grossunternehmer, Direktoren, Top Management • - leitende Angestellte, leitende Bürotätigkeiten • sonstige Bürotätigkeiten • Studierende • Freie Berufe • Meister • Landwirte, Fischer	• sonstige Angest. R & D • Ladenbesitzer, Handwerker • Grossunternehmer, Direktoren, Top Management • - leitende Angestellte, leitende Bürotätigkeiten • sonstige Bürotätigkeiten • Studierende • Freie Berufe • Meister • Landwirte, Fischer
Verhältnis Nutzer - Nichtnutzer	90 : 44 10 : 56	81 : 39 19 : 61
Fehlklassifikation	0.27	0.27
Verbesserung	0.1035	0.0637

Mathematisch-statistisch liefert CART die bessere Lösung: die Verbesserung ist mit 0.1035 deutlich höher als die von QUEST (0.0637). Die Fehlklassifikation ist gleich hoch und kann somit nicht zur Bewertung herangezogen werden.

Inhaltlich sieht man auch in diesem Beispiel einen gewissen „Rigorismus“ QUESTS: die Neigung, deutlich niedrig besetzte Gruppen herauszusegmentieren. Ein Blick in die Ursprungstabelle macht aber die Trennungen beider Algorithmen deutlich:

ABBILDUNG 128 (Einstufige) Berufssegmentierung mit CART und QUEST (N = 1413, Zeilen-%, schraffiert: Gruppen mit geringen Nutzeranteilen)

D15AR: Beruf * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

			Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)		
			Non User	User	Gesamt
D15AR: Beruf	nie erwerbstätig	Anzahl	86	32	118
		% von D15AR: Beruf	72,9%	27,1%	100,0%
	sonstige (Fach)Arbeiter	Anzahl	259	186	445
		% von D15AR: Beruf	58,2%	41,8%	100,0%
	sonst. Angest. Reise+Dienst.	Anzahl	86	115	201
		% von D15AR: Beruf	42,8%	57,2%	100,0%
	Meister	Anzahl	5	19	24
		% von D15AR: Beruf	20,8%	79,2%	100,0%
	Ladenbesitzer, Handwerker	Anzahl	11	17	28
		% von D15AR: Beruf	39,3%	60,7%	100,0%
	Landwirte, Fischer	Anzahl	3	11	14
		% von D15AR: Beruf	21,4%	78,6%	100,0%
	sonstige Bürotätigkeiten	Anzahl	18	149	167
		% von D15AR: Beruf	10,8%	89,2%	100,0%
	Büroangest. mit Leitungsfunktion	Anzahl	19	168	187
		% von D15AR: Beruf	10,2%	89,8%	100,0%
	Freie Berufe	Anzahl	3	35	38
		% von D15AR: Beruf	7,9%	92,1%	100,0%
	Grossunternehmermer, Top Management, Ltd. Angest.	Anzahl	7	56	63
		% von D15AR: Beruf	11,1%	88,9%	100,0%
	Student(in)	Anzahl	7	121	128
		% von D15AR: Beruf	5,5%	94,5%	100,0%
Gesamt		Anzahl	504	909	1413
		% von D15AR: Beruf	35,7%	64,3%	100,0%

Die prozentualen „Abstände“ der Gruppen zeigen die Differenz der PC-Nutzung zwischen den einzelnen Gruppen und somit die statistisch-logische Begründung. Betrachtet man die Nutzeranteile der ersten vier Variablenausprägungen des Berufs, ergibt sich (gerundet):

TABELLE 29 PROZENTSATZDIFFERENZEN AUSGEWÄHLTER NUTZERANTEILE NACH BERUF (N = 1413)

Kategorie 1	Kategorie 2	Differenz
nie erwerbstätig: 27 %	Arbeiter: 42 %	15 %
Arbeiter: 42 %	R&D-Angestellte: 57 %	15 %
Ladenbes./Handw. 61 %	Landw./Fischer: 79 %	18 %

Auf den ersten Blick sinnvoller wäre es, die Trennung des CART-Algorithmus zu verwenden - zwei Gründe sind dafür ausschlaggebend. Zum einen liegt der Verbesserungswert deutlich höher als bei QUEST, zum anderen ist die Trennung theoretisch besser begründbar, denn die segmentierten vier Berufsgruppen stehen nicht in einem unmittelbaren beruflichen Bezug zum PC - im Gegensatz z. B. zu Bürotätigkeiten.

Allerdings wäre bei der Auswahl eines Algorithmus die Frage offen, ob die Fehlklassifikation bei der Diskriminananalyse gleich hoch oder unterschiedlich hoch ausfällt. Dies läßt sich sehr leicht überprüfen:

ABBILDUNG 129 Diskriminanzanalyse: PC-Nutzung nach Alter, Haushaltsnettoeinkommen, Beruf (dichotomisiert, QUEST, N = 1152)

Klassifizierungsergebnisse[a]

		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Vorhergesagte Gruppenzugehörigkeit		
			Non User	User	Gesamt
Original	Anzahl	Non User	296	109	405
		User	183	564	747
		Ungruppierte Fälle	5	2	7
	%	Non User	73,1	26,9	100,0
		User	24,5	75,5	100,0
		Ungruppierte Fälle	71,4	28,6	100,0

a. 74,7% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Knapp 3/4 der Fälle wird hier richtig klassifiziert.

ABBILDUNG 130 Diskriminanzanalyse: PC-Nutzung nach Alter, Haushaltsnettoeinkommen, Beruf (dichotomisiert, CART, N = 1152)

Klassifizierungsergebnisse[a]

		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)	Vorhergesagte Gruppenzugehörigkeit		
			Non User	User	Gesamt
Original	Anzahl	Non User	337	68	405
		User	240	507	747
		Ungruppierte Fälle	7	0	7
	%	Non User	83,2	16,8	100,0
		User	32,1	67,9	100,0
		Ungruppierte Fälle	100,0	,0	100,0

a. 73,3% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

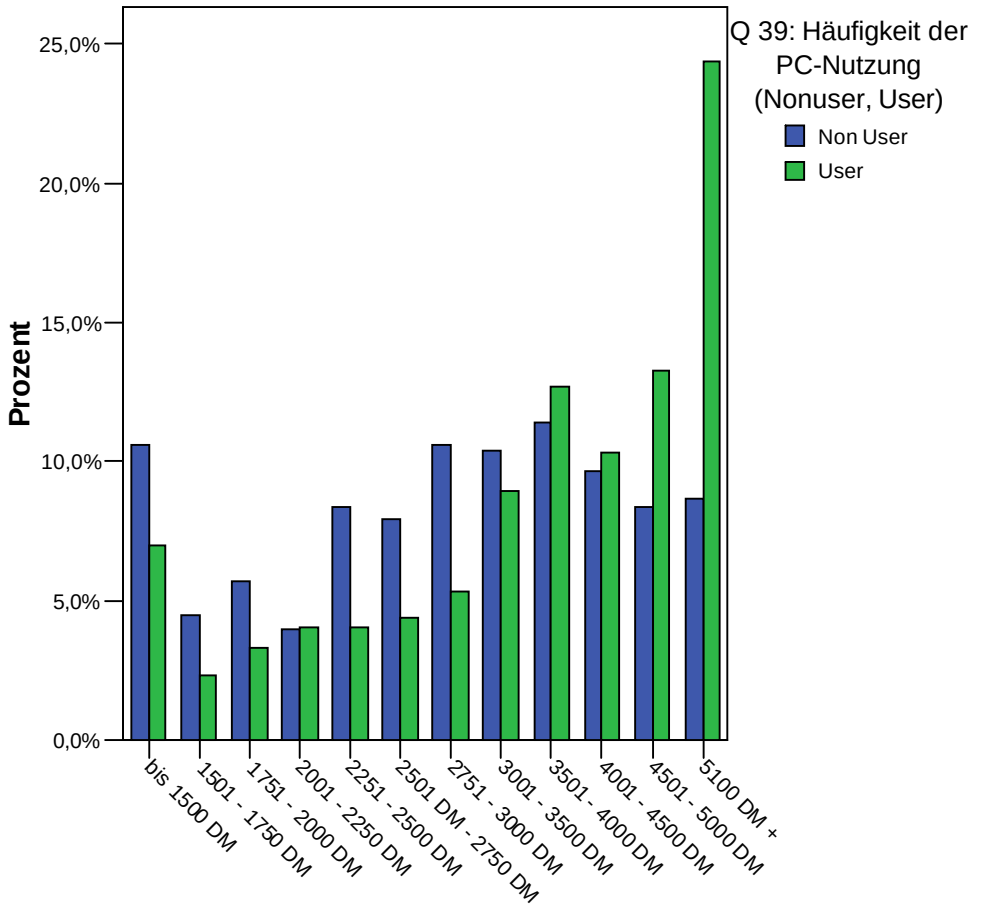
CARTs Fehlklassifikation liegt um 1.4 % schlechter als die von QUEST - ein Hinweis darauf, dass die Verwendung unterschiedlicher multivariater Verfahren doch recht komplex und nicht so ohne weiteres „durchschaubar“ ist. Eine bessere Fehlklassifikation von 1.4 % fällt nicht ins Gewicht - aber es zeigt, wie sich die unterschiedlichen unabhängigen Variablen doch beeinflussen können.

Wohlgemerkt - QUEST klassifiziert hier nicht „besser“ als CART, da beide Algorithmen nur den Beruf, nicht aber Alter und Haushaltsnettoeinkommen bei der Segmentierung berücksichtigten und das Ergebnis ist vernachlässigbar. Trotzdem zeigt es, dass es durchaus Sinn macht, statistischen Zahlen nicht so ohne weiteres Vertrauen zu schenken. Insgesamt ist die Fehlklassifikation mit den beiden anderen Verfahren vergleichbar: rund 3/4 aller Fälle wurde richtig klassifiziert - ein Ergebnis, dass die Entscheidungsbäume auch hier als gleichwertiges Verfahren rechtfertigt.

An dieser Stelle muss allerdings ein Bias eingeräumt werden: das Problem, dass die nachfolgend untersuchte ordinalskalierte Variable Haushaltsnettoeinkommen als metrisch unterstellt wurde. Somit bieten sich zwei Möglichkeiten an: entweder kann intervallskaliertes Niveau unterstellt werden - oder die Einkommensklassen werden zu zwei Gruppen dichotomisiert.

Um diese Frage zu klären, ist es sinnvoll, sich die Verteilung der Einkommensklassen über die PC-Nutzung zu vergegenwärtigen:

ABBILDUNG 131 Grafische Darstellung: PC-Nutzung nach Haushaltsnettoeinkommen (in %, N = 1413)

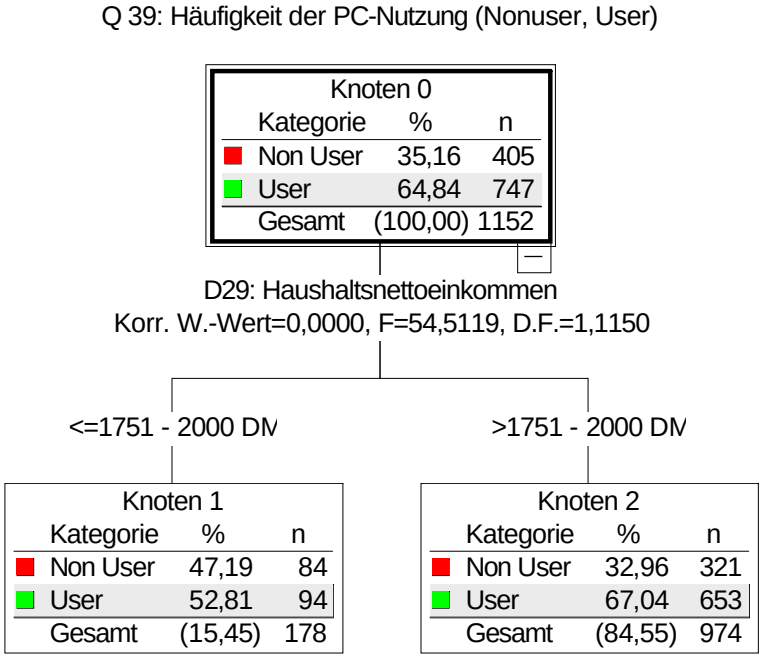


D29: Haushaltsnettoeinkommen

Erst in den höheren Altersgruppen liegt der Anteil der Nutzer deutlich über dem der Nichtnutzer (ab 3500 DM). Vorher sind die Anteile der Nichtnutzer zumeist deutlich überrepräsentiert (mit Ausnahme der Gruppe 2000 - 2250 DM). Somit kann kein monotoner Verlauf unterstellt werden.

Andererseits stellt sich die Frage, wo die Dichotomisierung erfolgen soll. Hier liefern CART und QUEST zwei völlig unterschiedliche Lösungen:

ABBILDUNG 132 BINÄRE HAUSHALTSNETTOEINKOMMENSSEGMENTIERUNG MIT QUEST (N = 1152)

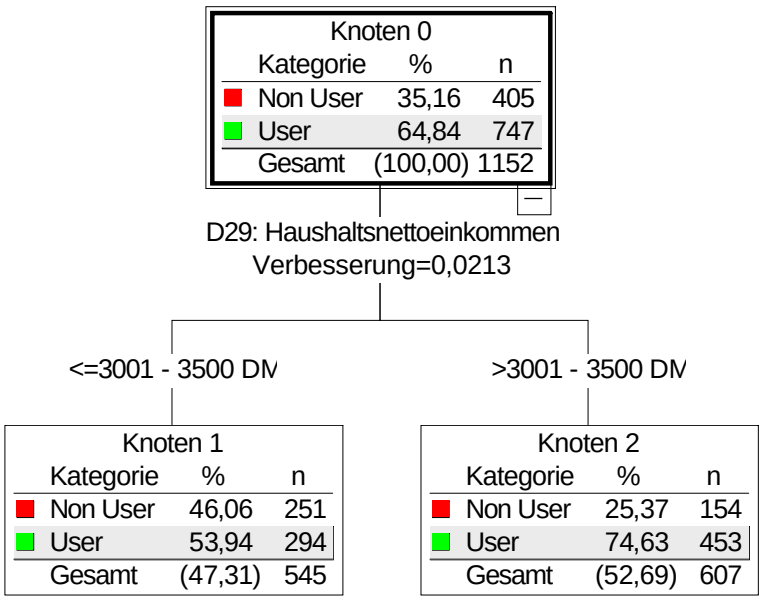


Die Lösung ist nicht sehr glücklich: die Trennung erfolgt bei 2000 DM - allerdings ist das Verhältnis der Knoten (Gesamt) mit 15 : 85 nicht sehr ausgewogen - auch hier wieder der Hinweis darauf, dass QUEST wohl versucht, kleinere Gruppen „herauszusegmentieren“ Das Verhältnis der Nutzer ist mit 52 : 67 nicht überragend.

CART segmentiert nahezu zwei anteilsmäßig große Unterknoten (Gesamt: 47 % vs. 53 %). Die Trennung erfolgt bei 3500 DM - genau an dem Punkt, wo auch die Anteile der Nutzer höher sind als die Nichtnutzer. Allerdings fällt die Verbesserung mit 0.02 recht gering aus. Die Fehlklassifikation beträgt in beiden Fällen 0.35 für den Gesamtbaum.

ABBILDUNG 133 BINÄRE HAUSHALTSNETTOEINKOMMENSSEGMENTIERUNG MIT CART (N = 1152)

Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)



Hier kann keine Entscheidung über ein „besseres“ oder „schlechteres“ Modell zur Dichotomisierung getroffen werden.[91] Da sich auch bei der Berufssegmentierung gezeigt hat, dass die isolierte Betrachtung von Variablen keine Entscheidung liefert, sollen auch hier wieder beide Fälle anhand der Diskriminanzanalyse durchgeführt werden. Anhand der Fehlklassifikation soll weiter entschieden werden, welche Berufssegmentierung für die Darstellung und Beschreibung des Diskriminanzmodells herangezogen wird.

Um zu prüfen, ob es zu grossen Abweichungen zwischen den Fehlklassifikationen der unterschiedlichen Diskriminanzmodelle kommt, werden alle möglichen Kombinationen in nachfolgender Tabelle dargestellt:

91. Natürlich wäre es möglich, dem CART-Vorschlag zu folgen, da an dem gefundenen Splitpunkt auch die Anteile der Nutzer höher sind als die der Nichtnutzer.

TABELLE 30 DISKRIMINANZANALYSE: VERGLEICH DER FEHLKLASSIFIKATIONSERGEBNISSE ZWISCHEN DEN DICHOTOMISIERTEN VARIABLEN HAUSHALTSNETTOEINKOMMEN UND BERUF

	Haushaltsnettoeinkommen CART (+/- 3500 DM)	Haushaltsnettoeinkommen QUEST (+/- 2000 DM
Berufssegmente CART (nie erwerbstätig, sonstige Facharbeiter, sonst. Angestellte Reise & Dienstleistung, Ladenbesitzer, Handwerker)	26.0 %	27.3 %
Berufssegmente QUEST (nie erwerbstätig, sonstige Facharbeiter)	25.2 %	26.4 %

Im Vergleich dazu stehen die für das (als metrisch unterstellte) Haushaltsnettoeinkommen für die CART-Segmentierung bei 26.7 %, für QUEST bei 25.3 %. Die Differenz zwischen dem „schlechtesten" (27.3 %) und dem „besten" (25.2 %) Ergebnis ist mit einer Differenz von rund 2 % als marginal zu bezeichnen (bei 1152 Personen entsprechen 2 % rund 23 Befragten, die anstatt falsch richtig klassifiziert wurden). Damit zeigt sich, dass für diesen einen Fall die Wahl der metrischen bzw. dichotomen) Haushaltsnettoeinkommensalternative bzw. Berufssegmentierungen kaum Unterschiede hinsichtlich der Fehlklassifikation macht.

Die inhaltliche (und Verteilungs-)Begründung für das Einkommen gibt der Trennung bei 3500 DM den Vorzug - die Nutzeranteile liegen in den besseren Einkommensklassen höher. Die Berufssegmentierung ist schwierig zu treffen: inhaltlich wäre die CART-Lösung mit vier Berufsgruppen mit überproportionalem Nichtnutzeranteil dem der QUEST-Lösung mit zwei stark unterrepräsentierten Nonuser-Berufsgruppen vorzuziehen - statistisch liefert für die Diskriminanzanalyse für dieses Segment aber die bessere Lösung.

Im Normalfall würde man der inhaltlichen Lösung den Vorzug geben - in dieser Arbeit geht es jedoch hauptsächlich um die Leistung und den Vergleich der verschiedenen Algorithmen und Verfahren. Deshalb soll im nachfolgenden die Lösung mit der geringsten Fehlklassifikation (für die Einkommenssegmentierung CART, QUEST für die Berufssegmentierung) herangezogen werden. Die Lösung ist insoweit inhaltlich vertretbar, da die Klassifizierungen nicht stark voneinander abweichen - es wäre allerdings für eine inhaltliche Betrachtung keine ausreichende Begründung.

ABBILDUNG 134 Diskriminanzanalyse: Schrittweises Vorgehen bei der Prüfung der unabhängigen Variablen Alter, Beruf und Haushaltsnettoeinkommen (N = 1152)

Schritt	Variablen	Lambda	Signifikanz
1	Beruf	0,81240342	0,00000
2	Alter	0,76782907	0,00000
3	HH-Nettoeinkommen	0,73954071	0,00000

In jedem Schritt wird eine Variable in die Analyse aufgenommen - in der Reihenfolge Beruf, Alter und Haushaltsnettoeinkommen. Die Ergebnisse sind höchstsignifikant.

ABBILDUNG 135 Diskriminanzanalyse: Eigenwerte der unabhängigen Variablen Alter, Beruf und Haushaltsnettoeinkommen (N = 1152

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,352[a]	100,0	100,0	,510

a. Die ersten 1 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

Die Eigenwerte geben die Güte des Modells an - je höher die Werte, desto besser die Erklärungskraft. Wenn sich beispielsweise bei zwei

Funktionen Eigenwerte von 2.0 und 0.5 ergeben, dann würde die erste Funktion 80 %, die zweite Funktion 20 % erklären.

ABBILDUNG 136 Diskriminanzanalyse: F und Signifikanz (N = 1152)

Wilks-Lambda				
Schritt	Anzahl der Variablen	Lambda	Exaktes F	Signifikanz
1	Beruf	0,81240342	265,552874	0,00000
2	Alter	0,76782907	173,713396	0,00000
3	HH-Nettoeinkommen	0,73954071	134,771602	0,00000

Die Variablen sind alle hochsignifikant, lassen sich also auf die Grundgesamtheit übertragen. Sie sind hier gut geeignet, die gefundenen Gruppen zu trennen. In der Spalte F wird die Wichtigkeit der unabhängigen Variablen angegeben - ebenso wie bei logistischer Regression und Entscheidungsbäumen ist der Beruf wichtigstes Merkmal, gefolgt von Alter und Haushaltsnettoeinkommen. Es ergibt sich leider nur eine Diskriminanzfunktion:

ABBILDUNG 137 Diskriminanzanalyse: Eigenwerte und WILKs LAMBDA (N = 1152)

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,352[a]	100,0	100,0	,510

a. Die ersten 1 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

0

Wilks' Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1	,740	346,532	3	,000

Die kanonische Korrelation ist mit 0.51 recht gut und zeigt an, dass auch hier die Variablen deutlich die Gruppen hinsichtlich der PC-Nutzung erklären kann.

Bei einer multivariaten Analyse wird in der Regel mehr als eine Funktion herausgearbeitet (Zahl der unabhängigen Variablen - 1) - und ist wahrscheinlich deutlicher mit vielen unabhängigen Variablen erreichbar. Erschwert wird es in diesem Fall noch durch die fehlenden Werte beim Haushaltsnettoeinkommen - ein Problem, dass sich auch für die logistische Regression ergibt.

3.2.3 Zusammenfassung

Es zeigt sich, dass Regressionsverfahren und Diskriminanzanalyse stark hypothesengeleitete Verfahren sind - im Gegensatz zu den Entscheidungsbäumen, die sowohl hypothesengeleitet als auch eher explorativ eingesetzt werden können. Sie können nur das Endmodell von Entscheidungsbäumen validieren, da sie viele unabhängige Variablen benötigen, um verschiedene (Diskriminanz-)Funktionen zu finden.

Interessanterweise unterscheiden sich die Klassifikationsergebnisse auch hier nur wenig, so dass kein Verfahren mathematisch-statistisch vorteilhafter oder schlechter gegenüber einem anderen wäre - verschiedene Logiken kommen somit zu einem recht ähnlichen Output.

Allerdings ist der Informationsgehalt der Entscheidungsbäume auch hier deutlich höher, da Gruppen nicht aufgrund von einer Funktion, sondern von der Überprüfung der unabhängigen Variablen auf jeder Stufe der Analyse gebildet werden. Anders ausgedrückt: Entscheidungsbäume führen stets zu einem differenzierten Ergebnis, wenn die Größe der Stichprobe und die unabhängigen Variablen eine gewisse Korrelation mit der Zielvariablen aufweisen. Logistische Regression und Diskriminanzanalyse stellen weitaus differenziertere Anforderungen an den Datensatz: es dürfen keine fehlenden Werte vorliegen, und die Größe der Stichprobe muß umfangreich sein.

3.3 Subordinierte Schichtungen der Entscheidungsbäume

Die subordinierten Schichtungen betreffen die Kultur- und Freizeitvariablen. In einem ersten Schritt werden die Ergebnisse aus der Gesamtstichprobe denen der Gruppe der bis 58jährigen gegenübergestellt. Der Vorteil liegt darin, altersspezifische Aktivitäten zu identifizieren. Liegt z. B. der Wert für eine Freizeitaktivität in der Gesamtstichprobe eher hoch, in der Gruppe der bis 58jährigen jedoch eher niedrig, ist das Alter und nicht die PC-Nutzung dafür verantwortlich.

TABELLE 31 WICHTIGE BIVARIATE ZUSAMMENHÄNGE (> 0.10) ZWISCHEN PC-NUTZUNG UND DEN KULTUR- UND FREIZEITVARIABLEN (PHI, CRAMERS V, UNSICHERHEITSKOEFFIZIENT (SIG = 0.000))[A]

unabhängige Variable	Skalenniveau	Kennzahlen Gesamt	Kennzahlen bis 58 Jahre
Art.v. Fernsehsendungen Soaps/Serien	nom.-dichotom	alle < 0.1	phi = -0.112 u = 0.100 phi = 0.100 u = 0.100
Häufigkeit DVD/Videos sehen	ordinal	v = 0.406	v = 0.223
Häufigkeit Radiohören	ordinal	v = 0.117	v = 0.136
Gelesene Bücher i. d. letzten 12 Monaten beruflich (v, u) Bildg (Pfl.) (v, u) Bildg (k.Pfl.)()v, u)	nom.-(MF)	 v = 0.333, u = 0.008 v = 0.329, u = 0.086) v = 0.243, u = 0.043	 v = 0.263, u = 0.062 v = 0.265, u = 0.063 v = 0.220. u = 0.041
Häufigkeit Musik hören	ordinal	v = 0.132	v = 0.126
Art v. Musik Oper(ette) Klassik Rock - Pop Hardrock Dance/House Techno Rap Jazz - Blues Volksmusik	nom.-dichotom	 ph = -0.128, u = 0.012) ph = 0.032, u = 0.00) ph = 0.346, u = 0.088 ph = 0.190, u = 0.027 ph = 0.205, u = 0.031 ph = 0.220, u = 0.037 ph = 0.194, , u = 0.029 ph = 0.137, u = 0.014 ph = -0.421, u = 0.133	 ph = - 0.11, u = .000[b] ph = -0.121, u = 0.01 ph = -0.121, u = 0.01 ph = -0.106, u = 0.00 ph = -0.093, u = 0.00 ph = -0.134, u = 0.01 ph = -0.118, u = 0.01 ph = -0.079, u = 0.00 ph = --0.252, u = 0.04
Kulturelle Aktivitäten (v) Kino Theater Sportveranst. Konzert Bibliothek Sehenswürdigk Museen Ausgrabungen	ordinal	 v = 0.488 v = 0.203 v = 0.261 v = 0.237 v = 0.278 v = 0.209 v = 0.158 v = 0.113	 v = 0.331 v = 0.206 v = 0.207 v = 0.205 v = 0.229 v = 0.236 v = 0.183 v = 0.121
Konzertbesuch (Phi, u) Rock - Pop Hardrock Jazz - Blues Volksmusik	nom.-dichotom	 ph = 0.248, u = 0.050) ph = 0.148, u = 0.021 ph = 0.106, u = 0.010 ph = -0.243, u = 0.043	 ph = 0.111, u = 0.012 ph = 0.106, u = 0.012 ph = 0.052, u = 0.003 ph = -0.207, u = 0.03

TABELLE 31 WICHTIGE BIVARIATE ZUSAMMENHÄNGE (> 0.10) ZWISCHEN PC-NUTZUNG UND DEN KULTUR- UND FREIZEITVARIABLEN (PHI, CRAMERS V, UNSICHERHEITSKOEFFIZIENT (SIG = 0.000))[A]

unabhängige Variable	Skalenniveau	Kennzahlen Gesamt	Kennzahlen bis 58 Jahre
Medienbesitz (Phi, u)	nom.-dichotom		
Videorecorder		ph = 0.369, u = 0.103	ph = 0.187, u = 0.034
Fotoapparat		ph = 0.194, u = 0.028	ph = 0.173, u = 0.03
Videokamera		ph = 0.257, u = 0.049	ph = 0.189, u = 0.030
Cassettenrec.		ph = 0.188, u = 0.026	ph = 0.170, u = 0.022
Stereoanlage		ph = 0.25, , u = 0.050	ph = 0.133, u = 0.013
Walk/Disc/mp3-Pl		ph = 0.386, u = 0.112	ph = 0.269, u = 0.058
Videospiele		ph = 0.232, u = 0.040	ph = 0.137, u = 0.020
Organizer/PDA		ph = 0.207, u = 0.039	ph = 0.176, u = 0.034
DVD-Player		ph = 0.256, u = 0.059	ph = 0.211, u = 0.047
Bücher		ph = 0.117, u = 0.010	ph = 0.154, u = 0.018
Lexika (Buch)		ph = 0.184, u = 0.025	ph = 0.203, u = 0.031
Lexika (CD-ROM)		ph = 0.379, u = 0.126	ph = 0.331, u = 0.116
Medien (CDs, ...)		ph = 0.205, u = 0.031	ph = 0.167, u = 0.021
Musikinstrument		ph = 0.202, u = 0.030	ph = 0.206, u = 0.037
Anzahl der Bücher (v)	ordinal	0.224	0.245

a. Variablen, bei denen ein Zusammenhangsmaß unsinnig ist (z. B. Vornamen) wurden nicht berechnet und nicht in die Tabelle aufgenommen.
b. nicht signifikant

Die größten Unterschiede ergeben sich zwischen den Musikstilen und dem (Medien-)Besitz eines Videorecorders (0.369 für die Gesamtstichprobe, 0.187 für die Jüngeren). Folglich ist auch die Häufigkeit des Video- bzw. DVD-Sehens (schraffierter Bereich) mit 0.406 (bis 58jährige: 0.223) relativ hoch Hier liegt eindeutig ein Alterseffekt zugrunde.

Von der hohen Korrelation von Phi = 0.346 in der Gesamtstichprobe für das Hören von Rock- bzw. Popmusik bleiben Phi = 0.121 in der Gruppe der bis 58jährigen übrig - ein Ergebnis, dass fast zu vernachlässigen ist und das auch die hohe Bedeutung der „alten" Ungleichheiten, bezogen auf Freizeit- und Kulturvariablen, hervorhebt (ebenso der Besuch von Pop- und Rockkonzerten: Gesamtstichprobe = 0.248, Jüngere = 0.111). Relativ stabil bleiben hingegen die Werte für Volksmusik (- 0.243 : - 0.207), was darauf hindeutet, dass eben

nicht nur Ältere diese Musikrichtung hören, sondern auch viele der mittleren Jahrgänge eine Vorliebe dafür besitzen. Dieser Musikstil ist jedoch in der Gruppe der bis 58jährigen, die einzige, die eine nennenswerte Korrelation aufweist.

Teilweise stark weichen die Werte der beruflich veranlaßten Weiterbildung ab. Sie liegen in der Gesamtstichprobe bei 0.333 (Jüngere: 0.263). Für die Pflichtweiterbildung liegen die Werte bei 0.329 (Gesamt) und 0.265 (bis 58jährige). Die freiwillige Weiterbildung unterscheidet sich kaum (0.243 : 0.220). Dies deutet darauf hin, dass Jüngere mehr in die Weiterbildung investieren müssen (z. B. durch den Berufseinstieg) als Ältere.

Bei den kulturellen Aktivitäten wirkt vor allem der Kinobesuch stark altersspezifisch (0.488 : 0.331). Alle anderen Aktivitäten sind - im Großen und Ganzen - altersunabhängig.

Nur der Besitz eines Musikinstruments, einer Stereoanlage, eines DVD-Players, Videorecorders oder Lexika bzw. Walk-, Disc- oder mp3-Players erreichen einen Zusammenhang über 0.2 - alle anderen Merkmale liegen darunter. In Teilen zeigen sich altersabhängige Effekte (z. B. Videospielkonsolen, Stereoanlage). Interessanterweise weichen die Werte für den Besitz eines Organizers/PDA kaum ab: 0.207 für alle Befragten, 0.176 für die bis 58jährigen. Auch gibt es kaum Unterschiede zwischen der Anzahl der Bücher im Haushalt (0.224 : 0.245).

Damit wird deutlich, dass Technik nicht nur von Jüngeren genutzt wird, sondern dass Ältere durchaus - im weitesten Sinne - Unterhaltungselektronik besitzen (die (Häufigkeit der) Nutzung ist natürlich davon unberührt). Bestimmte technische Geräte werden von Älteren jedoch weniger präferiert: dazu gehört neben dem PC z.B. ein Walk-, Disc- oder mp3-Player, also eher tragbare Geräte, die häufig außer-

halb der Wohnung eingesetzt werden (z. B. für Joggen, Fahrten in öffentlichen Verkehrsmitteln, u. ä.).

Deutlich wird, dass der PC-Nutzer nicht der eigenbrötlerische, kulturuninteressierte Mensch ist, der seine Wohnung nicht verläßt, sich Lebensmittel anliefern läßt und den ganzen Tag vor seinem Rechner sitzt. Diesen überzeichneten Typus mag es natürlich immer noch in Einzelfällen geben - und es existiert auch eine deutliche Affinität zu technischen Geräten (Videorecorder, Videospielkonsole, Walk-, Disc-, mp3-Player, etc.). Andererseits zeigen u. a. Weiterbildungsaktivitäten, Bibliotheks- und Kinobesuch, die Teilnahme an Sportveranstaltungen, der Besuch von Museen und Galerien, dass die Freizeit vielfältig genutzt wird - vor allem auch, da der PC bei Jüngeren zur selbstverständlichen Technologie wurde. Für die Anfangszeit des Personal Computers in den späten 70er/80er Jahren, wo Hard- und Software noch sehr teuer war und deshalb sich nur einige wenige Interessierte mit kryptischen Betriebssystemen, Programmiersprachen, etc. beschäftigten, mag dieses Bild noch Gültigkeit besitzen - vielleicht auch heute noch für den einen oder anderen Informatiker - aber die empirischen Ergebnisse dieser Arbeit widersprechen der Generalisierung dieses Typus.

Wie auch bei den sozialstrukturellen Merkmalen erfolgt die Auswahl der Variablen nach dem höchsten multivariaten Beitrag zum Gesamtmodell - ebenfalls wieder für die Entscheidungsbaumalgorithmen CART, QUEST und EXHAUSTIVE CHAID. In die Analyse gehen nachfolgende Variablen, die einen Zusammenhangswert > 0.2 aufweisen, ein.

TABELLE 32 SUBORDINIERTE BIVARIATE ZUSAMMENHÄNGE (> 0.20) ZWISCHEN PC-NUTZUNG (BIS 58 JAHRE) UND DEN KULTUR- UND FREIZEITVARIABLEN (PHI, CRAMERS V, UNSICHERHEITSKOEFFIZIENT (SIG = 0.000))[A]

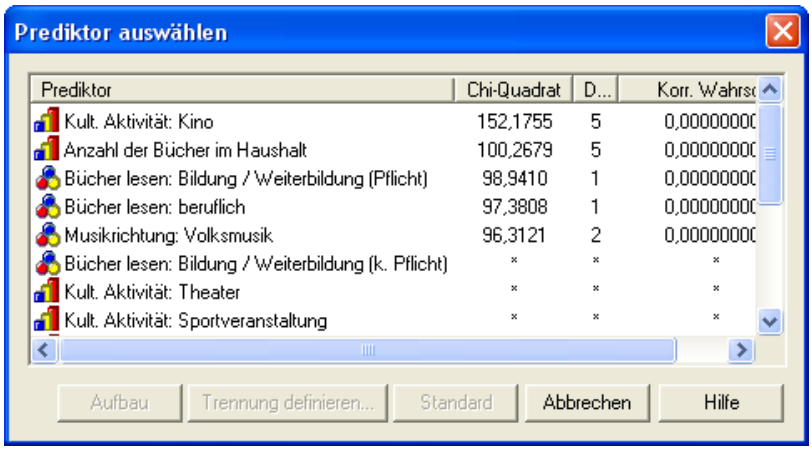
unabhängige Variable	Skalenniveau	Phi / Cramers v / Eta (Unsicherheitskoeff.)	
		Gesamt	bis 58 Jahre
Häufigkeit DVD/Videos sehen	ordinal	v = 0.406	v = 0.223
Gelesene Bücher i. d. letzten 12 Monaten	nom.-(MF)		
beruflich (v, u)		v = 0.333, u = 0.008	v = 0.26, u = 0.062
Bildg (Pfl.) (v, u)		v = 0.329, u = 0.086	v = 0.265, u = 0.063
Bildg (k.Pfl.) (v, u)		v = 0.24, u = (0.043	v = 0.220, u = 0.041
Art von Musik (Phi > (-)0.1, u)	nom.-dichotom		
Volksmusik		v = -0.421, u= 0.133	v = -0.252, u= 0.048
Kulturelle Aktivitäten (v)	ordinal		
Kino		v = 0.488	v = 0.331
Theater		v = 0.203	v = 0.206
Sportveranst.		v = 0.261	v = 0.207
Konzert		v = 0.237	v = 0.205
Bibliothek		v = 0.278	v = 0.229
Sehenswürdigk.		v = 0.209	v = 0.236
Konzertbesuch	nom.-dichotom		
Volksmusik		ph=-0.243, u=0.043	ph=-0.207, u=0.035
Medienbesitz	nom.-dichotom		
Walk/Disc/mp3-Pl		v = 0.386, u = 0.112	v = 0.269, u = 0.058
DVD-Player		v = 0.256, u = 0.059	v = 0.211, u = 0.047
Lexika (Buch)		v = 0.184, u = 0.025	v = 0.203, u = 0.03)
Lexika (CD-ROM)		v = 0.379, u = 0.126	v = 0.331, u = 0.116
Musikinstrument		v = 0.202, u = 0.03)	v = 0.206, u = 0.037
Anzahl der Bücher (v)	ordinal	v = 0.224	v = 0.245

a. Variablen, bei denen ein Zusammenhangsmaß unsinnig ist (z. B. Vornamen) wurden nicht berechnet und nicht in die Tabelle aufgenommen.

Die Variable „Lexika auf CD-ROM“ wurde aus der Analyse ausgeschlossen, da sie nur von PC-Besitzern genutzt werden kann. Das Er-

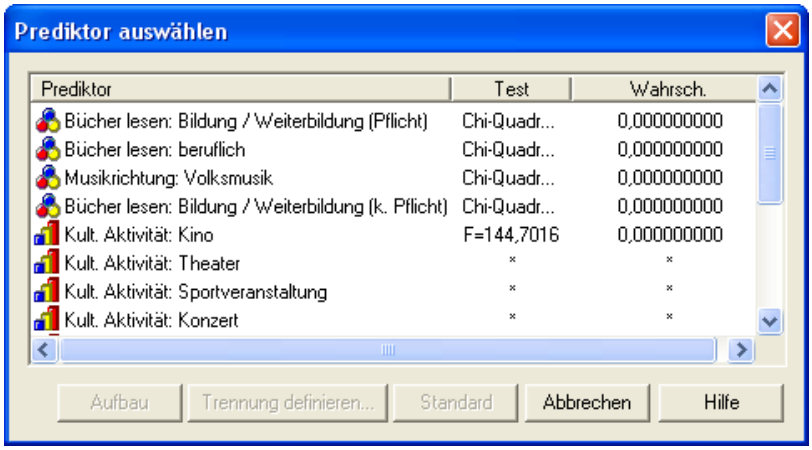
gebnis ist sehr interessant und verdeutlicht, dass eher Kultur- als Freizeitaktivitäten bei der PC-Nutzung im Vordergrund stehen.

ABBILDUNG 138 EXHAUSTIVE CHAID-Algorithmus: Wichtigste (subordinierte) Kultur- und Freizeitvariablen (Fehlklassifikation: 0.261)



Nur fünf der siebzehn Kultur- und Freizeitvariablen liefern einen wesentlichen Beitrag beim EXHAUSTIVE-CHAID-Modell (Fehlklassifikation: rund 26 %). Für QUEST ergibt sich:

ABBILDUNG 139 QUEST-Algorithmus: Wichtigste (subordinierte) Kultur- und Freizeitvariablen (Fehlklassifikation: 0.273)

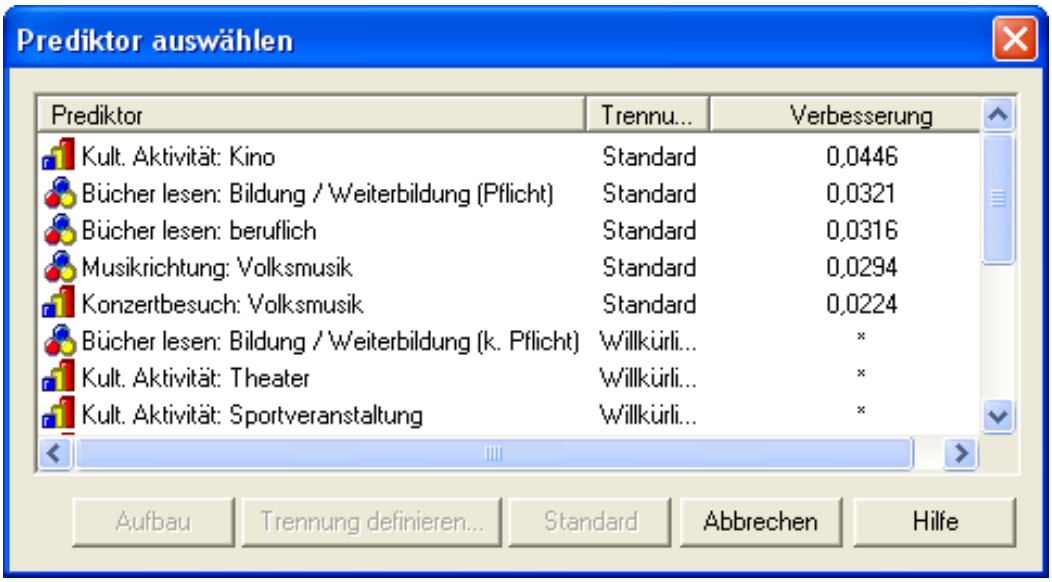


Die Fehlklassifikation ist mit 0.273 etwas schlechter als bei EXHAUSTIVE CHAID, allerdings unterscheiden sich die Variablen kaum: während

bei QUEST die nicht vom Arbeitgeber verpflichtende Weiterbildung als wichtige Variable herausgearbeitet wird, ist es bei EXHAUSTIVE CHAID die Anzahl der Bücher im Haushalt herangezogen.

CART ermittelt folgende Variablen:

ABBILDUNG 140 CART-Algorithmus: Wichtigste (subordinierte) Kultur- und Freizeitvariablen (Fehlklassifikation: 0.264)



Auch hier werden fünf wichtige Variablen gefunden. Zusammengefaßt ergibt sich:

TABELLE 33 ENTSCHEIDUNGSBAUMALGORITHMEN: DIE WICHTIGSTEN MULTIVARIAT ERMITTELTEN SUBORDINIERTEN FREIZEIT- UND KULTURVARIABLEN

	CART	QUEST	EXH. CHAID
Kulturelle Aktivität: Kinobesuch	ja	ja	ja
Bücher lesen: (Weiter-)Bildung Pflicht	ja	ja	ja
Bücher lesen: beruflich	ja	ja	ja
Musikrichtung: Volksmusik	ja	ja	ja
Konzertbesuch: Volksmusik	ja		
Bücher lesen: (Weiter-)Bildung keine Pflicht	-	ja	-
Anzahl der Bücher im Haushalt	-	-	ja

Vier von jeweils fünf identischen Variablen werden von allen Algorithmen zur Trennung der Bäume herangezogen - Weiterbildung - freiwillig oder vom Arbeitgeber verpflichtend -, Kinobesuch und Volksmusik hören. Die jeweils fünfte Variable variiert von Algorithmus zu Algorithmus: Volksmusik-Konzertbesuch (CART), private Weiterbildung (QUEST) und die Anzahl der Bücher im Haushalt (EXHAUSTIVE CHAID).

Auch wenn die Anzahl der Bücher keine Auskunft über ihren Inhalt gibt (es könnten sowohl Fachbücher für die Weiterbildung als auch „schöne Literatur“ oder Spieleanleitungen für PC-Spiele sein) scheinen es weniger Freizeitaktivitäten zur Zerstreuung als konkrete, evtl. verpflichtende Kulturaktivitäten zu sein, die die PC-Nutzer von den Nichtnutzern unterscheidet. Vor allem, wenn die Weiterbildungsbereitschaft einen derart großen Raum einnimmt. Kino und Volksmusik sind ganz typische Variablen, die in vielen Datensätzen zur Trennung von Gruppen beitragen. Obwohl sie deutlich altersspezifische Schwerpunkte besitzen (Kino eher bei den Jüngeren, Volksmusik eher bei den Älteren) ist es doch erstaunlich, dass diese Variablen bei den

bis 58jährigen eine so große Rolle spielen. Eine einfache Kreuztabelle verdeutlicht diesen Zusammenhang (aus Übersichtsgründen wurde auf eine recodierte Altersvariable zurückgegriffen) am Beispiel der Volksmusik:

ABBILDUNG 141 Altersgruppen und Volksmusikhören (N = 1382, Spalten-%)

Kreuztabelle

			Musikrichtung: Volksmusik		
			-	+	Gesamt
D11: Alter (gruppiert)	15 - 19 Jahre	Anzahl	107	5	112
		% von Musikrichtung: Volksmusik	10,5%	1,4%	8,1%
	20 - 29 Jahre	Anzahl	265	26	291
		% von Musikrichtung: Volksmusik	26,1%	7,1%	21,1%
	30 - 39 Jahre	Anzahl	323	68	391
		% von Musikrichtung: Volksmusik	31,8%	18,6%	28,3%
	40 - 49 Jahre	Anzahl	211	128	339
		% von Musikrichtung: Volksmusik	20,8%	35,0%	24,5%
	50 - 58 Jahre	Anzahl	110	139	249
		% von Musikrichtung: Volksmusik	10,8%	38,0%	18,0%
Gesamt		Anzahl	1016	366	1382
		% von Musikrichtung: Volksmusik	100,0%	100,0%	100,0%

Bereits in der Gruppe der ab 40jährigen überwiegt das Volksmusikhören deutlich (35 % der Volksmusikhörer vs. 20.5 % der Nichthörer). Der Prozentanteil steigt im Segment der 50 - 58jährigen nur noch geringfügig auf 38 % an.

Knapp ein Fünftel aller Volksmusikhörer ist jedoch zwischen 30 und 39 Jahre alt - ein Indiz dafür, dass diese Musikrichtung sich auch bei den eher Jüngeren einer gewissen Beliebtheit erfreut. Möglicherweise tra-

gen jüngere Moderatoren bzw. Interpreten wie Florian SILBEREISEN zu dieser Entwicklung bei.

Im direkten Vergleich der Altersgruppen zeigt sich, dass Volksmusik-Hören mit zunehmendem Alter deutlich ansteigt - nur im ältesten Segment (bei den 50 - 58jährigen) gibt es eine klare Mehrheit für diesen Musikstil. Allerdings scheint es nur bei den jüngsten Gruppen eine deutliche Ablehnung für diese Musik geben.

ABBILDUNG 142 Altersgruppen und Volksmusikhören (N = 1382, Zeilen-%)

Kreuztabelle

			Musikrichtung: Volksmusik		
			-	+	Gesamt
D11: Alter (gruppiert)	15 - 19 Jahre	Anzahl	107	5	112
		% von D11: Alter (gruppiert)	95,5%	4,5%	100,0%
	20 - 29 Jahre	Anzahl	265	26	291
		% von D11: Alter (gruppiert)	91,1%	8,9%	100,0%
	30 - 39 Jahre	Anzahl	323	68	391
		% von D11: Alter (gruppiert)	82,6%	17,4%	100,0%
	40 - 49 Jahre	Anzahl	211	128	339
		% von D11: Alter (gruppiert)	62,2%	37,8%	100,0%
	50 - 58 Jahre	Anzahl	110	139	249
		% von D11: Alter (gruppiert)	44,2%	55,8%	100,0%
Gesamt		Anzahl	1016	366	1382
		% von D11: Alter (gruppiert)	73,5%	26,5%	100,0%

Die Eindeutigkeit, mit der das Volksmusikhören stets nur den Befragten im Rentenalter zugeschrieben wird, ist somit nicht ganz richtig: vielmehr steigt der Anteil der Volksmusikhörer ab 40 Jahre deutlich

an. Auch wenn in diesem Teil der Auswertung nur die Befragten bis 58 Jahre herangezogen werden, zeigen sich doch deutliche Präferenzen auch bei jüngeren, wenn auch nicht den allerjüngsten Befragten, die deutlich mit dem Altersanstieg zunehmen.

Die gefundenen Kultur- und Freizeitaktivitäten machen deutlich, dass eine Segmentierung rein nach diesen Kategorien wenig ertragreich ist, da Alters- und Bildungseffekte wirken und die Aussagen wenig Sinn ergeben. Zwar ist dies auch bei Variablen wie Beruf und Einkommen der Fall, die Ergebnisse sind jedoch transparenter. Die Behauptung, „PC-Nutzer gehen gerne ins Kino, lehnen Volksmusik (sowohl die Musikrichtung als auch den Konzertbesuch) ab und bilden sich weiter“ ist recht pauschal und läßt sich schwer zu einem homogenen Bild zusammenführen. An diesem Punkt der Analyse erbringen die sozialstrukturellen Merkmale („Der PC wird in den jüngsten Befragtengruppen unabhängig vom Beruf stark genutzt, zwischen ca. 20 - 25 Jahren und ca. 40 Jahren entscheidet deutlich der Beruf, die über 40 - 58jährigen nutzen den PC, wenn das Haushaltsnettoeinkommen hoch und/oder der Beruf eine gewisse Nähe zur PC-Nutzung aufweist“) deutlich mehr, wenn auch noch nicht befriedigende Informationen über die PC-Nutzung. Aus diesem Grund ist eine weitergehende Analyse der gefundenen Entscheidungsbaumsegmente sinnvoll.

Interessant ist auch - und hier werden auch nochmals Kultur- und Freizeitvariablen in Frage gestellt - dass sehr häufig immer wieder die gleichen Variablen (z. B. Volksmusik = eher Ältere, Heavy Metal, Techno = eher Jüngere, Gartenzwerge = eher Kleinbürger), die sich teilweise schon zu Synonymen für bestimmten Lebenswelten und -stile herausgebildet haben. Kleinbürgerleben besteht aber ebensowenig nur aus Gartenzwergen wie Großbürgerleben aus Theaterbesuchen - was hier am Beispiel der Volksmusik deutlich wird. Natürlich wird

Volksmusik eher von den Älteren gehört - aber auch eine zunehmende Zahl von Jüngeren präferiert diese Musikrichtung. Auch die Fernsehmoderatoren bzw. Interpreten volkstümlicher Musik (Florian SILBEREISEN, Hansi HINTERSEER) sind heute jünger als die Moderatoren früherer Volksmusiksendungen (z. B. Karl MOIK), was ebenfalls dafür sprechen könnte, dass dies ein Versuch ist, auch jüngere Zuschauergruppen für volkstümliche Musik zu begeistern. Ein Zusammenhangswert ist schnell gebildet, eine multivariate Analyse bestätigt es nochmal. Gruppen werden etikettiert und gehen als Stereotype in die Literatur ein.

4 Multivariate Analyse II: Gruppenbildung

4.1 Methodisches Vorgehen bei der Gruppenbildung

In der Regel wird bei einer multivariaten Analyse das Ergebnis einer Methode - zum Beispiel zur Gruppenbildung (Clusteranalyse, Entscheidungsbaumalgorihmus, ...) zugrundegelegt, die methodisch oder theoretisch begründet werden.

Das könnte in diesem Fall geschehen - allerdings würde man die Möglichkeiten der unterschiedlichen Algorithmen ausser acht lassen. Gerade aber die unterschiedlichen Klassifizierungen eröffnen selbst dem statistischen Laien mit einem Blick, wo bestimmte Gruppen zu finden sind, die sehr hohe Anteile mit einem Merkmal (zum Beispiel der PC-Nutzung) aufweisen. Dies läßt sich zu einem Vorteil nutzen: durch die Möglichkeiten der Gewinnübersicht können Gruppen mit hohen bzw. geringen Nutzeranteilen über alle Algorithmen hinweg in homogeneren Clustern zusammengefaßt werden.

Da alle Algorithmen sich - aufgrund ihrer statistischen und theoretischen Grundlagen - unterscheiden, sollten sie auch zu etwas unter-

schiedlichen Segmenten gelangen. Warum diese Informationen nicht nutzen und die Segmente wieder zusammenfassen?

Wenn zum Beispiel in einer Gruppe, die sehr wenig mit PC-Nutzung assoziiert wird, eine bestimmte Altersgruppe oder bestimmte Einkommensklassen deutlich häufiger den PC nutzen als der Rest - warum sollte man auf diese Information verzichten?

Die durchschnittliche Nutzung der Reise- und Dienstleistungsberufe liegt bei 57.2 % - also eine eher unterdurchschnittliche Nutzung. Es gibt aber Reise- und Dienstleistungssegmente, die älter als 35 Jahre sind und mehr als 4500 DM verdienen, deren Nutzeranteil bei 90.5 % (CART) liegt. Warum sollte man dies unberücksichtigt lassen, nur weil man auf einen anderen Algorithmus zurückgegriffen hat, der dies nicht segmentiert?

Für dieses Beispiel wurde auf die Überlegungen GEIGERs zurückgegriffen, Berufe „aszendierend" (= aufsteigend) zusammenzufassen. Während GEIGER aus der Berufszählung von 1925 seine Gruppen sowohl explorativ als auch theoretisch, immer im Hinblick auf Mentalitäten fand, sollen hier die PC-Nutzeranteile explorativ nach der Höhe der Nutzeranteile gegliedert werden.

Die Gewinnübersicht der Entscheidungsbaumalgorithmen bietet hier eine gute Möglichkeit, Segmente mit ähnlich hohen Nutzeranteilen zu identifizieren und zusammenzufassen. Wie weiter oben gesehen, kommen die Algorithmen teilweise zu anderen Ergebnissen, da sie unterschiedliche Merkmale als Trennungsvariablen heranziehen (z. B. CART: Beruf, QUEST: Alter). Alle Algorithmen kommen jedoch zu Knoten, die unterschiedlich zusammengesetzt sind, aber einen ähnlich hohen Nutzeranteil aufweisen. Diese Gruppen können zusammengefaßt werden.

Da es sich in diesem Schritt der Untersuchung um eine deskriptive Analyse handelt und ein möglichst umfangreiches Schichtungsbild mit vielfältigen Gruppen entstehen soll, die in einem weiteren Schritt anhand der PC-Nutzeranteile zusammengefaßt werden, wurde die Größe des Hauptknotens mit mindestens 30, die der Unterknoten mit mindestens 20 angegeben (was für eine multivariate Analyse zu geringe Fallzahlen produziert).

Um das Ergebnis nicht zu unübersichtlich werden zu lassen wird ein Ausschnitt aus der Gewinnübersicht des CART- und EXHAUSTIVE-CHAID-Algorithmus illustrierend herangezogen:

TABELLE 34 EXHAUSTIVE-CHAID-ALGORITHMUS: GEWINNÜBERSICHT (AUSSCHNITT) FÜR PC-NUTZUNG (ABH. VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN U. BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN ≥ 30, UNTERKNOTEN ≥ 20 (N, %)

Knoten	Knoten: Anzahl	Knoten %	Gewinn Anzahl	Gewinn %	Treffer %
6 Studierende	128	9,1	121	13,3	94,5
9 Freie Berufe	38	2,7	35	3,9	92,1
5 Leitende Büroberufe	187	13,2	168	18,5	89,8
3 sonstige Büroberufe	167	11,8	149	16,4	89,2
2 Großunternehmer, Direktoren, Top Management	63	4,5	56	6,2	88,9

TABELLE 35 CART-ALGORITHMUS: GEWINNÜBERSICHT (AUSSCHNITT) FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN ≥ 30, UNTERKNOTEN ≥ 20, GEORDNET NACH TREFFERN (N, %)

Knoten	Knoten: Anzahl	Knoten: %	Gewinn Anzahl	Gewinn %	Treffer: %
11 Büroberufe, Freie Berufe, Studierende, Meister, Landwirte, Fischer bis 45 Jahre	456	32,3	426	46,9	93,4
24 Reise- u. Dienstleistungsangestellte, Handwerker, Ladenbesitzer, älter als 35 Jahre, Einkommen > 4500 DM	21	1,5	19	2,1	90,5

Die Tabelle vereinfacht sich, wenn der jeweilige Knoten herangezogen wird.

ABBILDUNG 143 CART-ALGORITHMUS: KNOTEN 11 FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTO-EINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN ≥ 30, UNTERKNOTEN ≥ 20 (N, %)

Knoten 11			
	Kategorie	%	n
	Non User	6,58	30
	User	93,42	426
	Gesamt	(32,27)	456

Die Spalte „Treffer %“ gibt den Anteil (93.4 % User) dieses Knotens an. Die Spalten „Knoten Anzahl“ und „Gewinn Anzahl“ geben mit N = 456 die Gesamtzahl der Befragten in Knoten 11 bzw. die Anzahl der User (N = 426) an - Knoten- und Gewinn-% die Anteile.

Gibt es weitere Segmente, die einen ähnlich hohen Useranteil (um 90 %) aufweisen? - Diese Segmente ließen sich zusammenfassen und könnten die Struktur der PC-Nutzer deutlich erhellen. Bei CART kommt überraschenderweise eine zweite, recht kleine Gruppe der über 35jährigen Reise- und Dienstleistungsangestellten, Handwerker und Ladenbesitzer, die mehr als 4500 DM verdienen, hinzu - eigentlich ein recht kleines Segment mit 21 Befragten, die wohl nur sehr schwer mit einem anderen multivariaten Verfahren identifiziert werden könnte, da es „quer“ zu den Logiken der PC-Nutzung liegt (PC-Nutzer = eher jünger, eher Büroberufe, jedoch höheres Einkommen).

In dieser Form können alle PC-Useranteile über alle Algorithmen zusammengefaßt werden. Ebenso ließen sich Segmente mit hohen, durchschnittlichen und niedrigen Anteilen finden. Dies stellt das nächste Problem dar: wieviele Gruppen werden gebildet und wo werden die Grenzen für die Gruppenbildung gezogen?

Eine Möglichkeit bietet ein theoriegeleitetes Vorgehen, das aus Ergebnissen anderer Untersuchungen die Anzahl und Grenzen der Gruppen angibt. Das ist hier nicht der Fall, da es keine Voruntersuchungen in dieser Form zu diesem Thema gibt. Deshalb können folgende Regeln zur Anwendung kommen:

- Die durchschnittliche Nutzung liegt bei 64.33 % (knapp 2/3): Segmente über diesen Nutzeranteilen repräsentieren eher PC-Nutzer, darunter eher PC-Nichtnutzer.
- Die gefundenen Gruppen sollten eine Gruppengröße >100 repräsentieren, um sinnvolle multivariate Aussagen treffen zu können.
- Grenzen können dort gebildet werden, wo „Lücken" (Prozentsatzdifferenzen) zwischen der Verteilung gefunden werden. Ergeben sich z. B. folgende Segmente: 90 %, 86 %, 81 %, 69 %, 67 %, ... Nutzeranteile, so könnte man eine Trennung zwischen 90 %, 86 % und 81 % einerseits und 69 % bzw. 67 % andererseits vornehmen, also zwei Gruppen bilden: eine mit stark überdurchschnittlicher, eine mit etwa durchschnittlichen Useranteilen,
- Gruppen < etwa 50 % Useranteile sind als bedroht anzusehen, was zukünftige gesellschaftliche Teilhabe angeht
- Gruppen < etwa 33 % Nutzeranteile sind als prekär hinsichtlich der gesellschaftlichen Teilhabe anzusehen - 2/3 dieser Befragten sind von zukünftigen technologischen Entwicklungen im Informationsbereich abgeschnitten

Die Definition dieser Regeln ist, wie jede Entscheidung, sowohl methodisch als auch theoretisch angreifbar. Am Beispiel der CART-Segmentierung soll der Punkt der Grenzziehung zwischen den Gruppen erläutert werden:

TABELLE 36 CART-ALGORITHMUS: GEWINNÜBERSICHT FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN $\geq$ 30, UNTERKNOTEN $\geq$ 20 (N, %)

Knoten	Knoten: %	Gewinn (%)	Treffer: %
11 Büroberufe, Studierende, Meister, Landwirte, Fischer bis 45 Jahre	32,3	46,9	93,4
24 Reise- u. Dienstleistungsangestellte, Handwerker, Ladenbesitzer, älter als 35 Jahre, Einkommen > 4500 DM	1,5	2,1	90,5
12	10,0	13,1	83,8
23	2,1	2,4	75,9
21	3,4	3,9	72,9
16	4,2	4,3	65,0
6	1,6	1,5	60,9
17	1,6	1,3	52,2
29	6,3	5,0	50,6
28	13,0	10,2	50,5
19	3,5	1,8	32,0
18	2,6	1,2	29,7
30	4,6	2,1	29,2
26	2,5	1,0	25,7
14	5,5	2,2	25,6
25	5,2	1,1	13,7

Knoten 11 und 24 enthalten ähnlich hohe Nutzeranteile um 90 % und liegen mit den Nutzeranteilen recht nahe beieinander (93.4, 90.5). Die Differenz zwischen Knoten 24 und Knoten 12 ist relativ hoch und veranlaßt dazu, eine zweite Gruppe zu bilden, die die Knoten 12, 21, 23 und 16 umfaßt. Diese beiden Gruppen repräsentieren überdurchschnittliche Nutzeranteile zwischen 65 % und 83 % bzw. > 90 %. Kno-

ten 6, 17, 29 und 28 repräsentieren Nutzeranteile zwischen 50 % und dem Durchschnitt mit 64.3 %. Die verbleibenden Knoten enthalten deutlich überproportionierte Nichtnutzeranteile.

Wie angedeutet - eine derartige Segmentierung ist möglich, sie stellt aber nicht den einzigen Weg dar, Gruppen zu bilden. Die Grenze bei der durchschnittlichen Nutzung anzusetzen (eher Nutzer - eher Non-User) scheint sinnvoll und wird weniger Diskussionen aufwerfen. Die jedoch recht explorative Trennung der Segmente aufgrund von größeren Lücken bei Prozentsatzdifferenzen ist durchaus diskussionsfähig, wird aber in dieser Arbeit so unterstellt: Ich weise darauf hin, dass jede andere Segmentierung ebenso diskussionswürdig ist. Der Vorteil liegt darin, dass alle Algorithmen Prozentsatzdifferenzen bei ähnlichen Prozentwerten aufweisen und somit dieses Vorgehen gestützt wird.

4.2 Beschreibung der Segmente

4.2.1 Gruppe 1: Stark überdurchschnittliche Nutzeramteile: (meist) PC-bezogene Berufe

Aus den Algorithmen ergeben sich folgende Segmente:

TABELLE 37 STARK ÜBERDURCHSCHNITTLICHE NUTZERANTEILE (> 88.9 %) FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN ≥ 30, UNTERKNOTEN ≥ 20 (N, TREFFER-%)

Algorithmus Knoten	Beschreibung	Knoten	Knoten Anzahl	Gewinn Anzahl	Treffer %
CART	Büroberufe, Freie Berufe, Studierende, Meister, Landwirte, Fischer bis 45 Jahre	11	456	426	93,4
CART	Reise- u. Dienstleistungsangestellte, Handwerker, Ladenbesitzer, älter als 35 Jahre, Einkommen > 4500 DM	24	21	19	90,5
QUEST	Büroberufe, Studierende, Freie Berufe, Meister, Landwirte, Fischer > 2750 DM, bis 54 Jahre	22	452	418	92,5
QUEST	Studierende, Leitende Büroberufe, sonstige Bürotätigkeiten< 1750 DM, < 54 Jahre	8	52	47	90,4
EXH. CHAID	Studierende	6	128	121	94,5
EXH. CHAID	Freie Berufe	9	38	35	92,1
EXH. CHAID	Leitende Büroberufe	5	187	168	89,8
EXH. CHAID	sonstige Büroberufe	3	167	149	89,2
EXH. CHAID	Großunternehmer, Direktoren, Top Management	2	63	56	88,9

Es gibt sehr viele Überschneidungen bei den Algorithmen (z. B. alle Büroberufe), aber auch einige Unterschiede in den Klassifikationen

Alter und Haushaltsnettoeinkommen. Die Gruppe läßt sich am besten folgendermaßen charakterisieren:

- CART: Büroberufe, Freie Berufe, Studierende, Meister, Landwirte, Fischer bis 45 Jahre
- QUEST: **Großunternehmer, Direktoren, Top Management, Leitende Büroberufe, sonstige Büroberufe, Studierende, Freie Berufe, Meister, Landwirte, Fischer > 2750 DM, bis 54 Jahre**
- CART: Reise- u. Dienstleistungsangestellte, Handwerker, Ladenbesitzer, älter als 35 Jahre, Einkommen > 4500 DM

Alle anderen Segmentierungen (vor allem die EXHAUSTIVE CHAID-Gruppen) sind bereits in anderen Gruppen definiert.

Deutlich wird an dieser Segmentierung, dass die PC-bezogenen Berufe nicht alleine ausschlaggebend für die PC-Nutzung sind, sondern auch Alter und Haushaltsnettoeinkommen eine gewisse Rolle spielen (QUEST-Segment bis 54 Jahre). Dies gilt auch für das zweite CART-Segment der über 35jährigen mit hohem Einkommen und Berufen, die wahrscheinlich weniger mit dem PC zu tun haben. Unbestritten spielt jedoch der Beruf für die stärkste Gruppe der PC-Nutzer eine überragende Rolle.

Mit dieser Synopse der Gruppen aus den einzelnen Algorithmen wird deutlich, dass es nicht den „besten“ oder „schlechtesten“ Algorithmus gibt. Auch wenn CART - für sich genommen - in diesem Fall die sicherlich sinnvollste Segmentierung findet, die gut interpretierbar ist. Allerdings wäre es bedauerlich, auf die Zusatzinformationen, die QUEST und EXHAUSTIVE CHAID (letzterer weiter unten) liefen, zu verzichten. Mit dieser Mischung aus multivariaten und deskriptiven Vorgehen sollen die Vorteile für die Lösung, was die Betrachtung mehrerer unabhängiger Variablen und die Verständlichkeit der Lösung angeht, deutlich erhöht werden. Sind diese Segmente definiert, werden sie weiter nach Kultur- und Freizeitvariablen charakterisiert.

Die Gruppen lassen sich recht einfach in SPSS abbilden: es besteht die Möglichkeit, die Syntax für jeden einzelnen gefundenen Knoten in SPSS zu transferieren.[92]

4.2.2 Gruppe 2: (über-)durchschnittliche Nutzeranteile (geringeres Alter, höheres Einkommen - bis 65 %)

TABELLE 38 (ÜBER-)DURCHSCHNITTLICHE (65 % - 80 %) NUTZERANTEILE FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN $\geq$ 30, UNTERKNOTEN $\geq$ 20 (N, TREFFER-%)

Algorithmus	Beschreibung	Knoten	Knoten Anzahl	Gewinn Anzahl	Treffer %
CART	Büroberufe, Freie Berufe, Studierende, Meister, Landwirte, Fischer 46 bis 56 Jahre	12	142	119	83,8
CART	Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker bis 35 Jahre, Einkommen > 4500 DM	23	29	22	75,9
CART	Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker, Arbeiter bis 21 Jahre, Einkommen < 4500 DM	21	48	35	72,9
CART	Arbeiter, nie erwerbstätig, bis 45 Jahre, Einkommen > 4500 DM	16	60	39	65,0
QUEST	Büroberufe, Meister, Landwirte, Fischer, Reise- und Dienstleistungsangestellte, bis 53 Jahre, bis 2250 DM	12	45	38	84,4
QUEST	Büroberufe, Studierende, Meister, Landwirte, Fischer bis 53 Jahre, bis 2750 DM	16	38	31	81,6
QUEST	Büroberufe, Meister, Landwirte, Fischer, FREIE BERUFE, älter als 53 Jahre	6	51	37	72,5

92. Answertree stellt im Fenster „Regeln" Syntaxen für SPSS, SAS und SQL für jeden einzelnen Knoten zur Verfügung. Diese Knoten können problemlos über die Zwischenablage in SPSS kopiert und dort weiterverwendet werden. Für SPSS werden IF-Befehle generiert.

TABELLE 38 (ÜBER-)DURCHSCHNITTLICHE (65 % - 80 %) NUTZERANTEILE FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN ≥ 30, UNTERKNOTEN ≥ 20 (N, TREFFER-%)

Algorithmus	Beschreibung	Knoten	Knoten Anzahl	Gewinn Anzahl	Treffer %
QUEST	Reise- u. Dienstleistungsangestellte, Ladenbesitzer, Handwerker bis 47 Jahre, > 2750 DM	27	134	92	68,7
QUEST	Arbeiter, nie erwerbstätig, bis 23 Jahre, > 3000 DM	29	32	21	65,6
EXH. CHAID	Reise- und Dienstleistungsangestellte, > 4000 DM	14	52	42	80,8
EXH. CHAID	Meister, Landwirte, Fischer ohne weitere Differenzierung	10	38	30	78,9
EXH. CHAID	Arbeiter bis 20 Jahre, ohne weitere Differenzierung	15	39	28	71,8

In diesem Segment wird noch deutlicher als in der stärksten Nutzergruppe, wie Einkommen und Alter die PC-Nutzung beeinflussen, wenn kein PC-bezogener Beruf vorliegt. In einigen dieser Segmente spielt auch das Einkommen keine Rolle, wenn es sich um jüngere Befragte handelt (z. B. EXH. CHAID, Knoten 15: Arbeiter bis 20 Jahre). Die Bedeutung des Alters sinkt, wenn das Einkommen hoch genug ist (z. B. EXHAUSTIVE CHAID, Knoten 14: Reise- und Dienstleistungsangestellte mit einem Einkommen > 4000 DM).

Auch hier lassen sich die Gruppen wieder vereinfachen:

- CART: Büroberufe, Freie Berufe, Studierende, Meister, Landwirte, Fischer 46 bis 56 Jahre
- QUEST: **Büroberufe, Meister, Landwirte, Fischer, Reise- und Dienstleistungsangestellte, bis 53 Jahre, bis 2250 DM**
- QUEST: Büroberufe, Studierende, Meister, Landwirte, Fischer bis 53 Jahre, bis 2750 DM
- QUEST: Büroberufe, Meister, Landwirte, Fischer, FREIE BERUFE, 53 Jahre +
- EXH. CHAID: Meister, Landwirte, Fischer ohne weitere Differenzierung

Die eher PC-bezogenen Berufe vereinen alle Büroberufe, Freien Berufe, Studierende, Meister Landwirte und Fischer - wenn sie über 45 Jah-

re bis 56 Jahre (CART), bis 53 Jahre alt sind, aber ein Einkommen bis 2750 DM beziehen (QUEST). Werden die Reise- und Dienstleistungsangestellten einbezogen, sinkt die Einkommensgrenze auf 2250 DM.

Hier finden sich auch zwei Segmente der Älteren: zum einen eher PC-bezogene Berufe (Büro, Freie Berufe, Meister, Landwirte, Fischer), die zwischen 53 und 58 Jahre alt sind und das von EXHAUSTIVE CHAID gefundene Segment der Meister, Landwirte und Fischer ohne weitere Differenzierung (vgl. Abbildung 38 auf Seite 320).

- CART: Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker bis 35 Jahre, Einkommen > 4500 DM
- QUEST: Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker bis 47 Jahre, > 2750 DM
- CART: Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker, Arbeiter bis 21 Jahre, Einkommen < 4500 DM
- CART: Arbeiter, nie erwerbstätig, bis 45 Jahre, Einkommen > 4500 DM
- QUEST: Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker bis 47 Jahre, > 2750 DM
- QUEST: Arbeiter, nie erwerbstätig, bis 23 Jahre, > 3000 DM
- EXHAUSTIVE CHAID: Reise- und Dienstleistungsangestellte, > 4000 DM
- EXHAUSTIVE CHAID: Arbeiter bis 20 Jahre, ohne weitere Differenzierung

Diese Differenzierung der weniger PC-bezogenen Berufe macht nochmals den Einfluß von Alter (z. B. EXHAUSTIVE CHAID: Arbeiter bis 20 Jahre, ohne weitere Differenzierungen) bzw. Haushaltsnettoeinkommen (z. B. EXHAUSTIVE CHAID: Reise- und Dienstleistungsangestellte > 4000 DM) deutlich. Der Anteil der Arbeiter liegt z. B. bei 41.8 % - also ein recht unterdurchschnittlicher Wert bei einer Durchschnittsnutzung von ca. 2/3.

Die weiteren Altersdifferenzierungen zeigen, dass der PC in ein bis zwei Jahrzehnten zum Alltagsgegenstand werden wird - allerdings mit einigen Gefahren des gesellschaftlichen Ausschlusses. So ist der Anteil der nie Erwerbstätigen mit 27 % über alle Altersgruppen sehr gering. Kommt noch ein geringes Einkommen dazu (unter 3000 DM bzw. 4500 DM), ist der gesellschaftliche Ausschluss vorprogrammiert -

nicht nur, was die PC-Nutzung angeht, sondern der Internetnutzung. Hier wird sich jedoch auch in den nächsten Jahren zeigen, ob sich neue Technologien (z. B. über den Fernseher) des Internetzugangs durchsetzen werden.

Die Probleme verschärfen sich heute schon: viele Unternehmen setzen häufig aus Kostengründen offene Stellen auf ihre Homepage oder in Job-Suchmaschinen, nicht mehr in die öffentlichen Tageszeitungen. Gerade für nicht Erwerbstätige, die auf Jobsuche sind, wird dies in Zukunft eine deutliche Barriere darstellen.

QUEST findet in diesem Fall etwas differenziertere Segmente als CART:

- CART: Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker bis 35 Jahre, Einkommen > 4500 DM
- QUEST: Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker bis 47 Jahre, > 2750 DM
- CART: Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker, Arbeiter bis 21 Jahre, Einkommen < 4500 DM

Die Berufsgruppen sind identisch, die QUEST-Lösung ist in diesem Fall allerdings etwas informativer, da sie eine breitere Gruppe (bis 47 anstatt bis 35 Jahre, Einkommen > 2750 DM anstatt > 4500 DM) liefert. Das zweite CART-Segment ist in einer QUEST-Gruppe enthalten.

- QUEST: Arbeiter, nie erwerbstätig, bis 23 Jahre, > 3000 DM
- EXHAUSTIVE CHAID: Reise- und Dienstleistungsangestellte, > 4000 DM
- EXHAUSTIVE CHAID: Arbeiter bis 20 Jahre, ohne weitere Differenzierung

An dieser Stelle wird nochmals deutlich, dass EXHAUSTIVE CHAID - obwohl die Segmentierung nicht sehr differenziert erfolgte - durchaus interessante Beiträge durch die Reise- und Dienstleistungsangestellten > 4000 DM und die Arbeiter bis 20 Jahre liefert: diese Gruppen müssen nicht weiter segmentiert werden, was zu einer Erleichterung des Verständnisses beiträgt.

Dieses Segment läßt sich weniger durch den Beruf, sondern mehr durch Alter und Einkommen charakterisieren. Vorsichtig ausgedrückt:

nicht der Beruf, sondern hohes Einkommen oder niedriges Lebensalter spielt hier eine Rolle für den Einsatz eines PCs.

4.2.3 Gruppe 3: unterdurchschnittliche Nutzeranteile (bedrohte Lagen 64 % - 33 %)

Diese Gruppe der (leicht) unterdurchschnittlichen Nutzeranteile zwischen 33 % und 64 % machen Lagen innerhalb der PC-Nutzer aus, die bedroht sind, von zukünftigen gesellschaftlichen Entwicklungen ausgeschlossen zu werden.

TABELLE 39 (UNTER-)DURCHSCHNITTLICHE (CA. 34 % - 64 %) NUTZERANTEILE FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN $\geq$ 30, UNTERKNOTEN $\geq$ 20 (N, TREFFER-%)

Algorithmus	Beschreibung	Knoten	Knoten Anzahl	Gewinn Anzahl	Treffer
CART	Büroberufe, Freie Berufe, Meister, Landwirte, Fischer > 55 Jahre	6	23	14	60,9
CART	Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker 46 bis 50 Jahre	17	23	12	52,2
CART	Reise- und Dienstleistungsangestellte, Arbeiter, Handwerker, Ladenbesitzer, 22 - 31 Jahre, < 2750 DM	29	89	45	50,6
CART	Reise- u. Dienstleistungsangestellte, Arbeiter, Handw., Ladenbesitzer, 22-45 Jahre, >2750 DM	28	184	93	50,5
QUEST	Arbeiter, nie erwerbstätig, 23-29 Jahre, >3500 DM	30	36	18	50,0
QUEST	Arbeiter, nie erwerbstätig, 41-42 Jahre, >2750 DM	31	24	12	50,0
QUEST	Arbeiter, nie erwerbstätig, bis 30 Jahre, <3500 DM	23	29	13	44,8
QUEST	Arbeiter, nie erwerbstätig, 30-40 Jahre, >2750 DM	25	135	59	43,7
QUEST	Arbeiter, nie erwerbstätig, 23-29 Jahre, >2750 DM	33	48	20	41,7

TABELLE 39 (UNTER-)DURCHSCHNITTLICHE (CA. 34 % - 64 %) NUTZERANTEILE FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN ≥ 30, UNTERKNOTEN ≥ 20 (N, TREFFER-%)

Algorithmus	Beschreibung	Knoten	Knoten Anzahl	Gewinn Anzahl	Treffer
QUEST	Reise- und Dienstleistungsangestellte, Ladenbesitzer, Handwerker, > 2750 DM, > 53 Jahre,	28	26	10	38,5
QUEST	Reise-, Dienstleistungsangestellte, Arbeiter, Ladenbesitzer, Handw., nie erwerbstätig, bis 53 Jahre, bis 2750 DM.	15	81	31	38,3
QUEST	Arbeiter, nie erwerbstätig, bis 53 Jahre, 1500 DM - 2250 DM	11	40	14	35,0
EXH. CHAID	Ladenbesitzer, Handwerker ohne weitere Differenzierung	7	28	17	60,7
EXH.C.	Reise- und Dienstleistungsangestellte bis 2250 DM	11	32	18	56,3
EXH. CHAID	Reise- und Dienstleistungsangestellte 2501 - 4500 DM	13	94	50	53,2
EXH. C	Arbeiter, 20 - 48 Jahre	16	311	140	45,0

Interessanterweise sind häufig die Ressourcen in Form von Haushaltsnettoeinkommen vorhanden (vor allem bei den Arbeitergruppen - z. B. QUEST-Knoten 25 und 31: Arbeiter und nie Erwerbstätige mit einem Haushaltsnettoeinkommen über 2750 DM). Obwohl hier auch niedrigere Einkommensgruppen vertreten sind, scheint das Haushaltsnettoeinkommen nicht das ausschlaggebende Kriterium zu sein, sondern möglicherweise eher eine bewußte Verweigerung - sowohl bei den Älteren als auch bei den Jüngeren, denn die Bandbreite des Alters reicht von 22 - 58 Jahre. Vor allem Reise- und Dienstleistungsangestellte, Arbeiter und nie Erwerbstätige finden sich in diesem Bereich wieder - und Büroangestellte, die älter als 55 Jahre sind. Etwas plakativ könnte man also diese Gruppe als „bewußte Verweigerer“ charakterisieren. Von den Berufen sind sie sich noch am ähnlichsten, Alter und Haushaltsnettoeinkommen differieren hier stark.

4.2.4 Gruppe 4: stark unterdurchschnittliche Nutzeranteile (prekäre Lagen)

Hier finden sich Befragte, wo nur maximal jeder Dritte je Segment den PC nutzt. Diese Gruppe könnte man - für zukünftige Entwicklungen - als prekär bezeichnen. Sie sind sehr wahrscheinlich von zukünftigen gesellschaftlichen Entwicklungen in Teilen ausgeschlossen. Zum einen, was Arbeitsplatzsuche angeht, zum anderen, was Qualifikationen für den Arbeitsmarkt angeht. Aber auch andere Möglichkeiten (Informationsbeschaffung über Internet, email, Newsletter, etc.) werden einem Großteil dieser Gruppe verbaut sein. Sie setzt sich zusammen aus:

TABELLE 40 STARK UNTERDURCHSCHNITTLICHE (BIS 33 %) NUTZERANTEILE FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN $\geq$ 30, UNTERKNOTEN $\geq$ 20 (N, TREFFER-%)

Algorithmus	Beschreibung	Knoten	Knoten Anzahl	Gewinn Anzahl	Treffer
CART	Arbeiter, 46 - 49 Jahre	19	50	16	32,0
CART	Reise- und Dienstleistungsangestellte, 51 Jahre und älter	18	37	11	29,7
CART	Reise- und Dienstleistungsangestellte, Arbeiter, Ladenbesitzer, Handwerker, 31 - 45 Jahre, bis 2750 DM	30	65	19	29,2
CART	Arbeiter, nie erwerbstätig, 55 Jahre und älter	26	35	9	25,7
CART	nie erwerbstätig, bis 45 Jahre, bis 4500 DM	14	78	20	25,6
CART	Arbeiter, nie erwerbstätig, 50 - 54 Jahre,	25	73	10	13,7
QUEST	Reise- und Dienstleistungsangestellte, Arbeiter, nie erwerbstätig, bis 53 Jahre, bis 1750 DM	7	67	20	29,9

TABELLE 40 STARK UNTERDURCHSCHNITTLICHE (BIS 33 %) NUTZERANTEILE FÜR PC-NUTZUNG (ABHÄNGIGE VARIABLE), ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (UNABHÄNGIGE VARIABLEN) FÜR 10 EBENEN, HAUPTKNOTEN ≥ 30, UNTERKNOTEN ≥ 20 (N, TREFFER-%)

Algorithmus	Beschreibung	Knoten	Knoten Anzahl	Gewinn Anzahl	Treffer
QUEST	Reise- und Dienstleistungsangestellte, Arbeiter, Ladenbesitzer, Handwerker, nie erwerbstätig, 54 Jahre und älter	5	70	18	25,7
QUEST	Arbeiter, nie erwerbstätig, 43 - 48 Jahre, > 2250 DM	34	53	10	18,9
EXH. CHAID	nie erwerbstätig ohne weitere Differenzierung	8	118	32	27.1
EXH. CHAID	Reise- und Dienstleistungsangestellte 2000 - 2750 DM	12	23	5	21.7
EXH. CHAID	Arbeiter, 49 Jahre und älter	17	95	18	18.9

Der Zusammenhang der vier Nutzergruppen macht deutlich, warum der Beruf eine so überragende Trennungsvariable ist: Büroberufe, Freie Berufe, Meister, Landwirte und Fischer tauchen nur in den überdurchschnittlichen Gruppen auf (abgesehen von einem kleinen Segment der 55 - 58jährigen bei der leicht unterdurchschnittlichen Nutzung). Die geringeren Useranteile finden sich fast ausschließlich bei den Arbeitern, Reise- und Dienstleistungsangestellten, nie Erwerbstätigen und Handwerkern bzw. Ladenbesitzern (mit eher geringerem Einkommen), wobei letztere zwei Segmente nur in den leicht unterdurchschnittlichen Usergruppen auftauchen.

Somit läßt sich feststellen, dass es grundsätzlich drei Berufsgruppen gibt, die den PC sehr gering nutzen: Arbeiter, Reise- und Dienstleistungsangestellte und nie Erwerbstätige - vor allem, wenn sie älter sind.

Bei den jüngeren Befragtengruppen scheint das Einkommen eine gewisse Rolle zu spielen (vgl. CART-Knoten 30, QUEST-Knoten 7): das Einkommen liegt mit 2750 bzw. 1750 DM nicht sehr hoch, es handelt sich aber mit um die jüngsten Befragtengruppen in diesem Segment.

Möglicherweise wirken hier Einkommensdefizite bei den Jüngeren und Desinteresse - vor allem bei den Älteren.

4.2.5 Zusammenfassung der Gruppen

Rein technisch ist es kein Problem, die segmentierten Gruppen in SPSS zu transferieren. Die Regeln lassen sich problemlos als SPSS- (SQL- oder SAS-) Syntax ausgeben - beispielsweise für die Reise- und Dienstleistungsangestellten bis 2250 DM Einkommen (EXHAUSTIVE CHAID-Algorithmus):

```
SELECT IF ((BERSTEL = 3)  AND  (HHNETTO <= 4)).
```

Die SPSS-Befehlszeile gibt an, dass alle Befragten aus der Variable BERSTEL (= Beruf), die den Wert 3 (= Reise- und Diensleistungsangestellte) aufweisen, herangezogen werden. Das Haushaltsnettoeinkommen ist in Gruppen zusammengefaßt, die vierte Gruppe enthält die Einkommenswerte 2000 - 2250 DM. Folglich werden alle Reise- und Dienstleistungsangestellten, die bis 2250 DM beziehen, ausgewählt.

Die meisen Bedingungen sind wesentlich komplexer. Hier ergibt sich auch für die Zusammenfassung ein kleines Problem, wenn man die teilweise etwas differierenden Ergebnisse der unterschiedlichen Algorithmen zusammenfassen will. Diese Arbeit erinnert etwas an die Klassifikation in der qualitativen Sozialforschung, wo Kategorien aufgrund von Inhalten überprüft werden (Typenbildung). Auch hier ist anzumerken, dass dieses Vorgehen durchaus kritikfähig ist - aber es gibt auch hier verschiedene Wege ans Ziel zu gelangen.

ABBILDUNG 144 Answertree-Segmente: PC-Nutzeranteile (N, %)

Answertree-Gruppen nach Nutzeranteilen

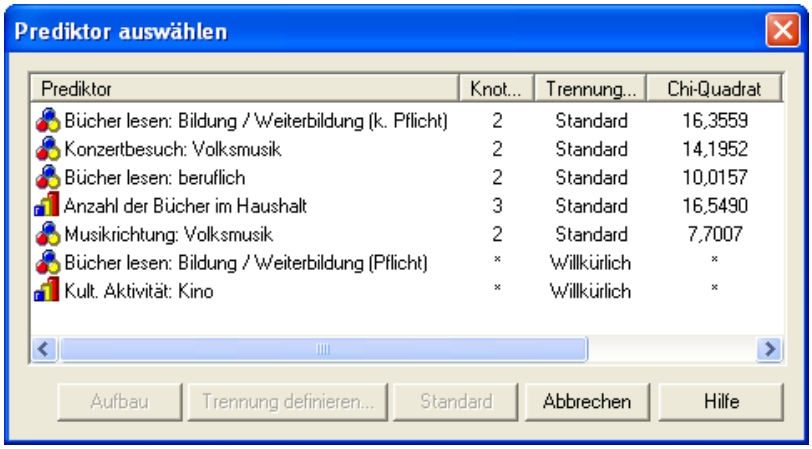
		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	86 - 100 % Nuzeranteil	574	40,4	40,4	40,4
	64 - 85 % Nutzeranteil	306	21,5	21,5	62,0
	34 - 63 % Nutzeranteil	424	29,9	29,9	91,8
	0 - 33 % Nutzeranteil	116	8,2	8,2	100,0
	Gesamt	1420	100,0	100,0	

Die Größe der Gruppen differiert deutlich, die Anteile liegen zwischen 8.2 % (0 - 33 % Nutzeranteil) und 40.4 % (86 - 100 % Nutzeranteil) . Die Grenzen wurden der besseren Verständlichkeit wegen gebildet. Die kleinste Gruppe umfaßt 116 Personen, die größte 574.

4.3 Beschreibung der Gruppen nach Kultur- und Freizeitvariablen

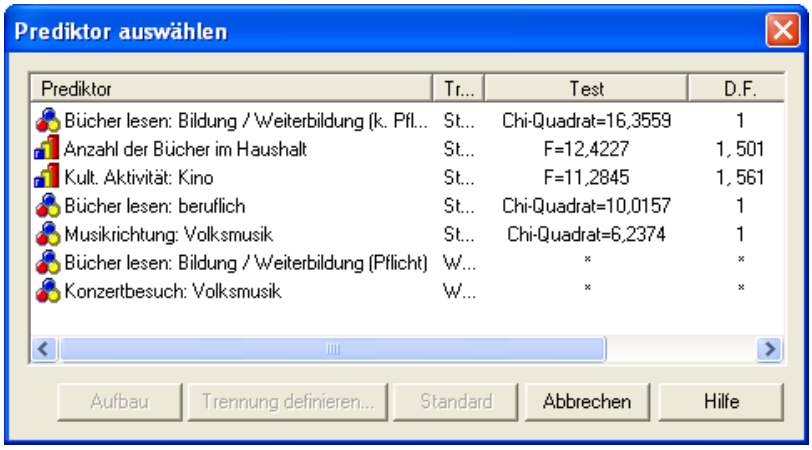
Für die Gruppe mit dem höchsten Nutzeranteilen ergeben sich bei den Kultur- und Freizeitvariablen folgende Prädiktoren:

ABBILDUNG 145 EXHAUSTIVE CHAID: Prädiktoren für Kultur- und Freizeitvariablen (Nutzersegment: 85 - 100 %)



Pflichtweiterbildung und Kinobesuch werden hier willkürlich getrennt, wichtigste Variablen sind die nicht verpflichtende Weiterbildung und der Besuch von Volksmusikkonzerten.

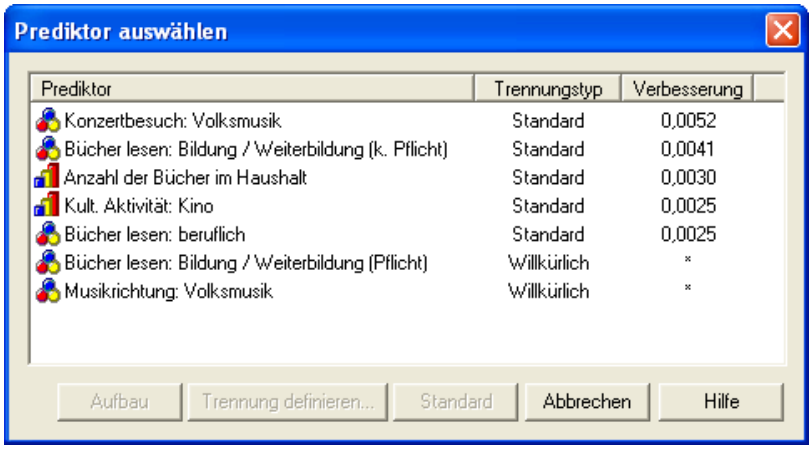
ABBILDUNG 146 QUEST: Prädiktoren für Kultur- und Freizeitvariablen (Nutzersegment: 85 - 100 %)



Leider mußte aufgrund der unzureichenden Exportmöglichkeiten der Prädiktorenansicht das Fenster etwas verkleinert werden: die zweite Spalte („Tr“) enthält die Trennungsinformationen die für die freiwillige Weiterbildung, die Anzahl der Bücher im Haushalt, Kinobesuch, berufliche Lektüre und Volksmusikhören als St(andardtrennung) erfolgt, die restlichen Variablen W(illkürlich), da keine Trennungen gefunden wurden. Auch hier ist die Fehlklassifikation mit 7.8 % unbefriedigend.

Die Beschreibung sehr homogener Gruppen mit einem Zielvariablenverhältnis von rund 90 : 10 ist schwierig, da die Kontrollgruppe der PC-Nichtnutzer recht klein ist. Hier bieten sich zwei Möglichkeiten des weiteren Vorgehens an: eine deskriptive (Kennwerte oder Grafiken) oder multivariate grafische Auswertung.

ABBILDUNG 147 CART: Prädiktoren für Kultur- und Freizeitvariablen (Nutzersegment: 86 - 100 %)



Die Variablen sind in der Reihenfolge der Verbesserung vom höchsten zum niedrigsten Verbesserungswert abgetragen. Dies wird auch aus der Spalte „Trennungstyp" ersichtlich. „Standard" bedeutet beim CART-Algorithmus den voreingestellten Gini-Wert, der sich in der Verbesserung ausdrückt.

Die letzten zwei Variablen weisen keine Verbesserung auf - somit kann auch kein Gini-Wert gebildet werden. Die Trennung erfolgt willkürlich, da es keine statistischen Kennzahlen gibt, die für die Variablen eine aussagekräftige Trennung herbeiführen könnten.

Dies kann ganz unterschiedliche Gründe haben: entweder lassen sich anhand der Variablen keine signifikanten Werte hinsichtlich der Zielvariablen ermitteln - oder sie werden erst sehr weit unten im Baum herangezogen, wo die Gruppen zu klein oder ebenfalls nicht mehr signifikant sind.

Die Verbesserungswerte sind nicht sehr hoch: Pflichtlektüre zur Weiterbildung und das Hören von Volksmusik trennen die Gruppen bei CART nicht, sondern werden willkürlich gesetzt. Die Fehlklassifikation

liegt bei rund 7.8 %, was sehr niedrig erscheint. Allerdings wurden alle Nichtnutzer als Nutzer klassifiziert, was nicht sehr hilfreich ist.

Es ergeben sich folgende Zusammenhangsmaße zwischen den Kultur- und Freizeitvariablen und den vier gefundenen Nutzersegmenten:

TABELLE 41 Nutzersegmente (0 - 33 %, 34 - 65 %, 66 - 84 %, 86 - 100 %): ZUSAMMENHÄNGE MIT DEN RELEVANTEN KULTUR- UND FREIZEITVARIABLEN (CRAMERS V, UNSICHERHEITSKOEFFIZIENT - N = 1420)

	Zusammenhang
Musikrichtung: Volksmusik	v = 0.312, u = 0.077
Bücher lesen: beruflich	v = 0.287, u = 0.089
Bücher lesen: (Weiter-)Bildung Pflicht	v = 0.281, u = 0.082
Bücher lesen: (Weiter-)Bildung keine Pflicht	v = 0.249, u = 0.057
Konzertbesuch: Volksmusik	v = 0.241, u = 0.057
Kulturelle Aktivität: Kinobesuch	v = 0.200, u = 0.047
Anzahl der Bücher im Haushalt	v = 0.151, u = 0.027

Das Hören von Volksmusik ergibt die größten Unterschiede zwischen den Nutzergruppen, die Anzahl der Bücher im Haushalt die geringsten. Die Weiterbildung spielt eine recht wichtige Rolle mit Zusammenhangswerten knapp unter 0.3. Der Volksmusikkonzertbesuch liefert einen relativ geringen Zusammenhangswert, da nur 498 Befragte antworteten.

Nochmals sei darauf verwiesen, dass viele „typische“ horizontale Kultur- und Freizeitvariablen wie Theater- oder Opernbesuch, das Hören von Klassischer Musik oder Operette hier nicht auftauchen. Das mag daran liegen, dass dies keine typische Sozialstrukturuntersuchung ist, sondern eine Betrachtung von PC-Nutzern. Hier könnte in zwei Rich-

tungen argumentiert werden: es könnten den PC-Nutzern bzw. Nichtnutzern - unabhängig von sozialstrukturellen Merkmalen - verschiedene Kulturstile unterstellt werden, die sich vor allem an Volksmusikhören bzw. beruflicher Weiterbildung orientieren: Nichtnutzer hören eher Volksmusik und sind weniger in Weiterbildungsaspekte eingebunden.

Die zweite Richtung, die hier unterstellt wird, verortet PC-Nutzergruppen sozialstrukturell und die Kulturstile als zweitrangige Komponente, die aber durchaus Gruppen typisieren kann.[93]

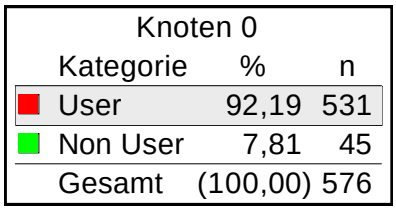
Auf jeden Fall läßt sich sagen, dass im Vergleich mit den klassischen Studien der Lebensstilforschung (z. B. SCHULZEs Erlebnisgesellschaft) sich die Dominanz von Kulturstilen hier in Teilen wiederspiegelt - vor allem Volksmusikhören als deutlich trennende Kategorie für die unterschiedlichen PC-Nutzertypen, die auch SCHULZE als „klassische" Variable herausgehoben hat.

Theoretisch zeigt sich somit eine Bestätigung GEIGERs, aber auch eine gewisse Verifizierung von SCHULZEs Modell, auch wenn der Aspekt des „Erlebnisses" heute durchaus diskussionsfähig ist.

Nachfolgend werden die einzelnen Nutzergruppen charakterisiert. Da der teilweise hohe Anteil an Nutzern bzw. Nichtnutzern eine weitergehende statistische Analyse erschwert, wurde auf das Verfahren der parallelen Boxplots zurückgegriffen. Am Beispiel der Gruppe mit den höchsten Nutzeranteilen soll dies verdeutlicht werden.

93. Natürlich gäbe es eine dritte Möglichkeit, die unterstellt, dass rein sozialstrukturelle Aspekte ausreichen, eine Variable zu untersuchen. Warum soll man aber auf Zusatzinformationen verzichten?

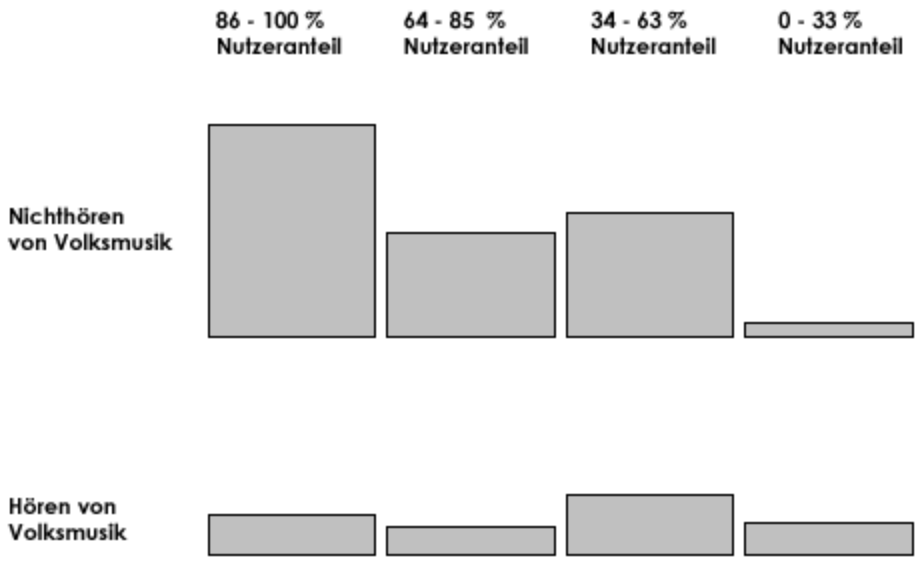
ABBILDUNG 148 Nutzeranteil 86 - 100 %: durchschnittliche Anteile der PC-Nutzer und Nichtnutzer (N = 576), %)



Weitergehende multivariate Untersuchungen der Kultur- und Freizeitvariablen macht es hier wenig Sinn: 45 Nichtnutzer weiter multivariat zu segmentieren, kann nur zu Artefakten führen - so dass deskriptive und grafische Verfahren herangezogen werden müssen.

Die grafische Lösung als Mosaic Plot liefert folgendes Ergebnis für das Hören von Volksmusik:

ABBILDUNG 149 Mosaic Plot: Volksmusikhören nach Nutzersegmenten



Horizontal wurden die Nutzersegmente (links finden sich die höhen Nutzeranteile, rechts die geringen), vertikal die Ausprägungen der Variablen „Volksmusik hören" abgetragen.[94]

Deutlich wird, dass vor allem in der ersten und dritten Gruppe eher die „Ablehner" von Volksmusik enthalten sind. Es scheint also kein kontinuierlicher Verlauf vorzuliegen („je höher der PC-Nutzeranteil, desto geringer die Neigung, Volksmusik zu hören"). Die Kreuztabelle liefert einen prozentualen Aufschluß, der die Grafik relativiert:

94. Bereits weiter oben wurde darauf verwiesen, dass fehlende Werte („missing values") in Mondrian nicht aus einer Analyse ausgeschlossen werden können. Für diese und alle weiteren Analysen wurden die Plots mit einem Bildbearbeitungsprogramm bearbeitet und somit die „unerwünschten" missings aus der Grafik ausgeschnitten.

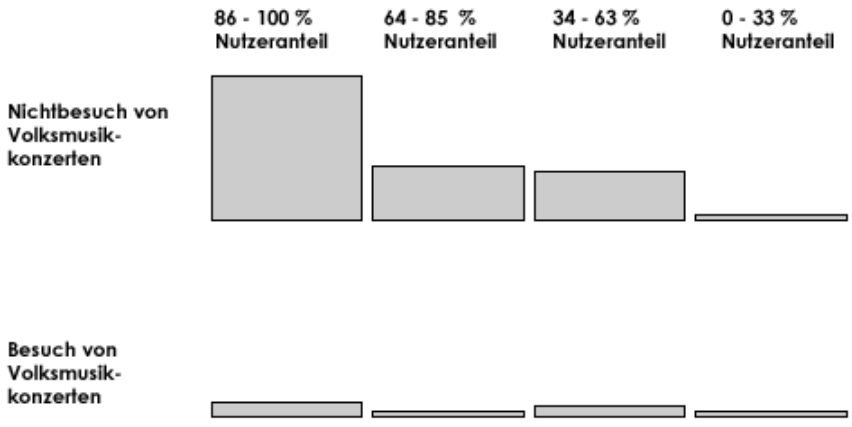
ABBILDUNG 150 Kreuztabelle: Volksmusikhören nach Nutzersegmenten (N = 1382, Spalten-%)

Musikrichtung: Volksmusik * Answertree-Gruppen nach Nutzeranteilen Kreuztabelle

			Answertree-Gruppen nach Nutzeranteilen				
			86 - 100 % Nuzeranteil	64 - 85 % Nutzeranteil	34 - 63 % Nutzeranteil	0 - 33 % Nutzeranteil	Gesamt
Musikrichtung: Volksmusik	-	Anzahl	471	234	276	35	1016
		% von Answertree-Gruppen nach Nutzeranteilen	83,5%	78,3%	67,2%	32,4%	73,5%
	+	Anzahl	93	65	135	73	366
		% von Answertree-Gruppen nach Nutzeranteilen	16,5%	21,7%	32,8%	67,6%	26,5%
Gesamt		Anzahl	564	299	411	108	1382
		% von Answertree-Gruppen nach Nutzeranteilen	100,0%	100,0%	100,0%	100,0%	100,0%

Durch die differierende Gruppengröße kommt es bei der Anzahl der Fälle zum Eindruck, dass in der dritten Gruppe der Anteil der Volksmusikhörer mit N = 135 (= 32.8 %) doppelt so hoch ist wie in Gruppe 2 (N = 65). Allerdings geben die Prozentwerte einen anderen Aufschluß, so dass der Verlauf - prozentual - tatsächlich von der höchsten bis zur geringsten Nutzerkategorie abnimmt.

ABBILDUNG 151 Mosaic Plot: Volksmusikkonzertbesuch nach Nutzersegmenten (N = 498)



Auch bei nachfolgender Grafik sind horizontal die Nutzergruppen, vertikal die Ausprägungen der Variable Volksmusikkonzertbesuch abgetragen. Die geringe Fallzahl ist auf einen Filter in Frage Q 47 zurückzuführen (dort wurde gefragt, ob Konzerte besucht werden. Nur diejenigen Fälle, die die Frage bejaht hatten, wurden in Frage Q 48 berücksichtigt).

ABBILDUNG 152 Kreuztabelle: Volksmusikkonzertbesuch nach Nutzersegmenten (N = 1382, Spalten-%)

Konzertbesuch: Volksmusik * Answertree-Gruppen nach Nutzeranteilen Kreuztabelle

			Answertree-Gruppen nach Nutzeranteilen				
			86 - 100 % Nuzeranteil	64 - 85 % Nutzeranteil	34 - 63 % Nutzeranteil	0 - 33 % Nutzeranteil	Gesamt
Konzertbesuch: Volksmusik	-	Anzahl	243	93	83	11	430
		% von Answertree-Gruppen nach Nutzeranteilen	90,7%	88,6%	79,8%	52,4%	86,3%
	+	Anzahl	25	12	21	10	68
		% von Answertree-Gruppen nach Nutzeranteilen	9,3%	11,4%	20,2%	47,6%	13,7%
Gesamt		Anzahl	268	105	104	21	498
		% von Answertree-Gruppen nach Nutzeranteilen	100,0%	100,0%	100,0%	100,0%	100,0%

68 Befragte besuchen Volksmusikkonzerte (Mehrfachnennungen mit anderen Musikrichtungen war möglich) , wobei im höchsten Nutzersegment absolut die meisten Konzertbesucher zu finden sind (N = 25). Insgesamt liegt der Anteil hier aber unter 10 %: im Gegensatz zur geringsten Nutzergruppe mit knapp 50 % (N = 10'). Hier wird es schwierig, sinnvoll Ergebnisse zu interpretieren, da die Zahlen sehr klein sind. Allerdings ergibt sich ein Cramers v von 0.247 (Sig. = .000) und ein ebenfalls signifikanter Unsicherheitskoeffizient von immerhin 0.057. Auch hier ist somit ein prozentualer kontinuierlicher Anstieg der Konzertbesucher für diese Musikrichtung gegeben - auch wenn es von den Absolutwerten her anders aussieht. Dieser Hinweis wird durch die Grafiken verdeutlicht und wird vielleicht bei der Interpretation der

Kreuztabelle nicht berücksichtigt, wo es sinnvollerweise mehr auf die Prozentsatzdifferenzen ankommt.

4.3.2 Kinobesuch

Die Häufigkeit des Kinobesuchs wurde ordinal (nie, 1 - 2mal, 4 - 6mal [sic!], 7 - 12mal, mehr als 12mal jährlich) erfragt.

ABBILDUNG 153 Kreuztabelle: Kinobesuch nach Nutzersegmenten (N = 1390, Spalten-%)

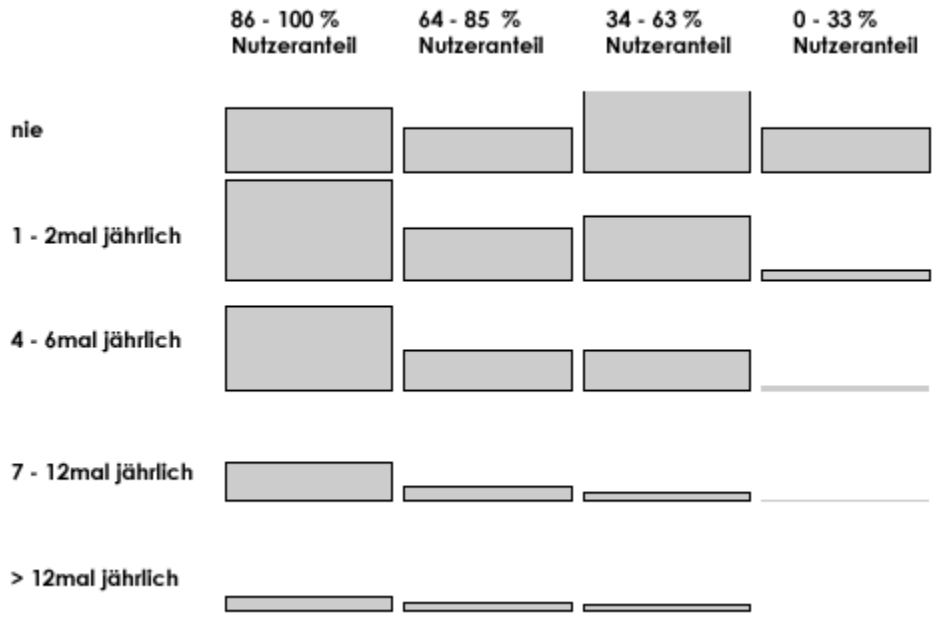
Kult. Aktivität: Kino * Answertree-Gruppen nach Nutzeranteilen Kreuztabelle

			Answertree-Gruppen nach Nutzeranteilen				
			86 - 100 % Nuzeranteil	64 - 85 % Nutzeranteil	34 - 63 % Nutzeranteil	0 - 33 % Nutzeranteil	Gesamt
Kult. Aktivität: Kino	nie	Anzahl	119	82	189	84	474
		% von Answertree-Gruppen nach Nutzeranteilen	21,2%	27,4%	45,4%	73,7%	34,1%
	ein bis zweimal jährlich	Anzahl	187	97	122	19	425
		% von Answertree-Gruppen nach Nutzeranteilen	33,3%	32,4%	29,3%	16,7%	30,6%
	vier bis sechsmal jährlich	Anzahl	157	75	76	9	317
		% von Answertree-Gruppen nach Nutzeranteilen	28,0%	25,1%	18,3%	7,9%	22,8%
	7 - 12mal jährlich	Anzahl	72	29	17	2	120
		% von Answertree-Gruppen nach Nutzeranteilen	12,8%	9,7%	4,1%	1,8%	8,6%
	> 12mal jährlich	Anzahl	26	16	12	0	54
		% von Answertree-Gruppen nach Nutzeranteilen	4,6%	5,4%	2,9%	,0%	3,9%
Gesamt		Anzahl	561	299	416	114	1390
		% von Answertree-Gruppen nach Nutzeranteilen	100,0%	100,0%	100,0%	100,0%	100,0%

Die Kreuztabellle liefert höchstsignifikante Zusammenhangswerte (KENDALLS tau b = -0.261, tau c = -0.248, SPEARMAN = -.0.306) . Dies ist ein geringerer Wert als bei der direkten Betrachtung von PC-Nutzung und Kinobesuch. Somit gibt es einen deutlichen Zusammenhang zwischen den Nutzergruppen und dem Kinobesuch. Das negative Vorzeichen bedeutet, dass geringe Nutzergruppen deutlich weniger häufig ins Kino gehen als hohe Nutzergruppen.

Allerdings darf auch an dieser Stelle nicht vergessen werden, dass sowohl hinter der PC-Nutzervariable als auch hinter dem Kinobesuch ein Alterseffekt wirkt - PC-Nutzer sind eher jünger und Kinobesucher ebenfalls. Somit wird auch an dieser Stelle deutlich, dass sozialstrukturelle Effekte immer im Hintergrund wirken - egal welche Kultur- und Freizeitvariablen man auch heranziehen mag.

ABBILDUNG 154 Mosaic Plot: Kinobesuch nach Nutzersegmenten (N = 1390)



Der Mosaic Plot verdeutlicht grafisch, wie das Ergebnis zustandekommt: deutlich zu sehen ist bei den Nutzersegmenten 3 und 4 die Präferenz, überhaupt nicht ins Kino zu gehen (45 % vs. 73 %). Diese Absolutwerte sind deutlich überrepräsentiert.

Der hohe Korrelationswert kommt vor allem durch die Kategorien 2, 3 und 4 zustande, der in der Gruppe 2 und vor allem im höchsten Nutzersegment deutlich überwiegt.

Der höchste Wert (mehr als 12mal jährlich ins Kino gehen) liegt prozentual im zweiten Segment (5.4 %) höher als im ersten (4.6 %), wobei sich die beiden Nutzergruppen mit der höchsten Nutzung kaum hinsichtlich der Kinopräferenz unterscheiden - im Gegensatz zu den Gruppen mit den niedrigen Nutzeranteilen (2.9 % vs. 0 %). Der (naheliegende) Schluß, hervorgerufen durch den höhen Korrelationswert, dass Gruppen mit hohen Nutzeranteilen auch deutlich häufiger mehr als 12mal jährlich ins Kino gehen (im Gegensatz zu den anderen Gruppen) stimmt so nicht: die durchschnittlichen Anteile (aus der Kreuztabelle ablesbar) in der Kategorie 5 liegt bei 3.9 %, wobei die Gruppen 1 und 2 bei etwa jeweils 5 % liegen. Der Hauptschwerpunkt befindet sich - wie beschrieben - bei den Kategorien 2, 3 und 4.

4.3.3 Anzahl der Bücher im Haushalt

Auch die Anzahl der Bücher wurde ordinal erfaßt, so dass auch ordinale Zusammenhangsmaße herangezogen werden können. Die Kreuztabelle stellt die Kategorien dar:

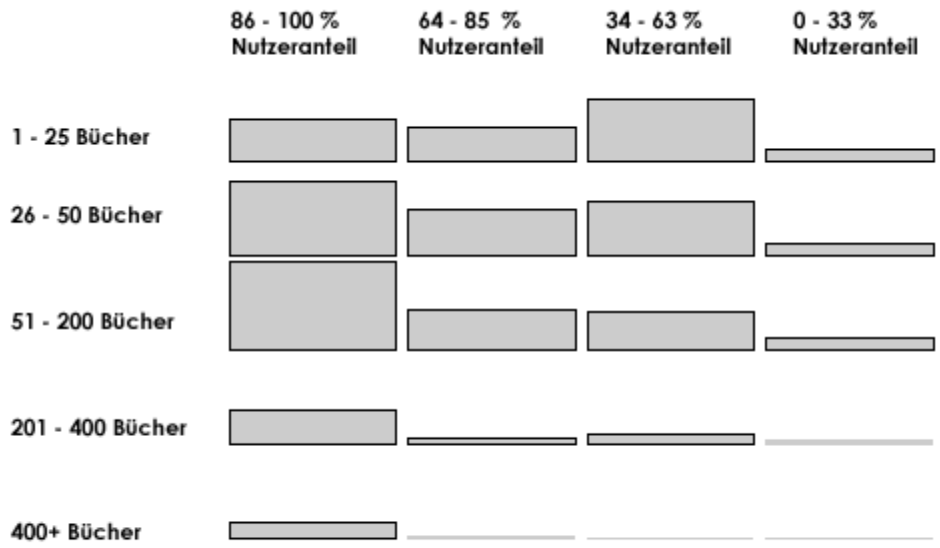
ABBILDUNG 155 Kreuztabelle: Anzahl der Bücher im Haushalt- nach Nutzersegmenten (N = 1176, Spalten-%)

Anzahl der Bücher im Haushalt * Answertree-Gruppen nach Nutzeranteilen Kreuztabelle

			Answertree-Gruppen nach Nutzeranteilen				
			86 - 100 % Nuzeranteil	64 - 85 % Nutzeranteil	34 - 63 % Nutzeranteil	0 - 33 % Nutzeranteil	Gesamt
Anzahl der Bücher im Haushalt	1 - 25 Bücher	Anzahl	84	69	121	25	299
		% von Answertree-Gruppen nach Nutzeranteilen	16,8%	26,5%	37,1%	28,1%	25,4%
	26 - 50 Bücher	Anzahl	145	92	108	26	371
		% von Answertree-Gruppen nach Nutzeranteilen	28,9%	35,4%	33,1%	29,2%	31,5%
	51 - 200 Bücher	Anzahl	171	78	74	27	350
		% von Answertree-Gruppen nach Nutzeranteilen	34,1%	30,0%	22,7%	30,3%	29,8%
	201 - 400 Bücher	Anzahl	68	15	20	9	112
		% von Answertree-Gruppen nach Nutzeranteilen	13,6%	5,8%	6,1%	10,1%	9,5%
	400+ Bücher	Anzahl	33	6	3	2	44
		% von Answertree-Gruppen nach Nutzeranteilen	6,6%	2,3%	,9%	2,2%	3,7%
Gesamt		Anzahl	501	260	326	89	1176
		% von Answertree-Gruppen nach Nutzeranteilen	100,0%	100,0%	100,0%	100,0%	100,0%

Die Differenz zu 1420 Befragten ergibt sich durch die Kategorien 6 („weiss nicht“) und 9 („keine Angabe“).

ABBILDUNG 156 Mosaic Plot: Anzahl der Bücher im Haushalt (ordinal) nach Nutzersegmenten (N = 1176)



Die Korrelation nach KENDALL (tau b = -0.190, tau c = -0.180, SPEARMAN = -0.221) fällt nicht so hoch aus wie beim Kinobesuch. Jedoch kann ganz grundsätzlich davon ausgegangen werden, dass Gruppen mit höheren Nutzeranteilen mehr Bücher besitzen. Interessant ist der höhe Anteil derjenigen, die die Antwort verweigerten bzw. nicht wußten, wieviele Bücher sie besitzen. Diese beiden Kategorien enthalten rund 250 Fälle.

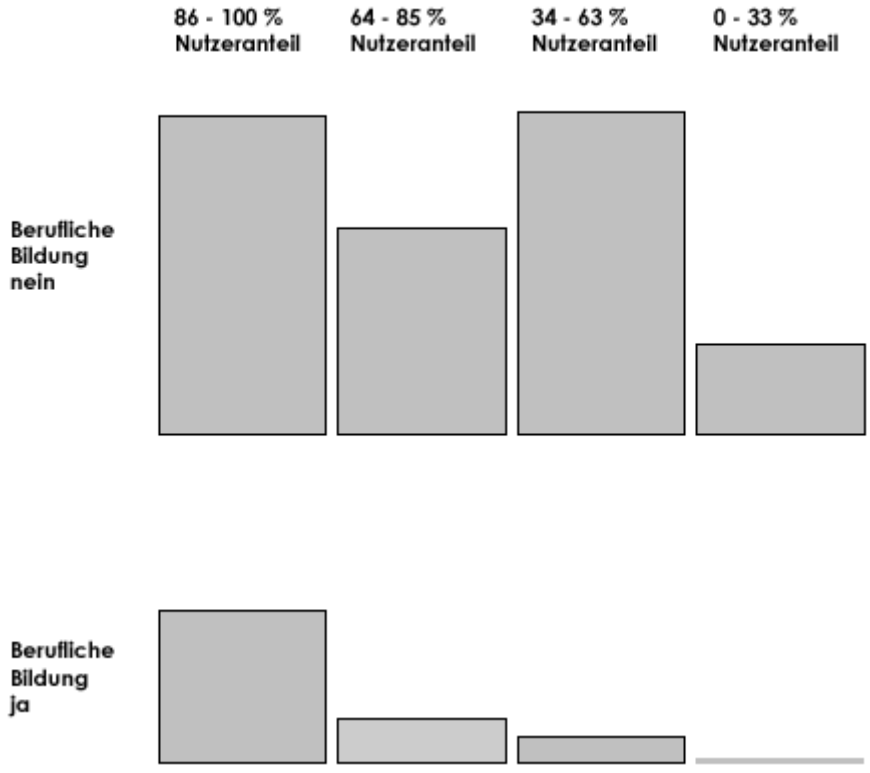
Da über den Inhalt der Bücher nichts bekannt ist (es kann sich sowohl um Spieleanleitungen als auch um Koch- oder Bastelbücher, Krimis, esoterische Literatur oder Belletristik handeln) kann nur auf die Vertrautheit im Umgang mit dem Kulturgut Buch geschlossen werden: das höchste Nutzersegment scheint darin mehr Erfahrung zu besitzen als die anderen Segmente - vor allem weil die Kategorien 4 und 5 (über 250 Bücher) deutlich höher als in den anderen Gruppen liegen. Damit werden Vorurteile (PC-Nutzer besitzen keine Bücher, Ideen der

Papierlosigkeit in Zeiten der elektronischen Datenverarbeitung) entkräftet. Mag es dahingestellt bleiben, ob in den 80er Jahren Menschen, die sich mit Computern in allen Formen beschäftigt haben, (keine) Leser von Büchern waren: heute stimmt diese Pauschalisierung nicht mehr.

4.3.4 Bildung und Weiterbildung

Der Anteil (Anzahl) derjenigen, die sich beruflich weiterbilden müssen, sinkt kontinuierlich von der ersten bis zur vierten Nutzerkategorie - was der Mosaic Plot gut verdeutlicht. Dahinter steht zum einen sicherlich ein Alterseffekt, da viele beim Übergang in den Beruf sich spezifische Kenntnisse aneignen müssen.

ABBILDUNG 157 Mosaic Plot: Berufliche Weiterbildung nach Nutzersegmenten (N = 1420)



Die waagrecht abgetragenen geben die Nutzergruppen, die vertikalen Kategorien die berufliche Bildung an. Im höchsten Nutzerseg-

ment wird berufliche Bildung deutlich häufiger als in den anderen Gruppen angestrebt.

Die Kreuztabelle liefert mit Cramers v = 0.287 (Unsicherheitskoeffizient: 0.089) deutliche Ergebnisse.

ABBILDUNG 158 Kreuztabelle: berufliche Weiterbildung nach Nutzersegmenten (N = 1420, Spalten-%)

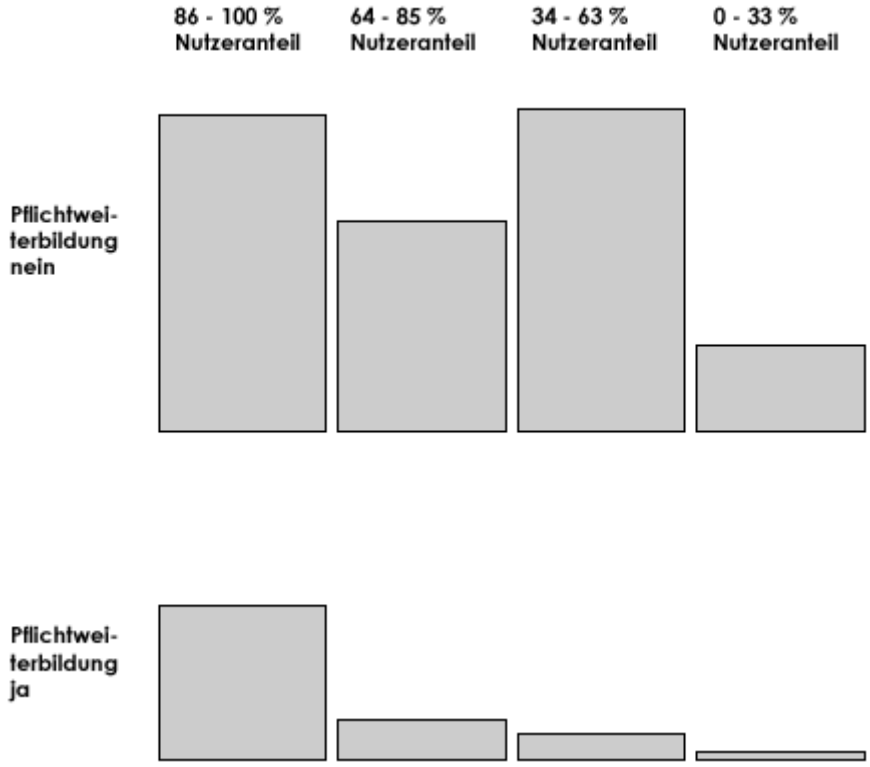
Bücher lesen: beruflich * Answertree-Gruppen nach Nutzeranteilen Kreuztabelle

			Answertree-Gruppen nach Nutzeranteilen				
			86 - 100 % Nuzeranteil	64 - 85 % Nutzeranteil	34 - 63 % Nutzeranteil	0 - 33 % Nutzeranteil	Gesamt
Bücher lesen: beruflich	-	Anzahl	388	252	392	111	1143
		% von Answertree-Gruppen nach Nutzeranteilen	67,6%	82,4%	92,5%	95,7%	80,5%
	+	Anzahl	186	54	32	5	277
		% von Answertree-Gruppen nach Nutzeranteilen	32,4%	17,6%	7,5%	4,3%	19,5%
Gesamt		Anzahl	574	306	424	116	1420
		% von Answertree-Gruppen nach Nutzeranteilen	100,0%	100,0%	100,0%	100,0%	100,0%

Während nahezu 1/3 der höchsten Usergruppe berufliche Bildung betreibt, sind es in der geringsten Kategorie gerade einmal vier Prozent. Die zweite Gruppe mit rund 18 % und die dritte Gruppe mit 7.5 % liegen ebenfalls weit hinter der ersten Gruppe.

Ein ähnliches Bild zeigt auch die Pflichtweiterbildung. Auch hier sinken die Absolutwerte von Segment zu Segment deutlich:

ABBILDUNG 159 Mosaic Plot: Pflichtweiterbildung nach Nutzersegmenten (N = 1420)



Mit 0.281 (Cramers v) bzw. 0.082 (Unsicherheitskoeffizient) wird ein ähnlich hohes Ergebnis wie bei der beruflichen Bildung erreicht

ABBILDUNG 160 Kreuztabelle: Pflichtweiterbildung nach Nutzersegmenten (N = 1420, Spalten-%)

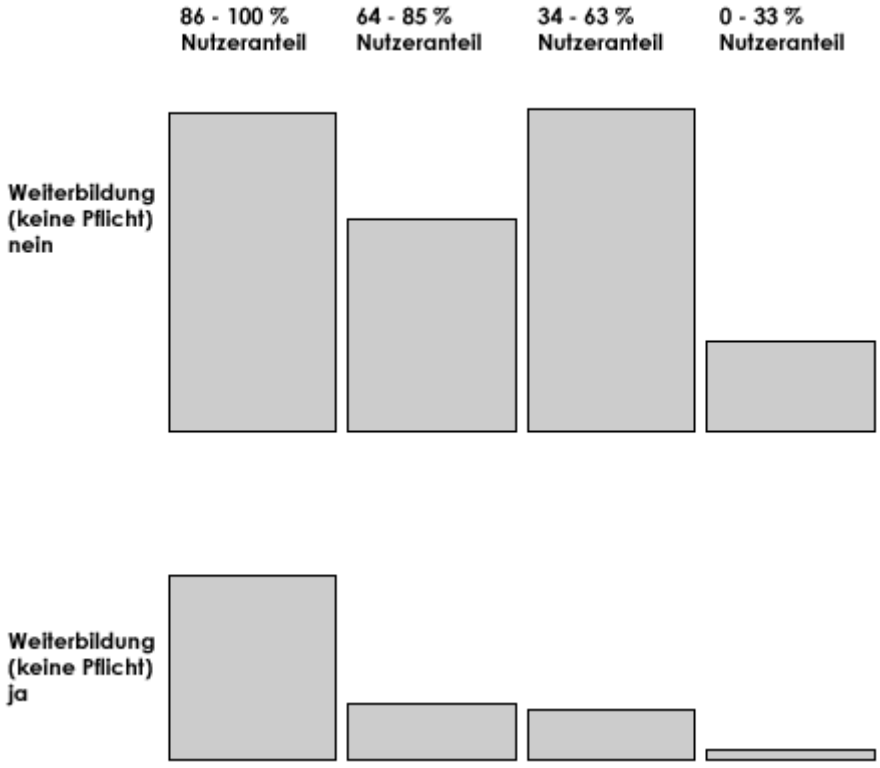
Bücher lesen: Bildung / Weiterbildung (Pflicht) * Answertree-Gruppen nach Nutzeranteilen Kreuztabelle

			Answertree-Gruppen nach Nutzeranteilen				
			86 - 100 % Nuzeranteil	64 - 85 % Nutzeranteil	34 - 63 % Nutzeranteil	0 - 33 % Nutzeranteil	Gesamt
Bücher lesen: Bildung / Weiterbildung (Pflicht)	-	Anzahl	385	256	391	106	1138
		% von Answertree-Gruppen nach Nutzeranteilen	67,1%	83,7%	92,2%	91,4%	80,1%
	+	Anzahl	189	50	33	10	282
		% von Answertree-Gruppen nach Nutzeranteilen	32,9%	16,3%	7,8%	8,6%	19,9%
Gesamt		Anzahl	574	306	424	116	1420
		% von Answertree-Gruppen nach Nutzeranteilen	100,0%	100,0%	100,0%	100,0%	100,0%

Die Ergebnisse unterscheiden sich im Vergleich mit der beruflichen Bildung kaum. Allerdings ist der Anteil im letzten Segment prozentual etwas höher - was bei der geringen Fallzahl kaum ins Gewicht fällt.

Im Bereich der freiwilligen Weiterbildung liegt das höchste Nutzersegment absolut deutlich vor den anderen Gruppen. Auch wenn die Zusammenhangswerte (Cramers v = 0.247, Unsicherheitskoeffizient 0.057) etwas sinken - alle drei Bereiche der Weiterbildung scheinen deutlich mit der PC-Nutzung verknüpft zu sein.

ABBILDUNG 161 Mosaic Plot: Freiwillige Weiterbildung nach Nutzersegmenten (N = 1420)



Das ist kein Wunder: bauen doch heute immer mehr Arbeitsplätze auf dem Umgang mit dem PC. Die Zeiten, in der ein Teil der Personalarbeit in Personalentwicklung und Weiterbildung investiert wurden, sind längst durch Globalisierung und damit verbundenem Kosten-

druck vorbei. Weiterbildung wird häufig nur noch in Form von Schulungen für PC-Programme (Office, SAP, etc.) gewährt - wenn es nicht Voraussetzung für eine Stelle ist.

ABBILDUNG 162 Kreuztabelle: Anzahl der Bücher im Haushalt- nach Nutzersegmenten (N = 1420, Spalten-%)

Bücher lesen: Bildung / Weiterbildung (k. Pflicht) * Answertree-Gruppen nach Nutzeranteilen Kreuztabelle

			Answertree-Gruppen nach Nutzeranteilen				
			86 - 100 % Nuzeranteil	64 - 85 % Nutzeranteil	34 - 63 % Nutzeranteil	0 - 33 % Nutzeranteil	Gesamt
Bücher lesen: Bildung / Weiterbildung (k. Pflicht)	-	Anzahl	363	242	367	103	1075
		% von Answertree-Gruppen nach Nutzeranteilen	63,2%	79,1%	86,6%	88,8%	75,7%
	+	Anzahl	211	64	57	13	345
		% von Answertree-Gruppen nach Nutzeranteilen	36,8%	20,9%	13,4%	11,2%	24,3%
Gesamt		Anzahl	574	306	424	116	1420
		% von Answertree-Gruppen nach Nutzeranteilen	100,0%	100,0%	100,0%	100,0%	100,0%

Alle drei Bereiche verdeutlichen (obwohl sie im Bereich der Kultur- und Freizeitvariablen liegen), wie eng sie mit sozialstrukturellen („alten“) Ungleichheiten verknüpft sind: je geringer die Useranteile für den Rechner, desto weniger „investiert“ ein Unternehmen in Weiterbildung seiner Mitarbeiter. Ebenso sind die Mitarbeiter nicht motiviert. Die Folgen können sich in Arbeitsplatzverlust und Langzeitarbeitslosigkeit (vor allem bei älteren, schlecht qualifizierten Menschen, vielleicht noch ohne PC-Kennnisse) auswirken. Das ist genau die Gruppe, die in diesem Segment besonders vertreten ist:

ABBILDUNG 163 Answertree-Gruppe 0 - 33 % nach Alter (N = 116)

Answertree-Gruppe 0 - 33 % Nutzeranteil * D11: Alter (gruppiert) Kreuztabelle

Anzahl

		D11: Alter (gruppiert)				
		15 - 19 Jahre	20 - 29 Jahre	40 - 49 Jahre	50 - 59 Jahre	Gesamt
Answertree-Gruppe 0 - 33 % Nutzeranteil	4,000	1	3	12	100	116
Gesamt		1	3	12	100	116

100 von 116 Befragten sind 50 Jahre und älter - der Rest sinkt kontinuierlich bis zur jüngsten Alterskategorie. Aus Sicht des Arbeitsmarktes sind diese Menschen bei einem Arbeitsplatzverlust chancenlos - denn auch die erreichten Schulabschlüsse sind überwiegend niedrig:

ABBILDUNG 164 Answertree-Gruppe 0 - 33 % nach Bildungsabschluss (N = 116)

Answertree-Gruppe 0 - 33 % Nutzeranteil * D8: Schulbildung der Befragten Kreuztabelle

			D8: Schulbildung der Befragten				
			bis 15 Jahre (Volks-, Hauptschule)	16 - 17 Jahre (Mittl Reife/erw. HS-Abschluss)	18 - 20 Jahre (Fachgeb.) Hochschulreife	21 Jahre + abgeschlossenes Hochschulstudium	Gesamt
Answertree-Gruppe 0 - 33 % Nutzeranteil	4,000	Anzahl	45	38	26	7	116
		% von Answertree-Gruppe 0 - 33 % Nutzeranteil	38,8%	32,8%	22,4%	6,0%	100,0%
Gesamt		Anzahl	45	38	26	7	116
		% von Answertree-Gruppe 0 - 33 % Nutzeranteil	38,8%	32,8%	22,4%	6,0%	100,0%

Rund 70 % der Befragten haben den Hauptschulabschluß oder die Mittlere Reife erreicht - Abschlüsse, mit denen junge Menschen heute kämpfen müssen, überhaupt einen Ausbildungs- oder Arbeitsplatz zu erhalten.

4.4 Fazit

Deutlich wird, wie obsolet isolierte Betrachtungen sog. „alter" und „neuer" Ungleichheiten sind. Es kommt immer auf die Verknüpfung an, wobei Variablen wie Bildung, Beruf oder Alter eine überragende Position in Untersuchungen über Informationstechnologie zukommt.

Diese Arbeit erhebt nicht den Anspruch, die deutsche Gesellschaft zu schichten oder ein neues Milieumodell zu entwickeln. Es gilt, nur die dominanten Variablen herauszuarbeiten, die für die Nutzung von Informationstechnologie wichtig sind, auch wenn die Arbeit in einigen Jahr(zehnt)en sicherlich veraltet sein wird, da Informationstechnologien weiter spezifiziert werden und immer neue Anforderungen

an die Nutzer stellt. Es ist deutlich geworden, dass Kultur- und Freizeitvariablen wenig zur Erklärungskraft beitragen. Gerade in dieser Untersuchung hat sich gezeigt, dass das Hochkultur- und Trivialschema alleine keinerlei Erklärungskraft besitzt - dominante Schichtungen sind nicht typischerweise Klassikhörer vs. Volksmusikhörern, Theaterbesucher vs. Gartenzwergbesitzer, sondern es sind - mit Ausnahme der Volksmusik - Weiterbildungsaspekte,. Diese lassen sich sehr leicht durch das Hinzulernen von Programmiersprachen, Softwareanwendungen oder anderen Bereichen erläutern, die sich auch im Begriff der „Wissensgesellschaft“ widerspiegeln.

5 Multivariate Analyse III: PC-Nutzung am Beispiel der ordinal und metrisch gemessenen PC-Nutzung

5.1 Deskription der Variablen und Recodierung

Die Variable PC-Nutzung wurde ursprünglich ordinal erfaßt und enthält folgende Ausprägungen:

ABBILDUNG 165 Häufigkeit der PC-Nutzung (ordinal, 6 Kategorien, N = 1420)

Q 39: Häufigkeit: PC-Nutzung

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	nie	504	35,5	35,7	35,7
	seltener als einmal im Monat	35	2,5	2,5	38,1
	ein- bis dreimal im Monat	47	3,3	3,3	41,5
	einmal die Woche	98	6,9	6,9	48,4
	mehrmals die Woche	283	19,9	20,0	68,4
	täglich	446	31,4	31,6	100,0
	Gesamt	1413	99,5	100,0	
Fehlend	weiss nicht	7	,5		
Gesamt		1420	100,0		

Obwohl die Variable wesentlich aussagekräftiger als die dichotomisierte Variante (PC-Nutzung: ja - nein) ist, zeigen sich deutliche Probleme für eine multivariate Analyse vor allem bei den Kategorien mit geringer Fallzahl („seltener als einmal im Monat“, „ein- bis dreimal im Monat“, „einmal die Woche“). Sowohl die Entscheidungsbäume als auch logistische Regresion und Diskriminanzanalyse kommen hier zu keinen oder unbefriedigenden Ergebnissen. Bei der logistischen Regression und der Diskriminanzanalyse kann keine Signifikanz berechnet werden. Da die Gruppen zu klein sind, bei den

Entscheidungsbäumen entstehen zwar große (signifikante) Bäume, diese haben aber eine recht hohe Fehlklassifikation von rund 0.44.

Aus diesem Grund wurden die Kategorien zu Nichtnutzern (Non User), Light User (bis mehrmals wöchentlich) und täglichen Nutzern (Heavy User) zusammengefaßt.

ABBILDUNG 166 Häufigkeit der PC-Nutzung (ordinal, 3 Kategorien, N = 1420)

Q 39: Häufigkeit der PC-Nutzung (Non-, Light-, Heavy) * Q 39: Häufigkeit der PC-Nutzung (Nonuser, User) Kreuztabelle

Anzahl

		Q 39: Häufigkeit der PC-Nutzung (Nonuser, User)		
		Non User	User	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Non-, Light-, Heavy)	Non-User	504	0	504
	Light User (bis mehrm. wöchentl.)	0	463	463
	Heavy User (täglich)	0	446	446
Gesamt		504	909	1413

Von den 1420 Personen haben sieben keine Angaben zur Häufigkeit gemacht, so dass 1413 Befragte in die Analyse eingehen. Die Hauptfragestellung lautet: wie klassifizieren die unterschiedlichen Entscheidungsbaumalgorithmen? Gibt es Unterschiede zur dichotomen Lösung? Welches Verfahren klassifiziert am „besten“ hinsichtlich einer inhaltlichen Aussage? - Da Diskriminanzanalyse und logistische Regression nur als Kontrollinstrumente dienen, werden die Ergebnisse in aller Kürze zusammengefaßt.

5.2 Diskriminanzanalyse

Als unabhängige Variablen dienen - wie im nominalen Fall - Alter, Haushaltsnettoeinkommen und Beruf. Die Diskriminanzanalyse enthält hier - im Gegensatz zum multivariaten Fall - zwei Funktionen:

ABBILDUNG 167 PC-Nutzung (ordinal): Diskriminanzfunktion (N = 1152)

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	,376[a]	95,4	95,4	,523
2	,018[a]	4,6	100,0	,134

a. Die ersten 2 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

Funktion 1 mit einer kanonischen Korrelation von .523 (und einer Varianzaufklärung von rund 95 %) ist deutlich höher und damit wesentlich aussagekräftiger als Funktion 2, die nur 0.134 aufweist. Funktion 2 mit einem Varianzanteil von rund 5 % trägt hier wenig zur Aufklärung bei. Den Zusammenhang zwischen den unabhängigen Variablen verdeutlicht nachfolgende Tabelle:

ABBILDUNG 168 PC-Nutzung ordinal: Zusammenhänge der unabhängigen Variablen (Diskriminanzanalyse)

Gemeinsam Matrizen innerhalb der Gruppen

		D 11 Alter	HH-Nettoeink - dichotome CART-Segmentierung bei 3500 DM	Beruf (dichotome QUEST-Segmentierung)
Korrelation	D 11 Alter	1,000	,157	,047
	HH-Nettoeink - dichotome CART-Segmentierung bei 3500 DM	,157	1,000	,084
	Beruf (dichotome QUEST-Segmentierung)	,047	,084	1,000

Alter und Haushaltsnettoeinkommen (.157) korrelieren deutlich höher als Alter und Beruf (0.047), allerdings ist der Zusammenhang nicht sehr stark.

WILKs LAMBDA ist hochsignifikant, d. h. die beiden gefundenen Funktionen trennen sehr gut, was bedeutet, dass sich auch die darauf beruhenden Gruppen deutlich unterscheiden werden. Der Chi Quadrat-Wert für FUnktion 1 ist mit 387 deutlich höher als für Funktion 2 mit rund 21.

ABBILDUNG 169 PC-Nutzung (ordinal): WILKs Lambda (Diskriminanzanalyse)

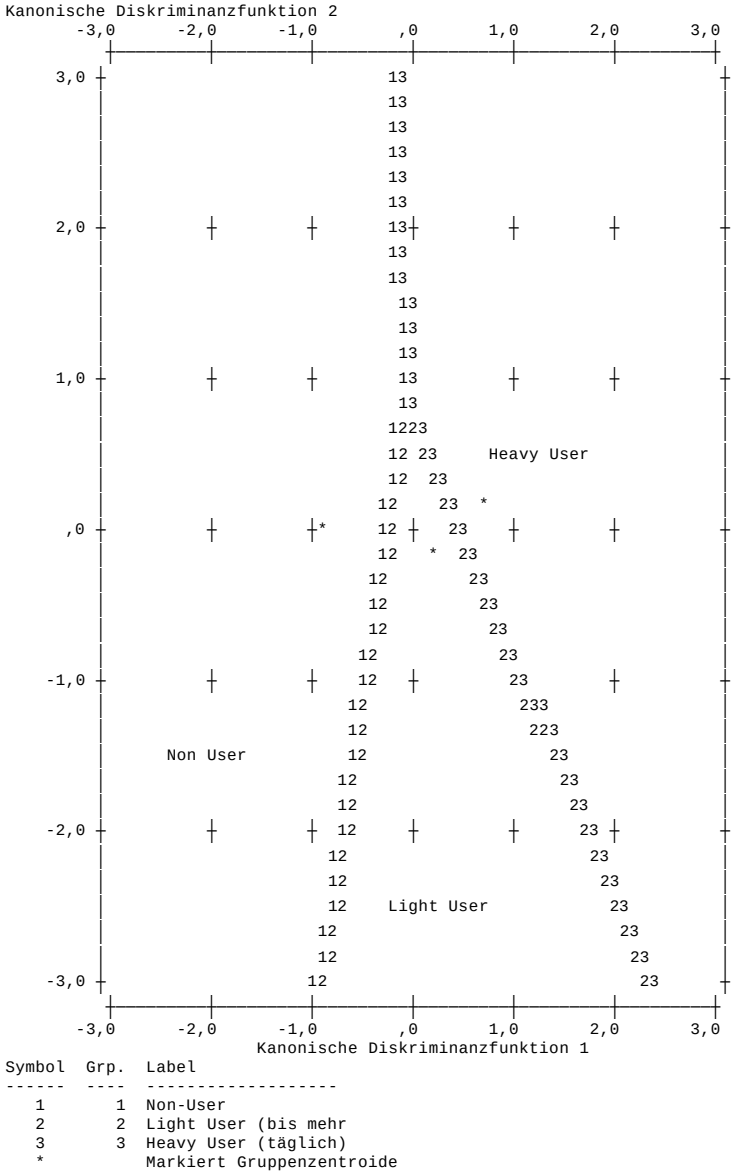
Wilks' Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1 bis 2	,714	387,173	6	,000
2	,982	20,691	2	,000

Die standardisierten Diskriminanzkoeffizienten bilden die Grundlage für die grafische Darstellung, der sog. „Territorial Map", die erst bei zwei Funkionen gebildet werden kann:

Die „Territorial Map" ist ein grafischer Output der Diskriminanzanalyse, der die Gruppen verortet:

ABBILDUNG 170 PC-Nutzung (ordinal): Territorial Map (Diskriminanzanalyse)



1, 2 und 3 bezeichen die jeweiligen Nutzergruppen, das * repräsentiert die Gruppenzentroide, d. h. das „Zentrum“ des jeweiligen Clusters. Die Werte sind z-transformiert, was am Wertebereich der x-Achse (-3 bis +3) erkennbar ist. Die Gruppe der Light User findet sich unten in der Mitte der Grafik innerhalb der Grenzen von „12“ und „23“. Analog hierzu liegen die Non User im linken Teil der Grafik, die Heavy User rechts.

Die Trennung der Gruppen ist deutlich, auch die Gruppenmittelwerte (*) liegen nicht sehr nahe an den Grenzen, was darauf hindeutet, dass die unabhängigen Variablen die PC-Nutzung gut erklären. Der Bereich von -3 bis +3 gibt die z-transformierte Lösung an, also eine mathematische Standardisierung, die die Lösung zwar inhaltlich nicht sehr anschaulich erklärt, dafür die Möglichkeit für Vergleiche zwischen verschiedenen Territorial Maps gewährleistet.

Funktion 1, die eher die Heavy User repräsentiert, ist waagrecht abgetragen und durch höheres Haushaltsnettoeinkommen, geringeres Alter und Büroberufe gekennzeichnet:

Folglich finden sich die Heavy User auch auf der rechten Seite der Grafik, die Nichtnutzer links.

Die Gruppenmittelwerte unterscheiden sich - bezogen auf die z-Transformation deutlich und zeigen, wie die unabhängigen Variablen die ordinal erfaßte PC-Nutzung beeinflussen. Trotzdem ergeben sich - anders als beim nominalen Fall - erheblich schlechtere Klassifikationsergebnisse:

ABBILDUNG 171 PC-Nutzung (ordinal): Fehlklassifikationsmatrix (Diskriminanzanalyse)

Klassifizierungsergebnisse[a]

			Vorhergesagte Gruppenzugehörigkeit			
		Q 39: Häufigkeit der PC-Nutzung (Non-, Light-, Heavy)	Non-User	Light User (bis mehrm. wöchentl.)	Heavy User (täglich)	Gesamt
Original	Anzahl	Non-User	278	17	106	401
		Light User (bis mehrm. wöchentl.)	117	44	217	378
		Heavy User (täglich)	46	32	295	373
		Ungruppierte Fälle	4	1	2	7
	%	Non-User	69,3	4,2	26,4	100,0
		Light User (bis mehrm. wöchentl.)	31,0	11,6	57,4	100,0
		Heavy User (täglich)	12,3	8,6	79,1	100,0
		Ungruppierte Fälle	57,1	14,3	28,6	100,0

a. 53,6% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

Waren es im nominalen Fall etwa 3/4 der Fälle, die richtig klassifiziert wurden, sind es hier rund 54 %, was an der Kategorie der Light User liegt. Je differenzierter eine Variable, desto schwieriger ist auch die Zuordbarkeit - was wieder die bereits diskutierte Frage aufwirft, ob es in jedem Falle sinnvoll ist, mit höheren Skalenniveaus zu arbeiten. Häufig werden (wie in diesem Fall) die Ergebnisse schlechter und die Aussagekraft ist fraglich, denn wie schnell kann jemand vom Light User zum Heavy User werden - oder umgekehrt. Der Unterschied zwischen den Kategorien, den PC „mehrmals die Woche“ oder „täglich“ zu nutzen ist eher trügerisch und könnte von Woche zu Woche differieren. Trotzdem muß sich der Befragte in eine Kategorie einordnen, was zu Artefakten führen kann. Hier scheint die Untersuchung der generellen Nutzung ein eindeutigeres Ergebnis zu erzielen. Obwohl davon auszugehen ist, dass das Ergebnis den anderen Verfahren ähnlich sein wird, ist die Klarheit und Differenziertheit der Gruppen wie bei den Entscheidungsbäumen nicht gegeben.

5.3 Ordinale Logistische Regression

Da die Zielvariable PC-Nutzung nun ordinal skaliert ist, kann auf die ordinale Regression in SPSS zurückgegriffen werden. Auch hier werden Alter, Haushaltsnettoeinkommen und Beruf als unabhängige Variablen herangezogen. Durch die Vielzahl der Kategorien (vor allem beim Alter) kommt es in diesem Fall zu einer Fehlermeldung, die besagt, dass 61.4 % aller Zellen unbesetzt ist. Somit ist das Ergebnis recht ungenau und muß mit großer Vorsicht interpretiert bzw. durch andere Verfahren gestützt oder verworfen werden. Hinzu kommt, dass - wie beim dichotomen Fall - nur 1152 Fälle eingehen, was auf fehlende Angaben beim Haushaltsnettoeinkommen zurückzuführen ist.

ABBILDUNG 172 PC-Nutzung (ordinal): Modellanpassung (Logistische Regression)

Information zur Modellanpassung

Modell	-2 Log-Likelihood	Chi-Quadrat	Freiheitsgrade	Sig.
Nur konstanter Term	2290,057			
Final	1811,542	478,514	22	,000

Verknüpfungsfunktion: Logit.

Ebenso wie beim nominalen Fall wird der Chi-Quadrat-Wert der konstanten und der finalen Lösung angezeigt. Der Chi-Quadrat-Wert ist signifikant - somit haben Alter, Haushaltsnettoeinkommen und Beruf durchaus einen Einfluß auf die ordinal gemessene PC-Nutzung.

Die Varianzaufklärung beträgt 0.382 (Nagelkerke).

ABBILDUNG 173 PC-Nutzung (ordinal): Pseudo R-Quadrat (Logistische Regression)

Pseudo R-Quadrat	
Cox und Snell	,340
Nagelkerke	,382
McFadden	,189

Verknüpfungsfunktion: Logit.

Deutlich wird, dass sich für den vorliegenden Fall die (ordinale) Regression weniger eignet, da sie auf kompletten Kreuztabellen basiert, bei denen wenige Werte fehlen. Wenn - wie in diesem Fall - 61.4 %, also weit über die Hälfte aller Zellen unbesetzt ist, kann dies nur zu Ungenauigkeiten führen - sowohl bei den Chi-Quadrat-Statistiken als auch bei den anderen Kennzahlen. Durch die hohe Zahl an fehlenden Werten zum Haushaltsnettoeinkommen wird eine Umkodierung der Werte kaum zu einer Verbesserung beitragen - vor allem, weil auch das Alter (in Lebensaltersjahren erfaßt) viele unbesetzte Zellen verantwortlich ist.

Die Fehlklassifikation wird in diesem Fall leider nicht von SPSS ausgegeben, es ist eine manuelle Berechnung nötig:

ABBILDUNG 174 PC-Nutzung (ordinal): Tatsächliche vs. vorhergesagte Kategorie (N = 1152)

Q 39: Häufigkeit der PC-Nutzung (Non-, Light-, Heavy) * Vorhergesagte Antwortkategorie Kreuztabelle

Anzahl

		Vorhergesagte Antwortkategorie			
		Non-User	Light User (bis mehrm. wöchentl.)	Heavy User (täglich)	Gesamt
Q 39: Häufigkeit der PC-Nutzung (Non-, Light-, Heavy)	Non-User	311	69	25	405
	Light User (bis mehrm. wöchentl.)	121	116	137	374
	Heavy User (täglich)	51	86	236	373
Gesamt		483	271	398	1152

Aus der Differenz der Randsummen (483 - 405, 374 - 271, 373 - 298 = 206) im Verhältnis zu 1152 Fällen, die in die Analyse eingingen, ergibt sich eine auf den ersten Blick eine erstaunlich geringe Fehlklassifikation von 17.88 %. Zieht man allerdings die nicht klassifizierten Fälle hinzu (N = 1420), ergibt sich

```
(206 fehlklass. Fälle + 268 nichtklass. Fälle) / 1420 = 33.4 %.
```

Somit klassifiziert im ordinalen Fall die Regression besser als die Diskriminanzanalyse. Auch hier ist jedoch die Lösung hinsichtlich differenzierter Einzelsegmente direkt nicht möglich.

5.4 Ordinale Entscheidungsbäume

Die Ergebnisse der Fehlklassifikation liegen zwischen 0.41 (EXHAUSTIVE CHAID), 0.42 (QUEST) und 0.47 (QUEST). Etwa 50 - 60 % werden somit durch die Algorithmen richtig klassifiziert - eine prozentual etwas bessere Lösung als die anderen Verfahren. Ein derart geringer Unterschied sollte allerdings nicht ausschlaggebend für den Einsatz eines Verfahrens sein.[95]

Deutlich wird allerdings auch hier, dass „Mischgruppen“ (in diesem Fall: Light User) von allen Verfahren schlechter klassifiziert werden - **müssen**, da die Merkmale (z. B. jünger, Büroberufe, hohes Einkommen) weniger gut getrennt werden können als im dichotomen Fall.

Ein wesentlicher Vorzug dieses Verfahrens liegt jedoch darin, dass nicht erst unabhängige Variablen dichotomisiert werden müssen, um sie überhaupt untersuchen zu können.

95. Mit hoher Wahrscheinlichkeit unterscheiden sich nominale und ordinale Bäume in diesem Beispiel nicht wesentlich aufgrund der kleinen Fallzahl voneinander.

5.4.1 EXHAUSTIVE CHAID-Algorithmus

(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)

Der Hauptknoten für alle drei Verfahren zeigt folgende Verteilung:

ABBILDUNG 175 PC-Nutzung (ordinal): Wurzelknoten (CART-Algorithmus, N = 1413)

Knoten 0

Kategorie	%	n
Non-User	35,67	504
Light User (bis mehrm. w öchentl.)	32,77	463
Heavy User (täglich)	31,56	446
Gesamt	(100,00)	1413

Die Kategorien sind etwa gleich groß, so dass eine Segmentierung einfacher ist als mit unterschiedlich großen Gruppen. Dies wird auch durch den ursprünglichen Wahrscheinlichkeitswert der Fehlklassifikation ausgedrückt, der bei rund 0.66 liegt. Während bei zwei Gruppen die Wahrscheinlichkeit = 0.5 beträgt, sind es bei 3 Gruppen rund 66 %. Anders ausgedrückt: Wenn ich einen PC-Nutzer zufällig zuordnen möchte, dann ist die Wahrscheinlichkeit ihn auch tatsächlich als PC-Nutzer zu kategorisieren, 1/3. Klassifiziere ich ihn falsch, liegt die Wahrscheinlichkeit bei 2/3, denn entweder ordne ich die Person den Light Usern oder den Non Usern zu.

Der Baum gestaltet sich - häufig wie bei EXHAUSTIVE CHAID - unübersichtlich - deshalb wieder die Tabellendarstellung:

TABELLE 42 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (EXHAUSTIVE CHAID, EINSTUFIG, N = 1413)

Segmente

Reise- und Dienstleistungsangestellte, Landwirte, Fischer

Knoten 1		
Kategorie	%	n
Non-User	40,93	88
Light User (bis mehrm. w öchentl.)	36,74	79
Heavy User (täglich)	22,33	48
Gesamt	(15,22)	215

Grossunternehmer, Direktoren, Top Management, Leitende Angestellte

Knoten 2		
Kategorie	%	n
Non-User	11,11	7
Light User (bis mehrm. w öchentl.)	22,22	14
Heavy User (täglich)	66,67	42
Gesamt	(4,46)	63

sonstige Bürotätigkeiten

Knoten 3		
Kategorie	%	n
Non-User	10,78	18
Light User (bis mehrm. w öchentl.)	31,74	53
Heavy User (täglich)	57,49	96
Gesamt	(11,82)	167

(sonstige Fach)Arbeiter

Knoten 4		
Kategorie	%	n
Non-User	57,75	257
Light User (bis mehrm. w öchentl.)	28,54	127
Heavy User (täglich)	13,71	61
Gesamt	(31,49)	445

Büroangestellte mit Leitungsfunktion

Knoten 5		
Kategorie	%	n
Non-User	10,16	19
Light User (bis mehrm. w öchentl.)	32,62	61
Heavy User (täglich)	57,22	107
Gesamt	(13,23)	187

Studierende

Knoten 6		
Kategorie	%	n
Non-User	5,47	7
Light User (bis mehrm. w öchentl.)	58,59	75
Heavy User (täglich)	35,94	46
Gesamt	(9,06)	128

Ladenbesitzer, Handwerker, Meister

Knoten 7		
Kategorie	%	n
Non-User	30,77	16
Light User (bis mehrm. w öchentl.)	42,31	22
Heavy User (täglich)	26,92	14
Gesamt	(3,68)	52

nie erwerbstätig

Knoten 8		
Kategorie	%	n
Non-User	72,03	85
Light User (bis mehrm. w öchentl.)	24,58	29
Heavy User (täglich)	3,39	4
Gesamt	(8,35)	118

Freie Berufe

Knoten 9		
Kategorie	%	n
Non-User	7,89	3
Light User (bis mehrm. w öchentl.)	18,42	7
Heavy User (täglich)	73,68	28
Gesamt	(2,69)	38

Die Lösung ist nicht sehr effizient, macht aber deutlich, dass Light User interessanterweise in hohem Maße bei den Studierenden (Anteil: 51

%) zu finden ist - die Gruppe, die mit die höchste Nutzeranteile besitzt. Auch Ladenbesitzer, Handwerker und Meister haben einen recht hohen Light User-Anteil (42 %). Hier könnte die überwiegend berufliche Nutzung eine Rolle spielen. Freie Berufe bzw. Top Manager, Direktoren und Großunternehmer haben mit rund 20 % den geringsten Light-User-Anteil - zum Teil ist die Erreichbarkeit für Anfragen am Wochenende bei Freiberuflern oder Selbständigen sicherlich wichtig. Insgesamt liegen die Anteile für die Büroberufe deutlich höher als im Durchschnitt.

TABELLE 43 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (EXHAUSTIVE CHAID, REISE- UND DIENSTLEISTUNGSBERUFE, LANDWIRTE FISCHER, N = 215)

Hauptknoten / Segmentbeschreibung	Unterknoten
Knoten 1 Kategorie / % / n Non-User 40,93 88 Light User (bis mehrm. w öchentl.) 36,74 79 Heavy User (täglich) 22,33 48 Gesamt (15,22) 215	
<= 2250 DM	Knoten 10 Kategorie / % / n Non-User 42,42 14 Light User (bis mehrm. w öchentl.) 27,27 9 Heavy User (täglich) 30,30 10 Gesamt (2,34) 33
2001 - 2250 DM; 4001 - 4500 DM; fehlende Werte	Knoten 11 Kategorie / % / n Non-User 51,24 62 Light User (bis mehrm. w öchentl.) 33,06 40 Heavy User (täglich) 15,70 19 Gesamt (8,56) 121
> 4500 DM	Knoten 12 Kategorie / % / n Non-User 19,67 12 Light User (bis mehrm. w öchentl.) 49,18 30 Heavy User (täglich) 31,15 19 Gesamt (4,32) 61

Zwei Berufssegmente werden weiter differenziert - die Reise- und Dienstleistungsberufe, Landwirte, Fischer und die Arbeiter - ähnlich wie im nominalen Fall.

Da die Knoten von oben und von links nach rechts durchnumeriert werden und die Segmentierung des Haushaltsnettoeinkommens erst auf der zweiten Stufe erfolgt, werden die Knoten mit den Nummern 10 bis 12 etikettiert.

Für diese Berufssegmente zeigen sich interessante Ergebnisse: das Haushaltsnettoeinkommen bestimmt vor allem in Knoten 12 die PC-Nutzung (rund 80 %) im Gegensatz zu den niedrigen Einkommensgruppen - ein Ergebnis, dass auch schon durch die nominale Segmentierung mit CART ersichtlich wurde. Bei den niedrigen Einkommensgruppen dominiert klar die Nichtnutzung.

TABELLE 44 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (EXHAUSTIVE CHAID, ARBEITER, N = 445)

Hauptknoten / Segmentbeschreibung	Unterknoten
Knoten 4 Kategorie % n Non-User 57,75 257 Light User (bis mehrm. wöchentl.) 28,54 127 Heavy User (täglich) 13,71 61 Gesamt (31,49) 445	
< 33 Jahre	Knoten 13 Kategorie % n Non-User 43,37 72 Light User (bis mehrm. wöchentl.) 39,16 65 Heavy User (täglich) 17,47 29 Gesamt (11,75) 166
33 - 48 Jahre	Knoten 14 Kategorie % n Non-User 58,70 108 Light User (bis mehrm. wöchentl.) 27,72 51 Heavy User (täglich) 13,59 25 Gesamt (13,02) 184
> 48 Jahre	Knoten 15 Kategorie % n Non-User 81,05 77 Light User (bis mehrm. wöchentl.) 11,58 11 Heavy User (täglich) 7,37 7 Gesamt (6,72) 95

Von Knoten 13 bis Knoten 15 steigt der Nichtnutzeranteil kontinuierlich von 43 % bis 81 % an.

Charakteristisch für die beiden Berufssegmente ist, dass sie - was die PC-Nutzung angeht - sehr inhomogen sind. Es gibt Teile, die den Rechner sehr intensiv nutzen, andere nicht. Auch hier der Hinweis, dass weder die Diskriminanzanalyse noch die logistische Regression Ähnliches leisten kann.

TABELLE 45 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (EXHAUSTIVE CHAID, ARBEITER BIS 33 JAHRE, N = 166)

Hauptknoten / Segmentbeschreibung	Unterknoten
Knoten 13 Kategorie / % / n Non-User 43,37 72 Light User (bis mehrm. w öchentl.) 39,16 65 Heavy User (täglich) 17,47 29 Gesamt (11,75) 166	
<= 2500 DM	Knoten 16 Kategorie / % / n Non-User 45,45 25 Light User (bis mehrm. w öchentl.) 41,82 23 Heavy User (täglich) 12,73 7 Gesamt (3,89) 55
2251 - 2500 DM; 3001 - 3500 DM; fehlt	Knoten 17 Kategorie / % / n Non-User 58,73 37 Light User (bis mehrm. w öchentl.) 31,75 20 Heavy User (täglich) 9,52 6 Gesamt (4,46) 63
> 3000 DM	Knoten 18 Kategorie / % / n Non-User 20,83 10 Light User (bis mehrm. w öchentl.) 45,83 22 Heavy User (täglich) 33,33 16 Gesamt (3,40) 48

Die Gruppen sind bereits recht klein - und das Ergebnis ist deshalb auch nicht sehr erhellend: aber auch in diesen Segmenten ist ein gewisser Einkommenseffekt sichtbar.

Die Fehlklassifikation liegt bei 0.41. Die Prädiktoren haben folgenden Einfluß auf die Zielvariable:

ABBILDUNG 176 PC-Nutzung (ordinal): Prädiktorwerte (EXHAUSTIVE CHAID-Algorithmus, N = 1413)

Der Chi-Quadrat-Wert des Berufs liegt mit 457 deutlich über dem des Haushaltsnettoeinkommens (rund 99) und des Altes (38).

Der eingesetzte Likelihood-Verhältnis-Test basiert auf der Chi-Quadrat-Statistik - folglich werden auch Chi-Quadrat-Werte ausgegeben. Der Output unterscheidet sich hier somit nicht vom nominalen Fall. Allerdings ist die Logik des Likelihood-Tests etwas anders.

Die Maximum-Likelihood-Funktion für den ordinalen Fall vergleicht zwei Modelle - z. B. ein Ausgangsmodell, bei der alle bis auf eine unabhängige Variable auf 0 gesetzt wird, um zu sehen, wie die eine, unabhängige die abhängige Variable „erklärt". Zum Vergleich wird dann das Gesamtmodell herangezogen. Aus dieser Differenz resultieren die Chi-Quadrat-Werte.

URBAN (1993: 53) bemerkt:

> „Die ML-Schätzung (Maximum Likelihood, Anm. S. L.) sucht nicht nach der kleinsten quadrierten Residuensumme, sondern sie wählt im Zuge einer schrittweisen Annäherung diejenigen Koeffizienten als optimale Schätzwerte aus, die unter der Annahme, sie wären identisch mit den wahren Parametern in der Grundgesamtheit, die beobachteten Stichprobenwerte mit der größten Wahrscheinlichkeit hervorbringen würden."

Einfacher erklärt bedeutet das, dass die Chi-Quadrat-Statistik durch den Vergleich der erwarteten und beobachteten Häufigkeiten zu ihren Ergebnissen kommt, die Maximum Likelihood-Schätzung jedoch dies unberücksichtigt läßt. Sie geht davon aus, dass das Modell der abhängigen und der unabhängigen Variablen bereits optimal ist („mit der Grundgesamtheit identisch ist"). Grundfrage ist somit, „... welche zugrundeliegenden Parameter die beobachteten Daten mit der größten Wahrscheinlichkeit hervorgebracht haben könnte" (URBAN (1993: 55).

5.4.2 QUEST-Algorithmus

QUEST	• vieldiskutierter, relativ neuer Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus (unabhängige Variablen) geeignet • nur für nominale Zielvariablen geeignet, die auch dichotom sein kann • gebräuchliche statistische Kennzahlen (Chi-Quadrat, F-Test) • erlaubt Ersatzprädiktoren, Pruning und a prioris

Der QUEST-Algorithmus klassifiziert in diesem Fall mit rund 0.47 am schlechtesten. Dies liegt daran, dass QUEST weder ordinale noch metrische, sondern nur nominale (streng genommen: dichotome) Zielvariablen zuläßt. Die Zielvariable konnte somit nicht als ordinale Variable definiert werden, sondern nur als Nominalvariable. Trotzdem soll das Beispiel - obwohl es nicht sehr überzeugend ist - in Ansätzen besprochen werden. Die Lösung wird dadurch erschwert, dass die Ordinalität (von Non User über Light User zu Heavy User) natürlich durch die fehlende Ordinalität nicht eingehalten wird und QUEST mit folgendem Knoten startet:

ABBILDUNG 177 PC-Nutzung (ordinal): Wurzelknoten (QUEST-Algorithmus, N = 1413)

Knoten 0		
Kategorie	%	n
Non-User	35,39	500
Heavy User (täglich)	31,56	446
Light User (bis mehrm. w öchentl.)	33,05	467
Gesamt	(100,00)	1413

Die Prädiktoren liefern folgende Werte:

ABBILDUNG 178 PC-Nutzung (ordinal): Prädiktorwerte (QUEST-Algorithmus, N = 1413)

Hier fällt sofort die Verwendung des F-Werts bei der ordinalen und metrischen Variablen auf[96], die alle in die Analyse eingehen (Trennungstyp = Standard, somit für nominale Variablen Chi-Quadrat). Im Gegensatz zu CART und EXHAUSTIVE CHAID ist die Reihenfolge der Prädiktoren vertauscht: das Alter steht hier an erster, bei den beiden anderen Algorithmen an letzter Stelle, der Beruf, hier an dritter Stelle, wird von den beiden anderen Verfahren als wichtigstes Kriterium herangezogen.

96. Sowohl bei BALTES-GÖTZ (2003: 57) als auch im SPSS-Answertree 3.0-Handbuch von SPSS (2001: 230) wird darauf hingewiesen, dass auch bei ordinalen unabhängigen Variablen der F-Wert der einfaktoriellen Varianzanalyse herangezogen wird.

Die Alterssegmentierung bei rund 54 Jahren ist zentral für diesen Algorithmus:

TABELLE 46 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (QUEST, N = 1413)

Hauptknoten / Segmentbeschreibung	Unterknoten
Knoten 0 Kategorie % n Non-User 35,39 500 Heavy User (täglich) 31,56 446 Light User (bis mehrm. w öchentl.) 33,05 467 Gesamt (100,00) 1413	
bis 54 Jahre	Knoten 1 Kategorie % n Non-User 33,59 434 Heavy User (täglich) 32,20 416 Light User (bis mehrm. w öchentl.) 34,21 442 Gesamt (91,44) 1292
55 Jahre +	Knoten 2 Kategorie % n Non-User 54,55 66 Heavy User (täglich) 24,79 30 Light User (bis mehrm. w öchentl.) 20,66 25 Gesamt (8,56) 121

Auch im ordinalen Fall segmentiert QUEST anders als die anderen Algorithmen - mit der Tendenz, „unliebsame“ Anteile hinsichtlich der Zielvariablen (Knoten 2, N = 121 oder 8.56 %) „auszusegmentieren“ - hier die Älteren. Die Lösung ist jedoch für beide Knoten nicht sehr effizient: im Knoten 1 ergibt sich ein Verhältnis von jeweils etwa 1/3, im Knoten 2 betragen die Anteile rund 55 : 25 : 20.

TABELLE 47 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (QUEST, BEFRAGTE BIS 54 JAHRE BZW. AB 55 JAHRE, N = 1413)

Hauptknoten / Segmentbeschreibung	Unterknoten
bis 54 Jahre Knoten 1 Kategorie / % / n Non-User 33,59 434 Heavy User (täglich) 32,20 416 Light User (bis mehrm. w öchentl.) 34,21 442 Gesamt (91,44) 1292	
<= 1500 DM	Knoten 3 Kategorie / % / n Non-User 40,23 35 Heavy User (täglich) 24,14 21 Light User (bis mehrm. w öchentl.) 35,63 31 Gesamt (6,16) 87
> 1500 DM	Knoten 4 Kategorie / % / n Non-User 33,11 399 Heavy User (täglich) 32,78 395 Light User (bis mehrm. w öchentl.) 34,11 411 Gesamt (85,28) 1205
55 Jahre + Knoten 2 Kategorie / % / n Non-User 54,55 66 Heavy User (täglich) 24,79 30 Light User (bis mehrm. w öchentl.) 20,66 25 Gesamt (8,56) 121	
Reise- und Dienstleistungsangestellte, (Fach-)Arbeiter, Ladenbesitzer, Handwerker, nie erwerbstätig, Landwirte, Fischer	Knoten 5 Kategorie / % / n Non-User 72,22 52 Heavy User (täglich) 11,11 8 Light User (bis mehrm. w öchentl.) 16,67 12 Gesamt (5,10) 72
Großunternehmer, Direktoren, Top Management, Leitende Angestellte, sonstige Bürotätigkeiten, freie Berufe, Meister, Büroangestellte mit Leitungsfunktion	Knoten 6 Kategorie / % / n Non-User 28,57 14 Heavy User (täglich) 44,90 22 Light User (bis mehrm. w öchentl.) 26,53 13 Gesamt (3,47) 49

Während die Einkommenssegmentierung der Jüngeren weniger deutliche Ergebnisse erbringt, wird zumindest bei den Berufen, die dem PC nicht sehr nahe stehen ein Anteil von über 70 % bei den Non-Usern heraussegmentiert (Knoten 5). Insgesamt ist diese Lösung allerdings nicht allzu überzeugend.

Insgesamt segmentiert QUEST 18 Knoten, wobei die Gruppe der Älteren - aufgrund der geringen Fallzahl - nur noch einmal segmentiert wird: nach dem Beruf. Hier finden sich einerseits die sonstigen Reise- und Dienstleistungsberufe, Arbeiter und nie Erwerbstätigen, andererseits die Gruppe der Studierenden (aufgrund der Einkommenssegmentierung von 1500 DM!), der sonstigen Bürotätigkeiten - und auch der Tätigkeiten mit Leitungsfunktion, wahrscheinlich Kleinunternehmer mit wenigen Angestellten Allerdings sind die Gruppen hier bereits relativ klein.

Die Segmente derjenigen, die über 1500 DM Haushaltsnettoeinkommen beziehen, ist nicht sehr aussagekräftig: die Trennung erfolgt bei 2250 DM und zeigt auch hier wieder ein Ungleichgewicht: auf der einen Seite stehen diejenigen, die zwischen 1500 DM und 2250 DM beziehen, auf der anderen Seite all diejenigen, die darüber Einkommen beziehen (117 : 1088 Befragten). Natürlich ist es auch möglich, durch die „Negativsegmente“ eine Verteilung zu beschreiben - es ist aber für die Interpretation nicht gerade einfach.

In der Logik von QUEST und dieses Entscheidungsbaums folgt daraus, dass die höhere Einkommensgruppe von 2250 DM und mehr wieder in ein kleines, eher Nichtnutzersegment und ein größeres Nutzersegment aufgeteilt wird: die Facharbeiter und nie Erwerbstätigen finden sich in Knoten 17 der nachfolgenden Abbildung, die restlichen Berufsgruppen in Knoten 18. Der Weg führt zu diesen Knoten über drei (!) Haushaltssegmentierungen:

ABBILDUNG 179 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (QUEST, BEFRAGTE BIS 54 JAHRE, HAUSHALTSNETTOEINKOMMEN > 2750 DM, N = 1413)

Knoten 17

Kategorie	%	n
Non-User	56,86	203
Heavy User (täglich)	13,73	49
Light User (bis mehrm. w öchentl.)	29,41	105
Gesamt	(25,27)	357

Knoten 18

Kategorie	%	n
Non-User	14,87	91
Heavy User (täglich)	47,55	291
Light User (bis mehrm. w öchentl.)	37,58	230
Gesamt	(43,31)	612

Auch hier fällt das „Ungleichgewicht“ von N = 357 (Knoten 17) zu N = 612 (Knoten 18) auf. Die Unterschiede (Non User: 57 : 14, Heavy User: 14 : 48, Light User: 30 : 38) sind aber deutlich.

5.4.3 CART-Algorithmus

CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und a prioris

CART kommt zu der gleichen Reihenfolge der Prädiktoren wie EXHAUSTIVE CHAID - Beruf, Einkommen und Alter.

TABELLE 48 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (CART, N = 1413)

Hauptknoten / Segmentbeschreibung	Unterknoten
Knoten 1 Kategorie % n Non-User 9,72 59 Light User (bis mehrm. w öchentl.) 36,57 222 Heavy User (täglich) 53,71 326 Gesamt (42,96) 607	KNOTEN 1: BERUFSSEGMENTIERUNG Grossunternehmer, Direktoren, Top Management, Leitende Angestellte, sonstige Bürotätigkeiten, Bürotätigkeiten mit Leitungsfunktion, Studierende, Freie Berufe, Meister
Grossunternehmer, Direktoren, Top Management, Leitende Angestellte, sonstige Bürotätigkeiten, Bürotätigkeiten mit Leitungsfunktion, Studierende, Freie Berufe, Meister	Knoten 3 Kategorie % n Non-User 10,33 47 Light User (bis mehrm. w öchentl.) 29,67 135 Heavy User (täglich) 60,00 273 Gesamt (32,20) 455
Studierende, Meister	Knoten 4 Kategorie % n Non-User 7,89 12 Light User (bis mehrm. w öchentl.) 57,24 87 Heavy User (täglich) 34,87 53 Gesamt (10,76) 152
Knoten 2 Kategorie % n Non-User 54,71 441 Light User (bis mehrm. w öchentl.) 30,40 245 Heavy User (täglich) 14,89 120 Gesamt (57,04) 806	KNOTEN 2: BERUFSSEGMENTIERUNG Reise- und Dienstleistungsangestellte, Arbeiter, Ladenbesitzer, Handwerker, nie erwerbstätig, Landwirte Fischer
<= 46.5 Jahre	Knoten 5 Kategorie % n Non-User 48,41 289 Light User (bis mehrm. w öchentl.) 34,84 208 Heavy User (täglich) 16,75 100 Gesamt (42,25) 597
> 46.5 Jahre	Knoten 6 Kategorie % n Non-User 72,73 152 Light User (bis mehrm. w öchentl.) 17,70 37 Heavy User (täglich) 9,57 20 Gesamt (14,79) 209

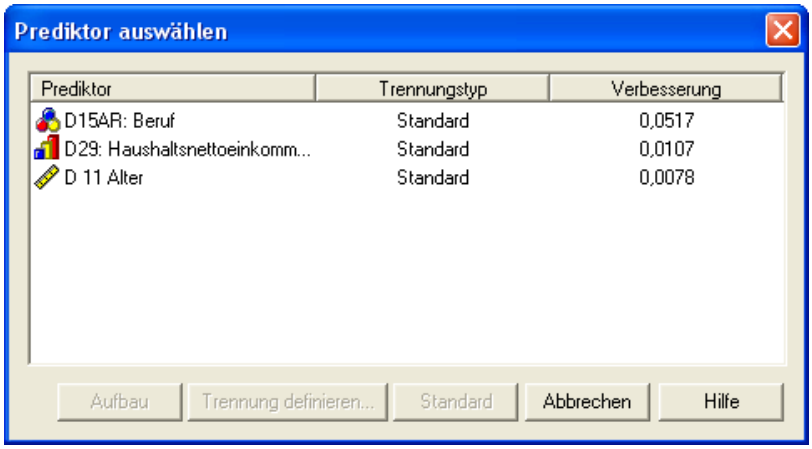
Die jüngeren (bis 47 Jahre alten) Befragten werden nochmal nach Haushaltsnettoeinkommen aufgesplittet:

TABELLE 49 ORDINAL SKALIERTE PC-NUTZUNG NACH ALTER, HAUSHALTSNETTOEINKOMMEN UND BERUF (CART, N = 597)

Hauptknoten / Segmentbeschreibung	Unterknoten
Knoten 5 Kategorie % n Non-User 48,41 289 Light User (bis mehrm. w öchentl.) 34,84 208 Heavy User (täglich) 16,75 100 Gesamt (42,25) 597	
<= 4500 DM	Knoten 11 Kategorie % n Non-User 53,68 255 Light User (bis mehrm. w öchentl.) 32,63 155 Heavy User (täglich) 13,68 65 Gesamt (33,62) 475
> 4500 DM	Knoten 12 Kategorie % n Non-User 27,87 34 Light User (bis mehrm. w öchentl.) 43,44 53 Heavy User (täglich) 28,69 35 Gesamt (8,63) 122

Das Haushaltsnettoeinkommen hat auch hier - vor allem in höheren Einkommensklassen - einen deutlichen Einfluss: rund 70 % Nutzer stehen bei höheren finanziellen Ressourcen rund 45 % User in niedrigeren Einkommensklassen gegenüber.

ABBILDUNG 180 PC-Nutzung (ordinal): Prädiktorwerte (CART-Algorithmus, N = 1413)



Allerdings sind die Verbesserungswerte wesentlich geringer als im binären Fall - sie beträgt insgesamt nur noch etwa 0.07.

Die Fehlklassifikation liegt bei 0.404 - wie angedeutet, kein überragendes, aber im Vergleich mit anderen Verfahren durchaus ebenbürtiges Ergebnis. Mit 0.4 ist die Fehlklassifikation genauso hoch wie bei EXHAUSTIVE CHAID.

Obwohl sich die nominalen und die ordinalen Bäume sehr ähneln, hat die Kategorie der Light User einen gewissen Störeffekt, der - obwohl die Variable statistisch differenzierter ist als die nominale Recodierung - das ganze Ergebnis schmälert. Inhaltlich stellt sich ebenfalls die Frage, was eine derartige Differenzierung bezwecken mag.

5.5 Metrische Entscheidungsbäume

Um didaktisch zu zeigen, wie Entscheidungsbäume klassifizieren, wird nochmals mit der ordinalen Variable eine metrische Analyse durchgeführt. Dies ist statistisch nicht erlaubt, sollte auch in der Praxis niemals so durchgeführt werden. Es dient nur dazu zu zeigen, wie dieses Verfahren anwendungsorientiert eingesetzt werden kann. Aufgrund der Einschränkungen von QUEST, wo nominale Zielvariablen unter-

stellt wenden, können nur CART und EXHAUSTIVE CHAID miteinander verglichen werden. Auch für diese Beispiele sind die Ergebnisse mit Vorsicht zu interpretieren - da die Zielvariable Non-, Light-, Heavy User Ordinalskalenniveau besitzt, jedoch nicht metrisch ist. Der Hauptknoten für beide Verfahren sieht etwas anders als im nominalen bzw. ordinalen Fall aus:

ABBILDUNG 181 PC-Nutzung (metrisch): Wurzelknoten (EXHAUSTIVE CHAID- bzw QUEST-Algorithmus, N = 1413)

Knoten 0	
Mittelwert	1,9618
Std.abw.	0,8176
n	1413
%	100,00
Vorhergesagt	1,9618

In beiden Fällen wird mit dem Mittelwert und mit der Standardabweichung für metrische Variablen gearbeitet. Die (ordinale) Skalierung lautet: 1 = Non User, 2 = Light User, 3 = Heavy User. Anhand der Fälle für jede Kategorie wird der gewichtete arithmetische Mittelwert gebildet, der mit 1.96 nahe bei den Light Usern liegt.

Folgende Rechnung liegt dem zugrunde:

```
Mittelwert = (500 * 1 + 467 * 2 + 446 * 3) / 1413.
```

Der vorhergesagte Mittelwert muss natürlich im Hauptknoten identisch mit dem tatsächlichen sein. Die Standardabweichung gibt an, wie „weit" die untersuchten Werte um den Mittelwert „streuen", also „durchschnittlich" von ihm abweichen. Es handelt sich hierbei um keinen standardisierten Wert, der von Verteilung zu Verteilung vergleichbar wäre. Würde man zum Beispiel die Standardabweichung von (metrisch gemessenen) Haushaltsnettoeinkommen untersuchen, wäre die Standardabweichung wesentlich höher, da Einkommen in

einem mindestens dreistelligen Bereich liegen, dadurch auch die Standardabweichung steigt.

Die Kennwerte für N und % geben nur die Zahl der Fälle (1413) und den jeweiligen Knotenanteil an - in diesem Fall natürlich 100 %, da alle Fälle in die Analyse eingehen.

Für den Wurzelknoten bedeutet das, dass die Standardabweichung nicht sehr weit um den Mittelwert streut.

ABBILDUNG 182 PC-Nutzung (ordinal): Wurzelknoten (CART-Algorithmus, N = 1413)

Knoten 0

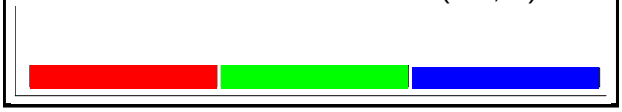
Kategorie	%	n
Non-User	35,39	500
Light User (bis mehrm. wöchentl.)	33,05	467
Heavy User (täglich)	31,56	446
Gesamt	(100,00)	1413

Dies läßt sich auch an der ordinalen Segmentierung zeigen: da die Anteile jeweils mit etwa 1/3 gleichverteilt sind, muß somit der Mittelwert etwa bei 2 liegen. Da die Kategorie 1 (Non User) mit 35 % den größten Anteil besitzt und die Kategorie der Heavy User (3) die kleinste ist, muss der Mittelwert etwas unter 2 liegen. Um diesen Sachverhalt auch grafisch darzustellen, ist es möglich, auch die Grafik innerhalb der Knoten anzuzeigen:

ABBILDUNG 183 PC-Nutzung (ordinal): Wurzelknoten (CART-Algorithmus, statistische und grafische Darstellung, N = 1413)

Knoten 0

Kategorie	%	n
Non-User	35,39	500
Light User (bis mehrm. wöchentl.)	33,05	467
Heavy User (täglich)	31,56	446
Gesamt	(100,00)	1413

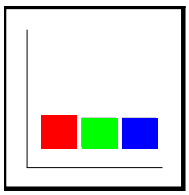


Bisher wurde auf diese Möglichkeit der Darstellung verzichtet, um die komplexen Bäume nicht durch weitere „Features“ zu verkomplizieren - denn im Mittelpunkt steht die Frage, wie die Algorithmen arbeiten und welche Ergebnisse sie erbringen.

Für dieses Kapitel, das „nur“ die Vorgehensweise bei der Generierung von Entscheidungsbäumen mit metrischen Variablen anhand der ordinal skalierten PC-Nutzung zeigt, soll neben den statistischen Ergebnissen auch die grafische Darstellung erläutert werden. Durch die Tatsache, dass als statistische Kennzahl „nur“ der Mittelwert und die Standardabweichung zentral im Mittelpunkt stehen, müssen auch nicht unterschiedliche Kennzahlen erläutert werden - somit können die Stärken der Grafiken aufgezeigt werden.

Daneben ist es auch möglich, die Ergebnisse nur grafisch darzustellen:

ABBILDUNG 184 PC-Nutzung (ordinal): Wurzelknoten (CART-Algorithmus, grafische Darstellung, N = 1413)



Somit kann die grafische und statistische Darstellung einige Vorteile bringen: auf den ersten Blick lassen sich hohe bzw. niedrige Anteile hinsichtlich der Zielvariablen identifizieren.

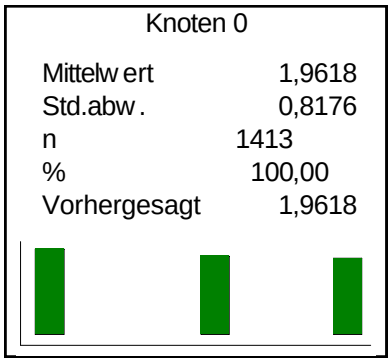
Neben dem bisher aufgezeigten Weg, die Entscheidungsbäume Ast für Ast zu interpretieren, gibt es ebenfalls die bereits angedeutete Möglichkeit, Entscheidungsbäume nach ihren Gewinnanteilen hinsichtlich der Zielkategorien zu untersuchen.

5.5.1 EXHAUSTIVE CHAID-Algorithmus

(EXHAUSTIVE) CHAID	• sehr verbreitet • segmentiert zwei oder mehr Unterknoten • für alle Skalenniveaus geeignet • gebräuchliche statistische Kennzahlen (Chi-Quadrat, Likelihood, F-Test)

Der Hauptknoten des EXHAUSTIVE CHAID-Baums liefert das bekannte Ergebnis:

ABBILDUNG 185 PC-Nutzung (metrisch): Wurzelknoten (EXHAUSTIVE CHAID- bzw QUEST-Algorithmus, statistische und grafische Darstellung, N = 1413)

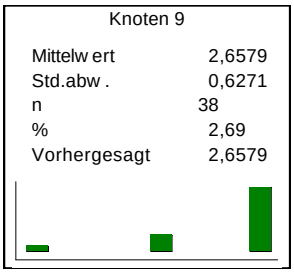
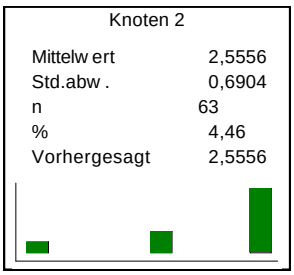
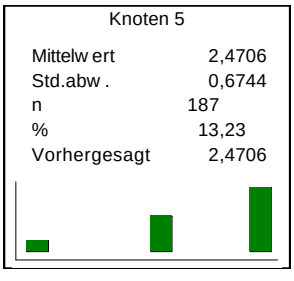


Der Mittelwert liegt - wie erläutert - bei 1.96. Die Balken repräsentieren die Anteile und sind in etwa gleich groß. EXHAUSTIVE CHAID findet 9 Berufssegmente, wobei zwei davon weiter unterteilt werden. Die Reise- und Dienstleistungsangestellten, Landwirte und Fischer, die zusammengefaßt wurden, werden weiter nach dem Einkommen segmentiert (bis 2250, 2251 bis 4500 DM und fehlende Werte bzw. über 4500 DM). Die Arbeiter werden weiter nach Alter (bis 33 Jahre, 34 - 48 Jahre, älter als 48 Jahre) differenziert. Somit ergeben sich keine neuen Ergebnisse im Vergleich zum nominalen Fall.

Die Gewinndarstellung läßt sich entweder als Gewinnübersicht - wie in Answertree ausgegeben - oder mit den einzelnen Knoteninformationen kombinieren. Dahinter könnte die Frage stehen: in welchem

Segment sind vor allem PC-Nutzer anzutreffen (z. B. für Marketingmaßnahmen)? Welche (berufsständischen) Zeitschriften könnten für Werbeaktivitäten genutzt werden, ohne allzugroße Streuung zu erzielen? Die Knoten mit den höchsten Gewinnanteilen könnten exemplarisch folgendermaßen dargestellt werden:

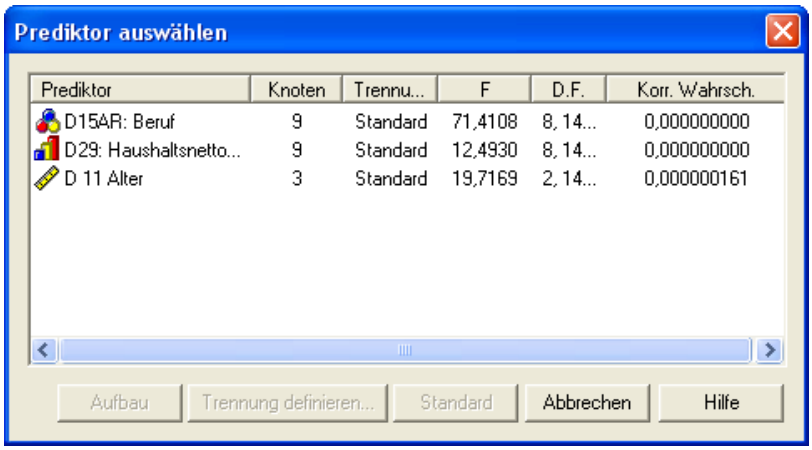
TABELLE 50 PC-Nutzung (metrisch): Knotenweise Gewinnübersicht (EXHAUSTIVE CHAID, N = 1413)

Knoten	Knoten: Anzahl	Knoten: %	Profit	Index (%)
Freie Berufe 	38	2,7	2,66	135,5
Grossunternehmer, Direktoren, Top Management, Leitende Angestellte 	63	4,5	2,56	130,3
Büroangestellte mit Leitungsfunktion 	187	13,2	2,47	125,9

Deutlich wird, dass die Anteile kontinuierlich von den Heavy Usern über die Light User zu den Non Usern sinken. Die Mittelwerte liegen mit 2.5 bis 2.7 deutlich höher als der Durchschnitt mit 1.96, die Standardabweichung ist mit rund 0.6 - 0.7 etwas niedriger als in der Gesamtstichprobe. Gerade dieser Punkt zeigt auch die Schwierigkeit dieses Beispiels auf, die Werte sinnvoll zu interpretieren. Was sagt eine Differenz des Mittelwerts um 0.1 oder 0.2 aus? - Im Prinzip überhaupt nichts. Somit ist die inhaltliche Interpretation dieses Beispiels nicht sinnvoll.

Die Fehlklassifikation liegt bei 0.45, darf aber durch die ordinale Zielvariable nicht interpretiert werden.

ABBILDUNG 186 PC-Nutzung (metrisch): Prädiktorwerte (EXHAUSTIVE CHAID-Algorithmus, N = 1413)



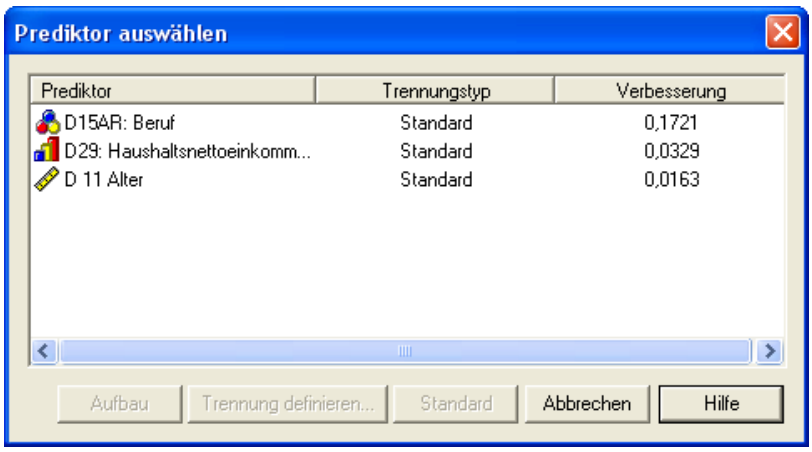
Die Reihenfolge der Prädiktoren wird durch den F-Wert der Varianzanalyse gebildet, der für den Beruf mit 71 deutlich höher ausfällt als für Haushaltsnettoeinkommen und Alter. Hierbei fällt auf, dass der F-Wert für Alter mit knapp 20 höher ist als für das Haushaltsnettoeinkommen - allerdings ist die Signifikanz, also die Sicherheit, mit der Ergebnisse auf die Grundgesamtheit übertragen werden können, geringer, was das Alter auf Platz 3 verweist.

5.5.2 CART-Algorithmus

CART (C&RT)	• vieldiskutierter Algorithmus • segmentiert immer nur zwei Unterknoten • für alle Skalenniveaus geeignet • weniger gebräuchliche Maßzahlen in den Sozialwissenschaften (Gini, Twoing, ordered twoing) • erlaubt Ersatzprädiktoren, Pruning und a prioris

Mit 0.46 Fehlklassifikation liegt die Güte beim CART-Baum etwas unter dem von EXHAUSTIVE CHAID. Der Wurzelknoten ist identisch. CART findet 12 Segmente, wobei die Reihenfolge der Prädiktoren Beruf, Haushaltsnettoeinkommen und Alter ebenfalls hier eingehalten werden.

ABBILDUNG 187 PC-Nutzung (metrisch): Prädiktorwerte (CART-Algorithmus, N = 1413)

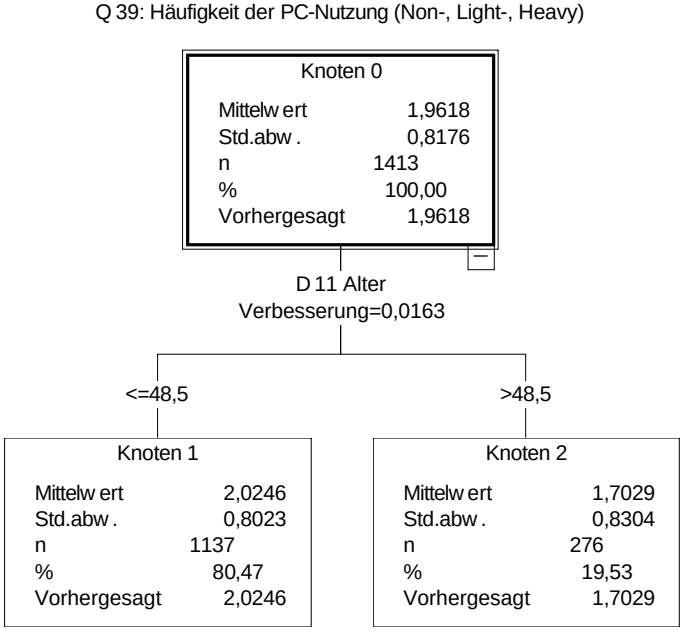


Während der Beruf deutlich mehr zur Verbesserung der Bäume beiträgt (0.17) ist der Anteil von Haushaltsnettoeinkommen mit 0.03 und Alter (0.01) eher bescheiden. Auch hier finden sich die gleichen Segmente wie im ordinalen Fall, weshalb nicht weiter auf die Darstellung des Baums eingegangen werden soll.

Die Risikoschätzung wird nach LSD (= Least Squared Deviation) berechnet. Um das Beispiel überschaubar zu halten und nicht nochmal die ganzen Berufskategorien zu zitieren, wurde ein Entscheidungs-

baum generiert, der auf der ersten Stufe das Alter zur Segmentierung heranzieht:

ABBILDUNG 188 BINÄRE ALTERSSEGMENTIERUNG MIT CART (N = 1413)



Die Berechnung erfolgt aufgrund der Standardabweichungen, die nur für den metrischen Fall Verwendung finden. Die Standardabweichungen der Unterknoten werden jeweils quadriert, mit den Knotenanteilen gewichtet und schließlich aufaddiert:

$$\text{Risikoschätzung} = \text{Standardabw. Knoten } 1^2 * \text{Prozentanteil} + \text{Standardabweichung Knoten } 2^2 * \text{Prozentanteil}$$

$$0.8023^2 * 0.8047 + 0.8304^2 * 0.1953 =$$

$$0.6437 * 0.8047 + 0.6896 * 0.1953 =$$

$$0.5180 + 0.1348 = 0.652$$

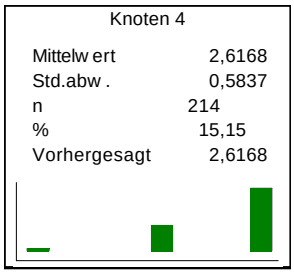
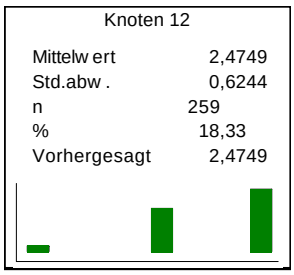
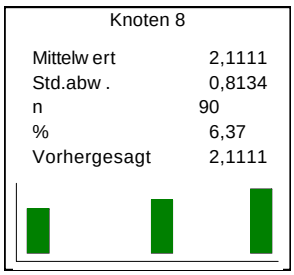
Ein Blick auf die Risikoschätzung von Answertree ergibt ein identisches Ergebnis (mit unerheblichen Rundungsdifferenzen):

```
Risikoschätzung 0,651768
```

Anders ausgedrückt: Mit der Quadrierung der Standardabweichung wird die Varianz ermittelt. Diese Varianz wird mit den jeweiligen Anteilen der Unterknoten gewichtet, so dass eine Art „gewichtete Varianz“ entsteht.

Die Gewinnübersicht für die „Top 3“ ist in nachfolgender Tabelle enthalten:

TABELLE 51 PC-Nutzung (metrisch): Knotenweise Gewinnübersicht (CART, N = 1413)

Knoten	Knoten: Anzahl	Knoten %	Profit	Index (%)
Großunternehmer, Direktoren, Top Management, Leitende Angestellte, sonstige Bürotätigkeiten, Büroangestellte mit Leitungsfunktion, Studierende, Freie Berufe, Meister, Haushaltsnettoeinkommen > 4500 DM 	214	15,1	2,62	133,4
Großunternehmer, Direktoren, Top Management, Leitende Angestellte, sonstige Bürotätigkeiten, Büroangestellte mit Leitungsfunktion, Studierende, Freie Berufe, Meister, Haushaltsnettoeinkommen bis 4500 DM, 18 - 45 Jahre 	259	18,3	2,47	126,2
Großunternehmer, Direktoren, Top Management, Leitende Angestellte, sonstige Bürotätigkeiten, Büroangestellte mit Leitungsfunktion, Studierende, Freie Berufe, Meister, Haushaltsnettoeinkommen bis 4500 DM, 46 Jahre und älter 	90	6,4	2,11	107,6

Auch hier sinken die Anteile monoton von den Heavy- über die Light- zu den Nonusern. Durch die erzwungene Dichotomisierung wird die Baumstruktur tiefer und ermöglicht einen höheren Informationsgehalt.

5.6 Zusammenfassung

Im ordinalen Fall wird durchgängig schlechter klassifiziert als mit nominal-dichotomen Zielvariablen (in diesem Fall auch für die als metrisch angenommene Zielvariable). Dies liegt an den „Mischtypen", die in die Analyse eingehen und somit schwerer zuzuordnen sind (in diesem Fall: Light User) - was auch Auswirkungen auf die beiden anderen Kategorien besitzt.

Das beste rechnerische Ergebnis liefert in diesem Fall die ordinale Regression - allerdings vor allem dadurch, dass die am schwierigsten zu klassifizierenden Fälle (fehlendes Haushaltsnettoeinkommen) von vornherein aus der Analyse ausgeschlossen werden, was die Zuordnung zu den Typen der PC-Nutzung vereinfacht - im Gegensatz z. B. zu den Entscheidungsbäumen.

Auch hier zeigt sich, dass die Verfahren ähnlich „gut" clustern. Eine Abweichung der Fehlklassifikation von einigen Prozent ist - aus meiner Sicht - vernachlässigbar.

Welches Verfahren man nun heranzieht - es hängt eher von persönlichen Vorlieben als von rechnerischer Exaktheit ab. Wenn alle drei Verfahren zu einem ähnlichen Ergebnis kommen, gibt es keine mathematisch begründbare Präferenz für ein bestimmtes Verfahren. Eher sollten hier inhaltliche Kriterien, z. B. die Verwendung von Ersatzprädiktoren bei bestimmten Entscheidungsbaumalgorithmen, herangezogen werden.

KAPITEL V Zusammenfassung, Kritik und Schluss

1 Zusammenfassung

PC-Nutzung und deren sozialstrukturelle Beschreibung, differenziert nach Kultur- und Freizeitvariablen, waren Ziel des Untersuchungsinteresses. Hierbei wurde theoretisch auf das Konzept der dominanten (subordinierten) Schichtungen von Theodor GEIGER zurückgegriffen, der unterstellte, dass es viele Schichtungslinien für eine Fragestellung gibt, aber nur die wichtigsten (dominantesten) zu einer empirischen Analyse herangezogen werden können.

Durch die Restriktionen des EUROBAROMETER-Datensatzes, der dieser Untersuchung zugrunde liegt und der vor allem nominale und ordinale Variablen enthält, wurden neben den deskriptiven Maßen die logistische Regression, die Diskriminanzanalyse und Entscheidungsbaumalgorithmen herangezogen. Letzteres Verfahren wurde in der Soziologie noch kaum angewandt, so daß eine Überprüfung durch die beiden anderen, „bewährten“ Verfahren sinnvoll erschien.

Grundsätzlich wurde davon ausgegangen, daß Kultur- und Freizeitvariablen den sozialstrukturellen („alte“) Ungleichheiten nachgeordnet werden, wobei vertikale sozialstrukturelle Variablen neben Bildung, Beruf und Einkommen auch Variablen wie Alter und Geschlecht beinhalten. Hier wird davon ausgegangen, dass gerade im Bereich der Informationstechnologie diese beiden Variablen deutlich zur Ungleichheit beitragen (z. B. durch die zunehmende Gefahr, bei Arbeitsplatzverlust in höherem Lebensalter noch eine Stelle zu finden). Somit konnten dominante Schichtungen nur sozialstrukturell begründet, subordinierte Linien können sowohl sozialstrukturell als auch kultur- oder freizeitbezogen sein.

Ziel war es nicht, ein neues (gesamtgesellschaftliches) Milieumodell zu entwickeln, sondern aufzuzeigen, wie PC-Nutzung in unterschiedlichen sozialstrukturellen Ressourcen zum Ausdruck kommt.

Der Vorteil dieser Untersuchung liegt darin, dass nicht nur PC-Nutzer befragt wurden, sondern auch Nichtnutzer - auf repräsentativer Basis. Somit können nicht nur Aussagen über die PC-Nutzung, sondern auch über die Nichtnutzung getroffen werden.

Als wichtigste Variablen haben sich das Alter, der Beruf, das Einkommen und der Bildungsgrad herauskristallisiert. Interessanterweise spielt das Geschlecht kaum eine Rolle für die generelle PC-Nutzung, die hier untersucht wird. Während in der Gesamtstichprobe das Alter als Variable eindeutig dominiert, was auch häufig das Ergebnis anderer Untersuchungen war, kann man das Ergebnis weiter differenzieren. So zeigt sich, dass in Altersgruppen bis etwa zum Renteneintritt der Beruf deutlich vor dem Alter dominiert - vor allem, wenn der ausgeübte Beruf eine gewisse Affinität zum PC-Einsatz besitzt.

Allerdings kann vor allem in jüngeren Alterskategorien das Alter oder auch das Einkommen ausschlaggebend für die PC-Nutzung sein - was sich recht allgemein zu folgenden Regeln verdichten läßt:

- Dominierend bei den Altersgruppen bis etwa 58 Jahre ist der Beruf
- In den jüngsten Altersgruppen gehört der PC wie selbstverständlich zum Leben - unabhängig von Beruf und Einkommen
- In mittleren Altersgruppen entscheidet häufig das Einkommen bei nicht PC-bezogenen Berufen über die Anschaffung: je höher das Einkommen desto eher erfolgt die Anschaffung

Vor allem weiterbildungsbezogene, nicht so sehr „klassische“ Kultur- und Freizeitvariablen erklären die PC-Nutzung: berufliche, Pflicht- und freiwillige Weiterbildung. Aber auch Kinobesuch und Volksmusikhören bzw. Volksmusikkonzerte besuchen tragen zu einer Erklärung der PC-Nutzung bei. Andere, typische Lebensstilvariablen, z. B. Musikrichtungen wie Klassische Musik haben keinen hohen Einfluß.

2 Kritik

2.1 Grafische Verfahren

Die hier vorgestellten, grafisch aufbereiteten, Ergebnisse sind durchaus verbesserungswürdig. Die Limitationen resultieren - wie beschrieben - aus der im Augenblick noch nicht vollständig entwickelten Software. Es ist jedoch davon auszugehen, dass sich dieses Manko in den nächsten Jahren ändern wird.

Diese neuen grafischen Methoden - seien es Parallel-, Mosaic- oder Spine Plots - stellen neue, interessante Möglichkeiten der Visualisierung zur Verfügung, die in den Sozialwissenschaften nicht ungenutzt bleiben sollten, beten sie doch die Chance, statistischen Laien empirische Ergebnisse transparenter zu machen.

Vor allem in einer Zeit, in der durch die Computernutzung die Möglichkeit geschaffen wird, umfangreiche Datenmengen zu verarbeiten, helfen gerade grafische Verfahren, einen „Überblick“ zu erhalten. Dies geschicht mit dem Ziel, eine Präsentation der Ergebnisse oder der weiteren Aufbereitung zu ermöglichen. Daraus ergeben sich auch neue Blickwinkel und Chancen für die Auswertung von Daten. Nicht nur, dass die Leser von Forschungsberichten aufbereitete grafische Ergebnisse erhalten, Grafiken sind in der Regel schneller und einfacher zu interpretieren als z. B. eine umfangreiche Kreuztabelle.

2.2 Entscheidungsbäume

Auch wenn beim einen oder anderen Leser der Eindruck entstanden ist, logistische Regression, Diskriminanzanalyse und Entscheidungsbäume seien substituierende Verfahren, so ist das nicht richtig. Entscheidungsbäume unterscheiden sich sehr deutlich von den anderen beiden Verfahren:

- Bei der logistischen Regression steht die Frage im Vordergrund, wie hoch der Einfluss von unabhängigen Variablen auf eine Zielvariable ist - sowohl als Einzel- als auch als Gruppeneffekt.
- Die Diskriminanzanalyse versucht, aufgrund von unabhängigen Variablen Gruppen zu finden und diese möglichst gut zu trennen
- Entscheidungsbaumalgorithmen versuchen, in jedem Schritt der Analyse für jedes Subsample die wichtigsten trennenden unabhängigen Variablen heranzuziehen.

Es wäre an dieser Stelle unredlich, das eine Verfahren gegen das andere „auszuspielen" in der Form des „besseren" oder „schlechteren" Verfahrens. Es ist vielmehr darauf hinzuweisen, welche Grundforschungsfrage beantwortet werden soll: ist es eher wichtiger generelle Effekte von unabhängigen Variablen zu erhalten, sollte auf jeden Fall die logistische Regression herangezogen werden. Ist es wichtiger, allgemein Segmente zu bilden kann auf die Diskriminanzanalyse zurückgegriffen werden.

Interessiert allerdings der „innere Zusammenhalt" von Gruppen, d. h., wie sich Subgruppen am besten durch die unabhängigen Variablen beschreiben lassen, so sind - von diesen drei Verfahren - Entscheidungsbäume zu bevorzugen. Darüberhinaus wird auch noch über die Prädiktorenansicht die Stärke (und damit auch die Reihenfolge) der unabhängigen Variablen insgesamt angegeben.

3 Schluss

Ziel der Arbeit ist es zu prüfen, ob Entscheidungsbaumalgorithmen ähnliche Ergebnisse wie Logistische Regression und Diskriminanzana-

lyse liefern und somit den Einsatz in den Sozialwissenschaften rechtfertigt oder nicht. Anhand der Frage der dominanten bzw. subordinierten Schichtungen bei GEIGER soll überprüft werden, ob sich die generelle PC-Nutzung heute noch sozialstrukturell erklären läßt.

Die methodischen Ergebnisse nochmals in aller Kürze:

- Baumalgorithmen liefern detailliertere Ergebnisse als die beiden anderen genannten Verfahren: Nicht nur die Stärke von unabhängigen auf eine abhängige Variable wird berechnet, sondern in welchem Segment sich dies auswirkt. Somit erhält man ein charakteristisches, gruppenspezifisches Ergebnis. Durch die verschiedenen Baumalgorithmen lassen sich innerhalb dieses Verfahrens Ergebnisse weiter validieren oder erweitern
- Entscheidungsbäume liefern ein grafisches Ergebnis, das auch statistische Laien nachvollziehen können
- Entscheidungsbäume lassen sich individuell steuern, sie lassen sich auch „manuell", Stufe für Stufe, erzeugen
- Somit ergeben sich deutlichere Segmente (z. B. junge (Fach-)Arbeiter, die sehr stark den PC nutzen im Gegensatz zu den Älteren dieser Berufsgruppe) bilden
- „Mischgruppen" hinsichtlich der unabhängigen Variable lassen sich nur von den Entscheidungsbäumen, nicht aber von anderen multivariaten Verfahren identifizieren

Die theoretischen Ergebnisse sind ebenfalls bereits weiter oben ausführlich dargestellt worden - deshalb auch hier nochmals kurz Stichpunkte:

- Die Wichtigkeit der unabhängigen Variablen, bezogen auf die PC-Nutzung ergeben sich durch die jeweiligen Zusammenhangswerte (in der Regel Phi, Cramers v, Unsicherheitskoeffizient). Es wird zwischen dominanten und subordinierten Schichtungen unterschieden. Dominant sind - in der Logik dieser Arbeit - ausschließlich sozialstrukturelle Variablen, die einen hohen Zusammenhangswert aufweisen. Subordiniert sind diejenigen Kultur- und Freizeitvariablen, die stark mit der PC-Nutzung korrelieren
- Als dominante Schichtungen können Beruf, Alter und Haushaltsnettoeinkommen angesehen werden. Der Bildungsgrad spielt ebenfalls eine große Rolle, korreliert aber sehr stark mit dem Beruf. Deshalb wurde diese Variable nicht herangezogen
- Als subordinierte Schichtungen liefern Kinobesuch, Volksmusikhören bzw. der Besuch von Volksmusikkonzerten sowie Weiterbildungsvariablen die höchsten Zusammenhänge

Es ist in diesem Zusammenhang nicht verwunderlich, dass gerade Weiterbildungskategorien PC-Nutzung erklären - ist der PC doch ein

relativ neues Medium, mit dem vor allem jüngere Menschen umgehen. Diese befinden sich möglicherweise in der Berufseintrittsphase, wo Weiterbildung (z. B. die Aneignung von Software) heute dazugehört.

Interessanterweise lassen sich die Kulturschemata wie das Hochkulturschema nicht direkt nachweisen - wohl aber einige Ergebnisse wie z. B. Volksmusikhören verifizieren. Allerdings ist die Frage nach PC-Nutzung keine Frage der sozialstrukturellen Schichtung einer Gesellschaft. Trotzdem bleibt es fraglich, ob derartige Kulturbegriffe heute angemessen sind, eine Gesellschaft alleinig zu beschreiben.

Insgesamt läßt sich feststellen, dass Entscheidungsbäume auf jeden Fall in die „bewährten" Methoden der multivariaten Statistik aufgenommen werden können. Sie haben in dieser Arbeit den überzeugenden Beweis erbracht, ähnliche Ergebnisse wie die anderen Verfahren zu liefern. Neben den o. g. Abgrenzungen zu den anderen Verfahren gibt es beim Einsatz - zusammenfassend - folgende Vorteile:

- alle Skalenniveaus können verzerrungsfrei von nahezu allen (eingesetzten) Algorithmen verarbeitet werden (Ausnahme: QUEST, der eine dichotome abhängige Variable voraussetzt)
- sowohl die Stärke als auch die besondere Bedeutung einzelner unabhängiger Variablen für Subsamples wird berücksichtigt: neben der Wichtigkeit für den Entscheidungsbaum insgesamt werden die einzelnen Gruppen charakterisiert
- es können sehr anschaulich Segmente mit hohen „Gewinn"anteilen identifiziert werden. Das Verfahren ist sehr transparent und klar strukturiert
- der ausgegebene Entscheidungsbaum kann auch von statistischen Laien verstanden werden, ohne auf „höhere Mathematik" (z. B. Logarithmen) zurückzugreifen

Es wäre aus meiner Sicht wünschenswert, wenn sowohl Entscheidungsbäume als auch die hier dargestellten grafischen Verfahren in den Sozialwissenschaften rege genutzt würden.

LITERATURVERZEICHNIS

a. Bücher, Aufsätze, Manuskripte

o. A. (2001): Grüne wollen langfristig 600 Mark Kindergeld, in: Sueddeutsche Zeitung vom 05. Juni 2001, S. 6

ADM et al. (2001): Standards zur Qualitätssicherung für Online-Befragungen, in: http://www.adm-ev.de [15.05.2001]

Bacher, Johann et al. (2004): SPSS Two Step Cluster - A First Evaluation, Arbeits- und Diskussionspapiere des Lehrstuhls für Soziologie 2004-2 der Universität Erlangen-Nürnberg, Nürnberg: Ms.

Bachmann, Siegfried (Hrsg.) (1995): Theodor Geiger - Soziologe in einer Zeit zwischen Pathos und Nüchternheit - Beiträge zu Leben und Werk, Berlin: Duncker & Humblot

Backhaus, Klaus et al. (2000): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, Springer: Berlin

Backhaus, Klaus et al. (2004): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, Berlin: Springer

Baltes-Götz, Bernhard (2001): Segmentierung und Klassifikation mit Answertree 2.1, in: http://www.uni-trier.de/urt/user/baltes/docs/at/v21/at21.pdf [28.03.2003]

Baltes-Götz, Bernhard (2004a): Entscheidungsbaumanalyse mit Answertree 3.1, in: http://www.uni-trier.de/urt/user/baltes/docs/at/v31/at31.pdf [28.01.2005]

Baltes-Götz, Bernhard (2004b): Logistische Regressionsanalyse mit SPSS, in: http://www.uni-trier.de/urt/user/baltes/docs/logist.pdf [14.03.2005]

Bandrilla, Wolfgang (1999): www-Umfragen - eine alternative Datenerhebungstechnik für die empirische Sozialforschung?, in: Batinic, Bernad/Werner, Andreas et al.: Online-Research. Methoden, Anwendungen und Ergebnisse, Göttingen: Hogrefe

Baudrillard, Jean (1991): Der symbolische Tausch und der Tod, München: Matthes & Seitz

Baudrillard, Jean (2001): Das System der Dinge. Über unser Verhältnis zu den alltäglichen Gegenständen, Frankfurt. Campus

Baur, Nina (2003): Bivariate Statistik, Drittvariablenkontrolle und das Ordinalskalenproblem. Eine Einführung in die Kaulsalanalyse und in den Umgang mit zweidimensionalen Häufigkeitsverteilungen, Bamberg: Bamberger Beiträge zur empirischen Sozialforschung, hrsg. von Gerhard Schulze und Nina Baur, Nr. 9/2003

BBDO Group Germany (Hrsg.) (2000): Wichtige Segmentierungsverfahren und aktuelle Marktstudien, in: http://www.bbdo.de [08.01.2001]

Beck, Ulrich (1983): Jenseits von Klasse und Stand?, in: Kreckel, Reinhard: Soziale Ungleichheit, Göttingen: Hogrefe

Beck, Ulrich (1986): Risikogesellschaft. Auf den Weg in eine andere Moderne, Frankfurt: Suhrkamp

Beck, Ulrich (1996): Das Zeitalter der Nebenfolgen und die Politisierung der Moderne, in: Beck, Ulrich et al.: Reflexive Modernisierung. Eine Kontroverse, Frankfurt: Suhrkamp

Benninghaus, Hans (1979): Statistik für Soziologen, Band 1: Deskriptive Statistik, Stuttgart: Teubner

Benjamin, Walter (1977): Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit, Frankfurt: Suhrkamp

Berger, Peter A. (1998): Soziale Mobilität, in: Schäfers, Bernhard und Zapf, Wolfgang: Handwörterbuch zur Gesellschaft Deutschlands, Bonn: Bundeszentrale für politische Bildung

Berry, Michael A. / Linoff, Gordon S. (2000): Mastering Data Mining. The Art and Science of Customer Relationship Management, New York: Wiley

Böhm, Wolfgang und Wehner, Josef (1990): Der symbolische Gehalt einer Technologie. Zur soziokulturellen Rahmung des Computers, in: Rammert, Wertner (Hrsg.): Computerwelten - Alltagswelten. Wie verändert der Computer die soziale Wirklichkeit?, S. 105ff.

Bourdieu, Pierre (1997 [1982]): Die Feinen Unterschiede. Kritik der gesellschaftlichen Urteilskraft, Frankfurt: Suhrkamp

Bourdieu, Pierre (1997): Das Elend der Welt. Zeugnisse und Diagnosen alltäglichen Leidens an der Gesellschaft. Konstanz: Universitätsverlag Konstanz

Breiman, Leo et al. (1984): Classification and Regression Trees, Belmont, California: Wadsworth

Breton, Philippe (2000): Hippies und Beatniks und Internet-Junkies. Eine Turbo-Subkultur, in: Le Monde Diplomatique, 31.10.2000, S. 2

Brosius, Felix (1998): SPSS 8 - Professionelle Statistik unter Windows, Bonn: MITP-Verlag

Brosius, Felix (2002): SPSS 11, Bonn: mitp-Verlag

Brosius, Gerhard (1989): SPSS/PC+ - Advanced Statistics und Tables, Hamburg: McGraw-Hill

Bühl, Achim (Hrsg.) (1999): Computerstile. Vom individuellen Umgang mit dem Pc im Alltag, Opladen/Wiesbaden: Westdeutscher Verlag

Bühl, Achim und Zöfel, Peter (2002a): SPSS 11. Einführung in die moderne Datenanalyse unter Windows, München: Addison Wesley, Reihe Pearson Studium

Bühl, Achim und Zöfel, Peter (2002b): Erweiterte Datenanalyse mit SPSS. Statistik und Data Mining, Wiesbden: Westdeutscher Verlag

Bundesminister für Bildung und Wissenschaft (1989): Das soziale Bild der Studentenschaft in der Bundesrepublik Deutschland - 12. Sozialerhebung des deutschen Studentenwerks, Bonn

Bundesministerium für Arbeit und Sozialordnung (Hrsg.) (2001): Lebenslagen in Deutschland. Der erste Armuts- und Reichtumsbericht der Bundesregierung. Band I: I: Bericht, Band II: Daten und Fakten, Bonn

Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (Hrsg.) (1995): Das soziale Bild der Studentenschaft in der Bundesrepublik Deutschland - 14. Sozialerhebung des Deutschen Studentenwerks, Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie

Carvin, Andy (2000): Beyond Access: Understanding the Digital Divide, in: http://www.benton.org/divide/thirdact/speech.html [18.02.2002]

ComCult Research (2000): Zielgruppen im Netz 2000, in: http://www.comcult.de/infopool/in_sozio.htm [31.10.00]

ComCult Research (2001): ComCult Panel Report: Online-Nutzung 2001, in: http://www.comcult.de [28.11.2001]

Costigan, James T. (2000): Forests, Trees and Internet Research, in: Steve Jones (Ed.): Doing Internet Research - Critical Issues and Methods for Examining The net, Thousand Oaks: Sage

Dahrendorf, Ralf (1965): Gesellschaft und Demokratie in Deutschland, Piper: München

Diaz-Bone, Rainer (o. J.): Eine kurze Einführung in die logistische Regression und binäre Logit-Analyse, in: http://www.agis.uni-hannover.de/eqqs/modulijk/logistische_regression.pdf [03.09.2004]

Diekmann, Andreas (1995): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, Reinbek: Rowohlt

Döllner, Jürgen (2003): Introduction to Visualization, Berlin: Ms

Durkheim, Emile (1991): Die Regeln der soziologischen Methode, Frankfurt: Suhrkamp

Eimeren, Birgit van et al. (1997): ARD-Onlinestudie 1997: Onlinenutzung in Deutschland, in: MediaPerspektiven 10/1997, S. 548ff.

Eimeren, Birgit van. et al. (1998): ARD/ZDF-Onlinestudie 1998: Onlinemedien gewinnen an Bedeutung, in: http://www.das-erste.de/studien [28.11.2000]

Eimeren, Birgit van et al. (1999a): ARD/ZDF-Onlinestudie 1999: Wird Online Alltagsmedium? in: MediaPerspektiven 8/1999, S. 401ff.

Eimeren, Birgit van. et. al. (1999b): Nichtnutzer von Online: Einstellungen und Zugangsbarrieren, in: MediaPerspektiven 8/1999, S. 415ff.

Eimeren, Birgit van et al. (1999c): Internet - (k)eine Männerdomäne, in: MediaPerspektiven 8/1999, S. 423ff.

Eimeren, Birgit van und GERHARD, Heinz (2000): ARD/ZDF-Onlinestudie 2000: Gebrauchswert entscheidet über Internenutzung, in: MediaPerspektiven 8/2000, S. 338ff.

Eimeren, Birgit van et al. (2001): ARD/ZDF-Onlinestudie 2001: Internetnutzung stark zweckgebunden, in: MediaPerspektiven 8/2001, S. 382ff.

Flaig, Berthold et al. (1997[3]): Alltagsästhetik und politische Kultur. Zur ästhetischen Dimension politischer Bildung und politischer Kommunikation, Bonn: Dietz

Friedrichs, Jürgen (1990[14]): Methoden empirischer Sozialforschung, Opladen: Westdeutscher Verlag

Fua, Ying-Huey Fua, Ward Matthew O. et al. (1999): Hierarchical Parallel Coordinates for Exploration of Large Datasets, in: http://delivery.acm.org/10.1145/320000/319355/p43-fua.pdf?key1=319355&key2=8151070011&coll=GUIDE&dl=GUIDE&CFID=31401988&CFTOKEN=80232452 [17.11.2004]

Funken, Michael (2000): Hartmut Winkler: Medientheorie der Computer, http://www.information-philosophie.de/philosophie/medientheoriewinkler.html [26.07.2001]

Gehring, Uwe UND Weins, Cornelia (1998): Grundkurs Statistik für Politologen, Opladen/Wiesbaden: Westdeutscher Verlag

Geiger, Theodor (1939a): Sociologi. Grundrids og Hovedproblemer, Kopenhagen

Geiger, Theodor (1939b): Soziologie. Grundriss und Hauptprobleme, § 29ff. Kultursoziologie, Siegen: Ms. [Rohübersetzung des dänischen Buchs durch die U-GH Siegen, die Seitennummerierung beginnt mit jedem Kapitel neu]

Geiger, Theodor (1948/49): Die Klassengesellschaft im Schmelztiegel, Köln und Hagen: Kiepenheuer

Geiger, Theodor (1950): A Radio Test of Musical Taste, in: The Public Opinion Quaterly, Vol. 14, S. 453ff.

Geiger, Theodor (1951): Soziale Umschichtungen in einer dänischen Mittelstadt, Kopenhagen: Acta Juttlandica

Geiger, Theodor (1955): Die Legende von der Massengesellschaft, in: Acta Sociologica I, 1, köbenhavn: Munksgaard, S. 73ff.

Geiger, Theodor (1962a [1931]): Zur Kritik der arbeiter-psychologischen Forschung, in: Geiger, Theodor: Arbeiten zur Soziologie, Neuwied: Luchterhand

Geiger, Theodor (1962b [1955]): Theorie der sozialen Schichtung, in: Geiger, Theodor: Arbeiten zur Soziologie, Neuwied: Luchterhand

Geiger, Theodor (1962c [1955]): Typologie und Mechanik der gesellschaftlichen Fluktuation, in: Geiger, Theodor: Arbeiten zur Soziologie, Neuwied: Luchterhand:

Geiger, Theodor (1964): Demokratie ohne Dogma. Die Gesellschaft zwischen Pathos und Nüchternheit, München: Szczesny-Verlag

Geiger, Theodor (1968 [1953]): Ideologie und Wahrheit. Eine soziologische Kritik des Denkens, Neuwied/Berlin: Luchterhand

Geiger, Theodor (1987a [1926]): Die Masse und ihre Aktion. Ein Beitrag zur Soziologie der Revolutionen, Stuttgart: Enke

Geiger, Theodor (1987b [1932]): Die soziale Schichtung des deutschen Volkes. Soziolgraphischer Versuch auf statistischer Grundlage, Stuttgart: Enke

Geiger, Theodor (1992 [1950]): Die soziale Herkunft der dänischen Studenten, Opladen: Leske & Budrich

Geiling, Heiko (1996): Das andere Hannover. Jugendkultur zwischen Rebellion und Integration in der Großstadt, Hannover: Offizin

Geissler, Rainer (1995): Die Bedeutung Theodor Geigers für die Sozialstrukturanalyse der modernern Gesellschaft, in: Bachmann, Siegfried (Hrsg.): Theodor Geiger - Soziologie in einer Zeit „zwischen Pathos und Nüchternheit“ - Beiträge zu Leben und Werk, Berlin: Duncker & Humblot, S. 273ff.

Geissler, Rainer (1996): Die Sozialstruktur Deutschlands. Zur gesellschaftlichen Entwicklung mit einer Zwischenbilanz zur Vereinigung, Bonn: Bundeszentrale für politische Bildung

Glaser, Hermann (1997): Deutsche Kultur. Ein historischer Überblick von 1945 bis zur Gegenwart, bonn: Bundeszentrale für politische Bildung

Grajczyk, Andreas und Mende, Annette (2000): Nichtnutzer von Online: Zugangsbarrieren bleiben bestehen, in: Media Perspektiven 8/2000, S. 350ff.

Grajczyk, Andreas und Mende, Annette (2001): Nichtnutzer von Online: Internet für den Alltag (noch) nicht wichtig, in: Media Perspektiven 8/2001, S. 398ff.

Grimm, Imre (2002): Menschen als Marken. Dr. Best ist tot - dabei dachten wir, Werbefiguren seien unsterblich, in: Hannoversche Allgemeine Zeitung vom 29. Juni 2002, S. 8

Grötker, Ralf (2001): Zugang macht noch keinen Surfer - Internet-Nichtnutzung in Zahlen: Ein Schnellkurs in Statistik-Deutung, in: http://www.heise.de/tp/deutsch/inhalt/konf/96161.html [21.09.2001]

Grüne, Heinz und Urlings, Stephan (1996): Motive der Onlinenutzung, in: MediaPerspektiven 9/1996, S. 493ff.

G+J Electronic Media Service (2000): Internetnutzung in Deutschland. Analyse der sechsten Erhebungswelle des GFK-Onlinemonitors, in: http//www.ems.guj.de [18.03.2002]

G+J Electronic Media Service (2001): GFK-Onlinemonitor Welle 7, in: http://www.ems.guj.de [18.03.2002]

Hauptmanns, Peter (1999): Grenzen und Chancen von quantitativen Befragungen mit Hilfe des Internet, in: Batinic, Bernad/Werner, Andreas et al.: Online Research. Methoden, Anwendungen und Ergebnisse, Göttingen: Hogrefe

Hoffmann, Jens (2002a): Auf der Suche nach der Struktur des Verbrechens - Theorien des Profilings, in: Musolff, Cornelia/Hoffmann, Jens (Hrsg.): Täterprofile bei Gewaltverbrechen. Mythos, Theorie und Praxis des Profiling, Springer: Berlin, S. 89ff.

Hoffmann, Jens (2002b): Fallanalyse im Einsatz, in: Musolff, Cornelia/ Hoffmann, Jens (Hrsg.): Täterprofile bei Gewaltverbrechen. Mythos, Theorie und Praxis des Profiling, Springer: Berlin, S. 89ff.

Hoffmann, Larissa (2001): Marktforschung und sozialwissenschaftliche Habitusforschung unter besonderer Berücksichtigung der sozialwissenschaftlichen Technikforschung, Hannover: unveröff. Diplomarbeit

Hradil, Stefan (1987): Sozialstrukturanalyse in einer fortgeschrittenen Gesellschaft. Von Klassen und Schichten zu Lagen und Milieus, Opladen: Leske & Budrich

Hradil, Stefan (1998): Die Seismographen der Modernisierung - Singles in Deutschland, in: Aus Politik und Zeitgeschichte, B 53/1998, S. 9ff.

Hradil, Stefan (1999): Soziale Ungleichheit in Deutschland, Opladen: Leske und Budrich

Hurrelmann, Klaus (1998[6]): Einführung in die Sozialisationstheorie - Über den Zusammenhang von Sozialstruktur und Persönlichkeit, Weinheim und Basel: Beltz

ICON ITS (Hrsg.) (1997): Brand Status Handbuch, Nürnberg: Ms.

Inhetveen, Katharina (1997): Musiksoziologie in der Bundesrepublik Deutschland, Opladen: Westdeutscher Verlag

Inselberg, Alfred (2000): Visualizing High Dimensional Datasets & Multivariate Relations. Tutorial for KDD 2000, in: http://delivery.acm.org/10.1145/350000/349102/p33-inselberg.pdf?key1=349102&key2=3536960011&coll=GUIDE&dl=GUIDE&CFID=31649855&CFTOKEN=59618714 [17.11.2004]

Inselberg, Alfred und Dimsdale, Bernard (1990): Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry, in: http://delivery.acm.org/10.1145/950000/949588/p361-inselberg.pdf?key1=949588&key2=0706960011&coll=GUIDE&dl=GUIDE&CFID=31649855&CFTOKEN=59618714 [17.11.2004]

Jann, Ben (2002): Einführung in die Statistik, München: Oldenbourg

Keim, Daniel (1997): Visual Techniques for Exploring Databases, in: http://www.dbs.infomatik.uni-muenchen.de/~daniel\kdd97.pdf

Klein, Naomi (2002): No Logo! Der Kampf der Global Players um Marktmacht. Ein Spiel mit vielen Verlierern und wenigen Gewinnern, Riemann: o. O.

Klein-Bölting, Udo und Busch, Oliver (o. J.): Markenführung im Digital Age, in: http://www.bbdo.de/bbdo-media/eBranding_ukb.pdf [22.01.2002]

Kneip, Ansbert (2000): Wer nicht drin ist, ist draussen, in: http://www.spielgel.de/reporter/0,1518,85860,00.html [08.08.2000]

König, René (1987): Soziologie in Deutschland - Begründer/Verächter/Verfechter, München und Wien: Carl Hanser Verlag

Koller, Barbara et al. (2003): Ältere ab 55 Jahren - Erwerbstätigkeit, Arbeitslosigkeit und Leistungen der Bundesanstalt für Arbeit, Nürnberg: IAB Werkstattbericht 5/2003

Konietzka, Dirk (1995): Lebensstile im sozialstrukturellen Kontext. Zur Analyse soziokultureller Ungleichheiten, Opladen: Leske & Budrich

Kraushaar, Wolfgang (1998): Frankfurter Schule und Studentenbewergung. Von der Flaschenpost zum Molotowcocktail 1946 - 1995, Band I: Chronik, Hamburg: Rogner und Bernhard

Kubicek, Herbert (2001): Internet für alle - zwischen Euphorie und Ignoranz, in: http://www.chancengleichheit-im-netz.de/ de/aktuell/ facts_and_figures.ppt [20.06.2002]

Kundera, Milan (1989): Der Scherz, München - Wien: Carl Hanser Verlag

Lash, Scott (1996 [1994]): Reflexivität und ihre Doppelungen: Struktur, Ästhetik und Gemeinschaft, in: Beck, Ulrich/Giddens, Anthony/Lash, Scott: Reflexive Modernisierung. Eine Kontroverse, Frankfurt: Suhrkamp

Last, Mark (2002a): Lecture No. 2 - The Role of Information Theory in Data Mining, in: http://www.ise.bgu.ac.il/courses/kdd/slides [26.01.2004]

Last, Mark (2002b): Lecture No. 3 - Decision Tree Learning I, in: http:// www.ise.bgu.ac.il/courses/kdd/slides [26.01.2004]

Last, Mark (2002c): Lecture No. 4 - Decision Tree Learning II, in: http:// www.ise.bgu.ac.il/courses/kdd/slides [26.01.2004]

Lepsius, Rainer M. (1966): Parteiensystem und Sozialstruktur: Zum Problem der Demokratisierung der deutschen Gesellschaft, in: Gerhard A. Ritter (Hrsg.): Deutsche Parteien vor 1918, Köln: Kiepenheuer und Witsch

Lessing, Theodor (1995): Die verfluchte Kultur, München: Matthes & Seitz

Lewis, Roger J. (2000): An Introduction to Classification and Regression Tree (CART) Analysis, in: http://www.saem.org/download/ lewis1.pdf [27.01.2004]

Lienert, Gustav A. (1973): Verteilungsfreie Methoden in der Biostatistik, Meisenheim am Glan: Verlag Anton Hain

Lim, Tjen-Sien und Loh, Wei-Yin (2000): A Comparison of prediction Accuracy, Complexity, and Training Time of Thirty-tree old and new Classification Algorithms, in: Machine Learing Vol. 20, S. 203ff.

Loh, Wei-Yin und Shih, Yu-Shan (1997): Split Selection Methodsfor Classification Trees, in: Statistica Sinica, Vol. 7, S. 815ff.

Lovink, Geert (1996): Der Computer: Medium oder Rechner? - Ein [sic!] Begegnung im Netz mit Hartmut Winkler, in: http://www.uni-paderborn.de/~winkler/lovink-9.html [12.03.2001]

Ludwig-Mayerhofer, Wolfgang (1990): Multivariate Logit-Modelle für odrdinalskalierte abhängige Variablen, in: ZA-Information 27, S. 62ff.

Ludwig-Mayerhofer, Wolfgang (1992): Statistik-Software zur Schätzung von Regressions-Modellen für ordinale abhängige Variablen, in: ZA-Information 31, S. 93ff.

Lyotard, Jean-Francois (1986): Das postmoderne Wissen. Ein Bericht, Wien: Böhlau

Maaz, Kai / Ringler, Dominik / Wenzke, Gerhard (2000): Generation N - Kinder und Jugendliche nutzen den Computer und das Internet, in: http://www.sensjs.berlin.de [18.03.2002]

MacElroy, Bill (o. J.): Sieben Techniken der Onlinebefragung, in: http://www.modalis.com/deutsch/news/7forms.html [21.05.01]

Matthiesen, Ulf (Hrsg.) (1998): Die Räume der Milieus. Neue Tendenzen in der sozial- und raumwissenschaftlichen Milieuforschung, in der Stadt- und Raumplanung, Berlin: Edition Sigma

McCormick, Bruce H. et al. (1987): Visualization in Scientific Computing, in: Computer Graphics 21(6)

Mertens, Wim (1983): American Minimal Music. La Monte Young, Terry Riley, Steve Reich, Philip Glass, London: Kahn & Averill

Meyer, Thomas (2001a): Die Soziologie Theodor Geigers - Emanzipation von der Ideologie, Wiesbaden: Westdeutscher Verlag

Meyer, Thomas (2001b): Das Konzept der Lebensstile in der Sozialstrukturforschung - eine kritische Bilanz, in: Soziale Welt Nr. 52, S. 255ff.

Morgaan, James N. und Sonquist, John (1963): Problems in the Analysis of Survey Data, and a Probposal, in: Journal of the American Statistical Association, Vol. 58, S. 415 ff.S.

Naumann, Johannes/Richter, Tobias (2001): Diagnose von Computer Literacy: Computerwissen, Computereinstellungen und Selbsteinschätzungen im multivariaten Kontext, in: http://www.uni-koeln.de/phil-fak/psych/allgemeine/downloads/neumann_richter_in_druck.pdf [28.03.2003]

Neville, Padraic G. (1999): Decision Trees for Predictive Modeling, in: http://stat.bus.utk.edu/datamining/decission%20trees%20for%20predictive%20 modelling%20(neville).pdf [27.01.2004]

Nguyen, Quang Vinh und Huang Mao Lin (2003): Space-Optimized Tree: a connection + [sic!] enclosure approach for the visualization of large hierarchies, in: Information Visualization, 2/2003, S. 3ff.

Nolte, Paul (2000):Die Ordnung der deutschen Gesellschaft. Selbstentwurf und Selbstbeschreibung im 20. Jahrhundert, München: C. H. Beck

Norusis, Marija N. (1998): Guide to Data Analysis, Upper Saddle River/ New Jersey: Prentice-Hall

Ostermaier, Albert (1999): The Making Of. Radio Noir. Stücke, Frankfurt/Main: Suhrkamp

Pampel, Fred (2000): Logistic Regression. A primer, Sage University Papers Series on Quantitativve Applications in the Social Sciences, Thousand Oaks: Sage

Perillieux, René, Bernnat, Rainer und Bauer, Marcus (2000): Digitale Spaltung in Deutschland. Ausgangssituation, internationaler Vergleich, Handlungsempfehlungen, in: http://www.digitale-chancen.de/transfer/downloads/md7.pdf [03.12.2001]

Quinlan, Ross (1993): C4.5: Programs for Machine Learning, Morgan Kaufman: San Mateo

Rammert, Werner (Hrsg.) (1990): Computerwelten - Alltagswelten. Wie verändert der Computer die soziale Wirklichkeit?, Opladen: Westdeutscher Verlag

Rammert, Werner/Böhm, Wolfgang/Olscha, Christa/Wehner, Wolfgang (1991): Vom Umgang mit Computern im Alltag - Fallstudien zur Kultivierung einer neuen Technik, Opladen: Westdeutscher Verlag

Reese-Schäfer, Walter (1988): Lyotard zur Einführung, Hamburg: Junius

Regionales Rechenzentrum für Niedersachsen (1996^2): Internet. Eine Einführung in die Nutzung der Internetdienste, Hannover: Ms.

Reijen, Willem van/Veerman, Dick (1988): Die Aufklärung, das Erhabene, Philosophie, Ästhetik. Interview mit Jean-Francois Lyotard, in: Reese-schäfer, Walter: Lyotard zur Einführung, Hamburg: Junius

Richter, Tobias/Naumann, Johannes/Groeben, Norbert (2001): Das Inventar zur Computerbildung (INCOBI): Ein Instrument zur Erfassung Computer Literacy und computerbezogenen Einstellungen bei Studierenden der Geistes- und Sozialwissenschaften, in: Psychologie in Erziehung und Unterricht, 48, S. 1ff.

Richter, Tobias/Naumann, Johannes/Noller, Stephan (1999): Computer Literacy und computerbezogene Einstellungen: Zur Vergleichbarkeit von Online- und Paper-Pencil-Erhebungen, in: http://www.dgof.de/tband99/pdfs/q-z/richter.pdf [22.06.2001]

Rodax, Klaus (1991): Theodor Geiger - Soziologie der Erziehung. Braunschweiger Schriften 1929 - 1933, Berlin: Duncker & Humblot

Rönsch, Horst Dieter (1973): Hermeneutik, in: Lexikon zur Soziologie, Opladen: Westdeutscher Verlag, S. 273

Scheid, Uwe (1999): Chattende Spieler, surfende Infosucher une shoppende Profis. Entwicklung einer Nutzertypologie für deutschsprachige Internetnutzer, in: http://www.scheid.com (Diplomarbeit) [18.03.2002]

Schelsky, Helmut (1979a [1953]): Die Bedeutung des Schichtungsbegriffs für die Analyse der gegenwärtigen deutschen Gesellschaft, in: Schelsky, Helmut: Auf der Suche nach Wirklichkeit. Gesammelte Aufsätze zur Soziologie der Bundesrepublik, München: Goldmann, S. 326ff.

Schelsky, Helmut (1979b [1956/61]): Gesellschaftlicher Wandel, in: Schelsky, Helmut: Auf der Suche nach Wirklichkeit. Gesammelte Aufsätze zur Soziologie der Bundesrepublik, München: Goldmann, S. 333ff.

Schelsky, Helmut (1979c [1961]): Die Bedeutung des Klassenbegriffs für die Analyse unserer Gesellschaft, in: Schelsky, Helmut: Auf der Suche nach Wirklichkeit. Gesammelte Aufsätze zur Soziologie der Bundesrepublik, München: Goldmann, S. 350ff.

Schelsky, Helmut (1979d [1955]): Über das Restaurative in unserer Zeit, in: Schelsky, Helmut: Auf der Suche nach Wirklichkeit. Gesammelte Aufsätze zur Soziologie der Bundesrepublik, München: Goldmann, S. 410ff.

Schieb, Jörg (o. J.): Aufgepasst, sonst werden Sie obdachlos. Ohne Internet ist der einzelne schon bald verloren, in: http/www.evita.de/artikel/1,,11004,00.html

Schnell, Rainer (1999): Graphisch gestützte Datenanalyse, München: Oldenbourg

Scholz, Joachim (2001): Das Zahlenspiel, in: http://www.emar.de [20.02.2001]

Schroth, Yvonne (1999): Dominante Kriterien der Sozialstruktur. Zur Aktualität der Schichtungstheorie von Theodor Geiger, Münster: Lit-Verlag

Schuck-Wersig, Petra (1999): Das Internet und seine Nutzer. Übersicht über den Stand der Untersuchungen zur Online-Nutzung in Deutschland, in: http://www.kommwiss.fu-berlin.de/~gwersigßforschung/news/nutzerst.htm [20.10.2000]

Schulze, Gerhard (1988): Alltagsästhetik und Lebenssituation. Eine Analyse kultureller Segmentierungen in der Bundesrepublik Deutschland, in: Soeffner, H.-G (Hrsg.): Kultur und Alltag [Soziale Welt, Sonderband 6], Göttingen: Otto Schwartz, S. 71ff.

Schulze, Gerhard (1990): Die Transformation sozialer Milieusin der Bundesrepublik Deutschland, in: Berger Peter/Hradil, Stefan (Hrsg.): Lebenslagen, Lebensläufe, Lebensstile [Soziale Welt, Sonderband 7], Göttingen: Otto Schwartz, S. 409ff.

Schulze, Gerhard (1992): Die Erlebnisgesellschaft. Kultursoziologie der Gegenwart, Frankfurt/Main: Campus

Schwingel, Markus (1993): Analytik der Kämpfe. Macht und Herrschaft in der Soziologie Bourdieus, Hamburg: Argument-Verlag

Serner, Walter (1981): Letzte Lockerung. Ein Handbrevier für Hochstapler, München: Renner

Sevenonemedia (Hrsg.) (o. J.): Semiometrie. der Zielgruppe auf der Spur, in: http://www.sevenonemedia.de [08.05.2002]

Sevenonemedia (Hrsg.) (2002): Semi0metrie: Innovatoren im Fokus, in: http://www.sevenonmedia.de [08.05.2002]

Snyder, Mark (1987): Public Appearances Private Realities. The Psychology of Self-Monitoring, o. O., W. H. Freeman & Company

Spence, Robert (o. J.): Rapid, serial and visual: A presentation technique with potential, in: http://www.iis.ee.ic.ac.uk/~rspence [31.10.2003]

Spence, Robert (2001): Information visualization, New York: ACM

Spiegel-Verlag (Hrsg.) (1997): Online - Offline, Hamburg: Spiegel-Verlag

Spiegel-Verlag (Hrsg.) (2000): Die Internet-Deutschen, in: http://www.spiegel.de/Reporter [01.08.2000]

SPSS (Hrsg.) (o. J.): AnswerTree Algorithm Summary, in: http://stat.bus.utk.edu/datamining/atalgwp_0599.pdf [28.01.2004]

SPSS (Hrsg.) (2001a): Answertree 3.0 User´s Guide, Chikago: o. V.

SPSS (Hrsg.) (2001b): Answertree 3.0 Benutzerhandbuch, Chikago: o. V.

SPSS (Hrsg.) (2002): Clementine 7.0 Users Guide, Chikago: o. V.

Spss Inc. und Recognition Systems Inc. (Hrsg) (1995): Neural Connection 1.0 User's Guide, Chikago: o. V.

Statistisches Bundesamt (Hrsg.) (2003): Datenreport 2002. Zahlen und Fakten über die Bundesrepublik Deutschland, Bonn: Bundeszentrale für politische Bildung

Statistisches Bundesamt (Hrsg.) (2004): Datenreport 2004. Zahlen und Fakten über die Bundesrepublik Deutschland, Bonn: Bundeszentrale für politische Bildung

Sutton, Clifton D. (2003): Understanding CART, in: http://www.galaxy.gmu.edu/stats/syllabi/INFT979.spring2003.html [30.08.2003]

Tenfelde, Klaus (1997): Milieus, politische Sozialisation und Generationenkonflikte im 20. Jahrhundert, Bonn: Gesprächskreis Geschichte (Heft 19)

Thearling, Kurt et al. (2001): Visualizing Data Mining Models, in: Fayyad, Usama et al. (Hrsg.): Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufman: o. O.

Theus, Martin (1996): Theorie und Anwendung interaktiver statistischer GrafikAugsburg: Wissner

TMS Emnid (2001): Der Verweigerer-Atlas - Basiserhebung, Hamburg: Ms.

TMS Emnid (2002): (N)Onliner - Atlas 2002. Eine Typographie des digitalen Grabens durch Deutschland, Hamburg

Trappe, Paul (1978): Theodor Geiger, in: Käsler, Dirk (Hrsg.): Klassiker des soziologischen Denkens, Zweiter Band: von Weber bis Mannheim, München: Beck, S. 254ff.

Trappe, Paul (1993): Begrüssung der Gäste und Einführung zum Gedenksymposium, in: Fazis, Urs/Nett, Jachen C. (Hrsg.): Gesellschaftstheorie und Normentheorie. Theodor Geiger Symposium, Basel

Tufte, Edward (1983): The visual display of Quantitative Information, Cheshire, Conneticut: Graphics Press

Tufte, Edward (1990): Envisioning Information,Cheshire, Conneticut: Graphics Press

Tufte, Edward (1997): Visual Explanations. Images and Quantities, Evidence and Narrative, Cheshire, Coneticut: Graphics Press

Turkle, Sherry (1986 [1984]). Die Wunschmaschine. Der Computer als zweites Ich, Reinbek: Rowohlt

Turkle, Sherry (1999 [1995]): Leben im Netz. Identität im Zeitalter des Internet, Reinbek: Rowohlt

Two Crows Corporation (Hrsg.) (1999): Introduction to Data Mining and Knowlede Discovery, in: http://www.twocrows.com/intro-dm.pdf [29.08.2003]

Urban, Dieter (1993): Logit-Analyse. Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen, Stuttgart: Gustav Fischer Verlag

Vester, Michael (1994): Die verwandelte Klassengesellschaft. Modernisierung der Sozialstruktur und Wandel der Mentalitäten in Westdeutschland, in: Mörth, Ingo und Fröhlich, Gerhard (1994): Das symbolische Kapital der Lebensstile. Zur Kultursoziologie der Moderne nach Pierre Bourdieu, Frankfurt/New York: Campus, S. 129ff.

Vester, Michael (1995): Deutschlands feine Unterschiede. Mentalitäten und Modernisierung in Ost- und Westdeutschland, in: Aus Politik und Zeitgeschichte B 20/95 vom 12. Mai 1995, Bonn: Bundeszentrale für politische Bildung, S. 16ff.

Vester, Michael (1997b): Soziale Milieus und Individualisierung. Mentalitäten und Konfliktlinien im historischen Wandel, in: Beck, Ulrich/ Sopp, Peter (Hrsg.): Individualisierung und Integration. Neuer Integrationismus oder neue Konfliktlinien?, Opladen: Leske und Budrich, S. 102 - 125

Vester, Michael (1998): Klassengesellschaft ohne Klassen. Auflösung oder Transformation der industriegesellschaftlichen Sozialstruktur?, in: Berger, Peter A./Vester, Michael (Hrsg.): Alte Ungleichheiten - neue Spaltungen, Opladen: Leske und Budrich, i. E.

Vester, Michael/Clemens, Bärbel/Geiling, Heiko/Herrmann, Thomas/ Müler, Dagmar/von Oertzen, Peter (1989^2): Der Wandel der Sozialstruktur und die Entstehung neuer gesellschaftlich-politischer sozialer Milieus in der Bundesrepublik Deutschland. Forschungsantrag an die Stiftung Volkswagenwerk vom 12. Oktober 1987, Hannover: Ms.

Vester, Michael/von Oertzen, Peter/Geiling, Heiko/Hermann, Thomas/Müller, Dagmar (1992): Neue soziale Milieus und pluralisierte Klassengesellschaft. Endbericht des Forschungsprojektes "Der Wandel der Sozialstruktur und die Entstehung neuer gesellschaftlich-politischer Milieus", Hannover: Ms.

Vester, Michael/von Oertzen, Peter/Geiling, Heiko/Hermann, Thomas/Müller, Dagmar (1993): Soziale Milieus im gesellschaftlichen Strukturwandel. Zwischen Integration und Ausgrenzung, Köln: Bund-Verlag

Vester, Michael/von Oertzen, Peter/Geiling, Heiko/Hermann, Thomas/Müller, Dagmar (2001): Soziale Milieus im gesellschaftlichen Strukturwandel. Zwischen Integration und Ausgrenzung, Frankfurt: Suhrkamp

Vester, Michael und Schwarzer, Thomas (1996): Arbeitnehmermentalitäten in technologischen Revolutionen, Hannover: Ms.

Ward, Matthew O. (1994): XmdvTool: integrating multiple methods for visualizing multivariate data, in: http://delivery.acm.org/10.1145/960000/951146/p326-ward.pdf?key1=951146&key2=6032070011&coll=GUIDE&dl=GUIDE&CFID=31402930&CFTOKEN=10285024

Wegman, Edward J. (1999): Data Mining and Visualization: Some Strategies, in: http://www.stat.fi/isi99/proceedings/arkisto/varastowegm1028.pdf [31.08.2003]

Wegman, Edward J. (2003a): Statistical Data Mining of Massive Data, Lecture Administrative, in: http://www.galaxy.gmu.edu/stats/syllabi/INFT979.spring2003.html [30.08.2003]

Wegman, Edward J. (2003b): Statistical Data Mining of Massive Data, Lecture Visual Complexity, in: http://www.galaxy.gmu.edu/stats/syllabi/INFT979.spring2003.html [30.08.2003]

Wegman, Edward J. (2003c): Statistical Data Mining of Massive Data, Lecture Artificial Neural Networks, in: http://www.galaxy.gmu.edu/stats/syllabi/INFT979.spring2003.html [30.08.2003]

Wiggershaus, Rolf (1988): Die Frankfurter Schule. Geschichte - Theoretische Entwicklung - politische Bedeutung, München: dtv

Wilkinson, Leland (1992): Tree Structured Data Analysis: AID, CHAID and CART, in: http://stat.bus.utk.edu/datamining/tree%20structured%20data%20analysis%20(spss).pdf [28.01.2004]

Witten, Ian und Frank, Eibe (2001): Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen, München: Hanser

LITERATURVERZEICHNIS

B. Onlinequellen

Die Onlinequellen beziehen sich auf spezielle Links, die für die Arbeit verwendet wurden, die nicht im Literaturverzeichnis unter A genannt sind und die für die Erstellung der Arbeit benutzt wurden. Dabei wurde auf allgemeine Links, wie z. B. Suchmaschinen, verzichtet. Um dem Leser die Inhalte der Seiten zu verdeutlichen, wurde eine Kurzbeschreibung angefügt.

http://www.gesis.org/za, http://www.gesis.org/zuma [02.08.2000]

Die ZUMA in Mannheim ist Ansprechpartner für sozialwissenschaftliche Fragestellungen, das Zentralarchiv für empirirische Sozialforschung in Köln ist eine Stelle, die Datensätze archiviert und der interessierten Öffentlichkeit für Sekundäranalysen zur Verfügung stellt.

http://davis.wpi.edu/~xdmv/ [28.10.2004]

Homepage des Xdmv-Tool-Projekts, einem Programm zur Darstellung multivariater Grafiken.

http://www.xlstat.com [20.10.2004]

Homepage des Excel-Statistik-Pakets XLSTAT, einem Zusatztool zu Excel, das sowohl deskriptive als auch multivariate Statistiken und Grafiken beinhaltet.

http://www.online-forschung.de [01.02.2001]

Portalseite zur Onlineforschung mit zahlreichen Links, Gastartikeln, ein Link zur GIRL-Mailingliste (deutschsprachige Mailiingliste zur Onlineforschung).

http://www.rulequest.com [28.01.2004]

Auf dieser Seite gibt es Informationen, Download- und Bestellmöglichkeiten für den von Ross QUINLAN entwickelten C4.5-Algorithmus

http://stats.math.uni-augsburg.de/mondrian/mondrian.html [16.06.2005]

Homepage des Mondrian-Projekts in Deutschland mit Downloadmöglichkeiten des jeweils aktuellen Programms für unterschiedliche Betriebssysteme.

http//spitswww.uvt.nl/~vermunt [28.10.2004]

Homepage von J. K. VERMUNT, dem Entwickler des LEM-Programms. Auf dieser Seite findet sich auch weiterführende Literatur zu dem Thema.

LEBENSLAUF

PERSÖNLICHE DATEN

Name:	Stefan Lebert
Geburtstag/-ort:	31. Januar 1966 in Nürnberg
Familienstand:	geschieden
Staatsangehörigkeit:	deutsch

SCHULAUSBILDUNG

1972 - 1981	Grund- /Hauptschule in Nürnberg und Wendelstein
1981 - 1983	Wirtschaftsschule Fürth
1983 - 1986	FOS Nürnberg, Wirtschaft

BERUFLICHER WERDEGANG

1986 - 1992	Studium der Betriebswirtschaft an der FH Nürnberg
1992 - 1999	Studium der Sozialwissenschaften an den Universitäten ERlangen - Nürnberg und Hannover
1999 - 2000	Marktforschung
2000 - 2005	Wissenschaftlicher Mitarbeiter am Lehrstuhl für Soziologie und empirische Sozialfrorschung der Universität Augsburg