Diplomarbeit

Aufbereitung des Sozio-ökonomischen Panels für die Datenanalyse

Tobias Malbrecht

Diplomarbeit
an der Fakultät für Informatik
der Technischen Universität Dortmund

Dortmund, den 11. März 2008

Betreuer:

Prof. Dr. Katharina Morik
Dipl.-Inform. Ingo Mierswa

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die empirische Forschung nimmt in vielen Wissenschaften eine wichtige und in ihrer Bedeutung zunehmende Rolle ein. Dies gilt auch im Bereich der Sozial- und Wirtschaftswissenschaften. Die empirische sozio-ökonomische Forschung liefert Erkenntnisse, die dazu beitragen, gesellschaftliche Probleme aufzudecken, ihre Gründe und Ursachen zu erforschen und zu erkennen sowie geeignete politische Maßnahmen, seien es nun gesellschafts-, sozial-, bildungs- oder wirtschaftspolitische, zu initiieren und erfolgreich durchzuführen. Die empirische sozio-ökonomische Forschung ist vielfach in der Lage, aus der Betrachtung vergangener Vorgänge Muster zu erkennen und aus diesen Modelle abzuleiten oder aber theoretisch erstellte Modelle empirisch zu untermauern bzw. zu falsifizieren.

Paneldatensätze sind dabei in vielen Fällen Grundlage und wichtige Basis der empirischen sozial- und wirtschaftswissenschaftlichen Forschung. Vor allem Haushaltspaneldatensätze, in denen Daten privater Haushalte und einzelner Personen, also Daten einzelner Wirtschaftssubjekte erfasst werden, werden in diesen Wissenschaftsbereichen häufig verwendet. Mittlerweile existiert in vielen Ländern mindestens einer dieser Datensätze, wie der von Haisken-DeNew (2001) gegebene Überblick über die weltweit wichtigsten dieser Haushaltspaneldatensätze zeigt. Unter den Haushaltspaneldatensätzen ist in Deutschland vor allem das Sozio-oekonomische Panel (SOEP) bekannt und wird weltweit zu empirischer Forschung hinsichtlich wirtschaftlicher und soziologischer Fragestellungen verwendet.

Methodisch bedient sich die empirische Forschung auf Paneldatensätzen zumeist spezialisierter statistischer Modelle und darauf basierender ökonometrischer Verfahren. Beispiele hierfür sind etwa die logistische Regression bzw. die Cox-Regression[1], die etwa von Winkelmann und Winkelmann (1998) und Biewen und Wilke (2005) bei ihren empirischen Untersuchungen zu Aspekten der Arbeitslosigkeit verwendet wurden. Verfahren des maschinellen Lernens wurden und werden auf Paneldaten - vor allem auch auf sozio-ökonomischen Paneldaten - hingegen kaum angewendet. Dies ist insofern bedauerlich, als dass auch maschinelle Lernverfahren oft eine detaillierte, explorative Aufdeckung von Mustern und Zusammenhängen in Daten und die Bildung intuitiv verständlicher Modelle gestatten.

Die vorliegende Arbeit untersucht daher, wie Verfahren des Data Mining und des maschinellen Lernens auf sozio-ökonomischen Paneldaten angewendet werden können. Sie besitzt damit einen interdisziplinären Fokus im Spannungsfeld zwischen Sozial- und Wirtschaftswissenschaft, Statistik und Informatik (vgl. Abbildung 1.1). Durch die angedeutete Dominanz ökonometrischer Verfahren in den empirischen Sozial- und Wirtschaftswissenschaften steht in dieser Beziehung die Informatik bislang eher im Abseits - trotz des großen Wachstums der aus diesem Bereich erwachsenen Anwendung von Data-Mining-Verfahren, wie es etwa von Mitchell (1999) zutreffend erkannt wird. Diese Arbeit soll zum einen dazu beitragen, die Rolle der Informatik innerhalb der Sozial- und Wirtschaftswissenschaften zu stärken,

[1] Die logistische Regression ist ein Verfahren zur Bestimmung des Einflusses von Variablen auf eine binäre, abhängige Variable. Sie wird u.a. von Kleinbaum und Klein (2002) erläutert. Die Cox-Regression ermöglicht die Bestimmung von Einflüssen auf die Verweildauer in Zuständen, etwa den der Arbeitslosigkeit. Sie geht auf Cox (1972) zurück.

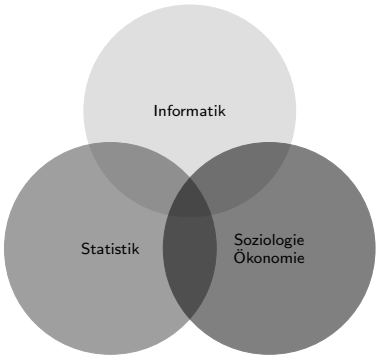


Abbildung 1.1: Interdisziplinärer Fokus der Arbeit

indem auch Methoden aus dem Bereich des Data Mining bzw. maschinellen Lernens auf Daten aus diesen Wissenschaften erfolgreich angewendet und damit für eine Anwendung in Forschungsvorhaben propagiert werden. Zum anderen soll jedoch auch der Informatik die explorative sozial- und wirtschaftswissenschaftliche Forschung als neues Anwendungsgebiet informatischer Methoden vorgeschlagen werden.

Um dieser Intention gerecht zu werden, wird in dieser Arbeit exemplarisch die Bearbeitung einer sozio-ökonomischen Fragestellung anhand der Anwendung informatischer Datenanalysemethoden auf einem Haushaltspaneldatensatz durchgeführt. Diese Arbeit konzentriert sich dabei zum einen auf die wesentlichen Schritte, die vor der eigentlichen Anwendung von Datenanalyseverfahren stehen und zu deren Durchführung unabdingbar sind. Dies umfasst vor allem die Auswahl der Daten, ihre Vorverarbeitung und Transformation. Ein weiterer Schwerpunkt dieser Arbeit ist dann die Anwendung exemplarisch ausgewählter Data-Mining-Verfahren bzw. maschineller Lernverfahren.

Als Datensatz anhand dessen die Vorbereitungen und Analysen exemplarisch beschrieben werden, wurde das bereits erwähnte Sozio-oekonomische Panel herangezogen, da es aufgrund des mittlerweile sehr langen Beobachtungszeitraumes und der enormen Abdeckung verschiedener Themenbereiche eine in Deutschland herausragende Stellung einnimmt und auch weltweit in der sozio-ökonomischen Forschung viel Beachtung und Verwendung erfährt. Die in der Arbeit beschriebene Vorgehensweise bei Vorverarbeitung und Analyse orientiert sich daher auch am SOEP und den daraus resultierenden speziellen Gegebenheiten. In den meisten Fällen ist die Verfahrensweise jedoch zumindest für Paneldaten allgemein gültig und daher leicht auch auf andere Paneldatensätze übertragbar.

Neben der Selektion eines Datensatzes musste außerdem eine Selektion einer sozio-ökonomischen Fragestellung, d.h. der zu analysierenden Themenaspekte, erfolgen, auf die in dieser Arbeit exemplarisch fokussiert werden sollte. Thematisch wurde hier die *Arbeitslosigkeit* aufgegriffen. Dies ist durch folgende Punkte zu begründen. Das Thema *Arbeitslosigkeit* hat in den meisten Wirtschaftsnationen eine hohe gesellschaftliche Relevanz und prägt vielfach die politische Diskussion. Dies ist auch und in besonderem Maße in Deutschland der Fall. Eine hohe Relevanz ist insofern nicht verwunderlich, als dass gerade die Arbeitslosigkeit zum einen für Individuen gravierende negative Folgen finanzieller, sozialer und psychologischer Natur mit sich bringt und zum zweiten auch volkswirtschaftlich als problematisch anzusehen ist. Um dem Problem Arbeitslosigkeit entgegenzuwirken, ist es daher zunächst von Nöten, bestimmende Einflussfaktoren der Arbeitslosigkeit zu identifizieren.

Es sollten dazu sowohl Faktoren auf Mikroebene, d.h. das einzelne Wirtschaftssubjekt betreffende Faktoren wie möglicherweise der Bildungsgrad einer Person, als auch Faktoren auf Makroebene, d.h. gesamtwirtschaftliche Faktoren wie z.B. Konjunkturschwankungen, erkannt werden, die einen bestimmenden Einfluss auf die Arbeitslosigkeit haben. Anschließend können daraus Schlüsse gezogen werden, welche Maßnahmen ergriffen werden müssten, um der Arbeitslosigkeit entgegenzuwirken.

Auf methodischer Ebene ist das Thema Arbeitslosigkeit insofern interessant, als dass es sich bei der Arbeitslosigkeit um ein dynamisches Phänomen handelt. Somit unterliegen nicht nur die potentiell zu erkennenden Einflussfaktoren einem zeitlichen Verlauf, sondern ebenfalls die zu untersuchende Zielgröße. Wie in dieser Arbeit zu verfolgen ist, erhöht dies die Komplexität des Analyseprozesses signifikant.

1.1 Aufbau der Arbeit

Die empirische Forschung versucht, durch Beobachtung Wissen über die reale Welt zu erwerben. Da ein direkter Erwerb dieses Wissens aufgrund der Komplexität der realen Welt meist nicht möglich ist, bedient sich die empirische Forschung eines Umweges, der sich grob in drei Schritte oder Phasen unterteilen lässt, wie in Abbildung 1.2 anschaulich zu erkennen ist. Diese Phasen sind zum einen die Datenerhebung, die Analyse der Daten und die

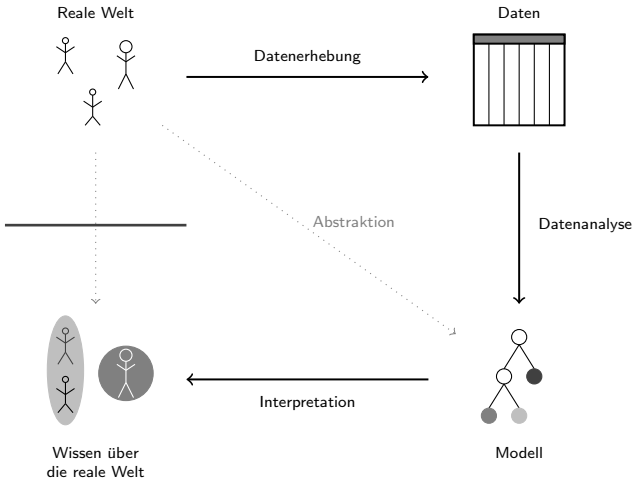


Abbildung 1.2: Empirischer Erwerb von Wissen über die reale Welt

Interpretation der Ergebnisse der Datenanalyse. Auch der Aufbau dieser Arbeit ist anhand dieser drei Schritte gut zu verdeutlichen. Zu Beginn erfolgt in Kapitel 2 eine grundlegende Einführung in Aspekte von Datenerhebungen im Bereich sozio-ökonomischer Forschung unter besonderer Berücksichtigung von Paneldaten. Kapitel 3 stellt danach das in dieser Arbeit betrachtete Sozio-oekonomische Panel vor und geht dabei vor allem auf die Methodik der Studie und die Struktur der in der Studie erhobenen Daten, die die Grundlage dieser Arbeit bilden, ein. Dieser Teil der Arbeit stellt damit vor allem für Leser aus dem Bereich der Informatik die Grundlagen von Datenerhebungen aus den fachfremden Sozial- und Wirtschaftswissenschaften sowie des Sozio-oekonomischen Panels bereit.

Der anschließende Hauptteil der Arbeit beschreibt dann die angekündigte exemplarische

Bearbeitung einer ökonomischen Fragestellung unter Einsatz von Verfahren der Datenanalyse. Kapitel 4 grenzt dazu die in dieser Arbeit untersuchte Fragestellung inhaltlich ein und umreißt die daraus resultierenden Datenanalyseaufgaben. Die folgenden Kapitel widmen sich dann der Analyse der Daten des SOEP hinsichtlich der definierten Aufgaben in Bezug auf die ausgewählte Fragestellung. Die eigentliche Analyse - d.h. die Anwendung eines Datenanalyseverfahrens bzw. das Data Mining im engeren Sinn - ist dabei lediglich ein Teilschritt eines umfassenden Prozesses, den Fayyad et al. (1996) als *Wissensentdeckung in Datenbanken* (*engl.* Knowledge Discovery in Databases) einführte, und der in Abbildung 1.3 dargestellt ist. Auch die in dieser Arbeit durchgeführte Analyse von Paneldaten des

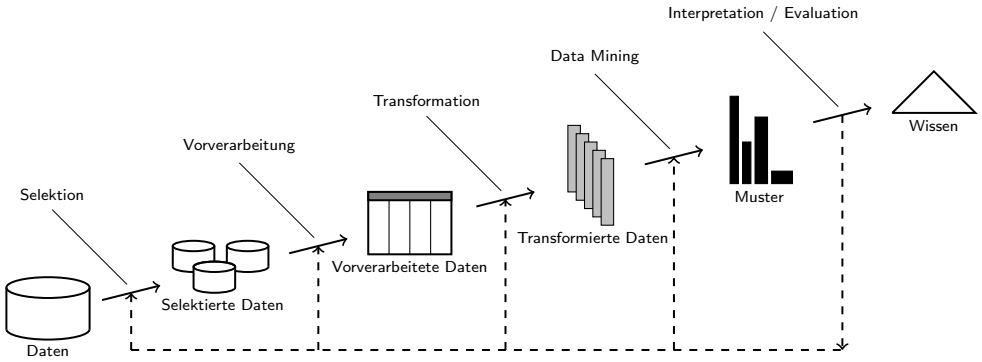


Abbildung 1.3: Prozess der Wissensentdeckung in Datenbanken (nach Fayyad et al. (1996))

SOEP erfordert die explizite Berücksichtigung und Durchführung aller Komponenten dieses Prozesses. Die Strukturierung des Prozesses eignet sich daher in hohem Maße für die Strukturierung des Hauptteils dieser Arbeit. Kapitel 5 erläutert, wie Daten aus dem SOEP für die vorgestellten Analyseaufgaben selektiert, aus dem SOEP-Datensatz extrahiert und zur gemeinsamen Analyse verknüpft werden. Darauf aufbauend erklärt Kapitel 6 die weiterhin notwendigen Schritte zur Vorverarbeitung der Daten und erläutert die sinnvolle bzw. für die im Rahmen dieser Arbeit verwendeten Analyseverfahren sogar notwendige Transformation der Paneldaten. Auf die folgende Datenanalyse hinleitend, beschreibt Kapitel 7 kurz die vorverarbeiteten Daten und analysiert deskriptiv zeitliche Effekte hinsichtlich der zu bearbeitenden Analyseaufgaben. Auch Kapitel 8 leistet vornehmlich einen deskriptiven Beitrag, indem Gewichte einzelner Einflussgrößen hinsichtlich des Themenaspekts Arbeitslosigkeit mit unterschiedlichen Verfahren berechnet werden. Kapitel 9 und 10 haben dann die Beschreibung der eigentlichen Anwendung von Datenanalyseverfahren zum Inhalt. Sie erläutern die verwendeten Verfahren des Data Mining bzw. des maschinellen Lernens, die zur Erkennung von Zusammenhängen hinsichtlich der betrachteten Fragestellung herangezogen wurden, und präsentieren die bei der Anwendung der Verfahren erzielten Ergebnisse.

Den letzten Teil dieser Arbeit bildet Kapitel 11, welches eine Interpretation der vorherig vorgestellten Ergebnisse enthält. Dieses Kapitel leitet außerdem ein Fazit dieser Arbeit ab und nennt Ideen und Vorschläge zur weiteren Vertiefung der in dieser Arbeit verfolgten Betrachtungen.

1.2 Verwendete Werkzeuge

Eine Analyse von Daten, wie sie im Rahmen dieser Arbeit durchgeführt wurde, obliegt dem Einsatz geeigneter Werkzeuge, d.h. geeigneter Software. In Bezug auf Datenanalyseprozesse liegt mit der frei verfügbaren Software RapidMiner (früher Yale) eine umfassende Umgebung vor. Diese ist durch ihr modulares Operatorkonzept in der Lage, den gesamten oben erwähnten Datenanalyseprozess durch entsprechende Komponenten zu unterstützen und gewährleistet zudem eine leichte Adaptabilität hinsichtlich veränderter Anwendungsintentionen bzw. hinsichtlich der Anwendung unterschiedlicher Analyseverfahren. Eine Beschreibung der Funktionalität von RapidMiner und einiger exemplarischer Datenanalyseszenarien findet sich in Mierswa et al. (2006). Speziell bezüglich der Bearbeitung von Paneldaten fehlte RapidMiner jedoch bislang die dafür notwendige Funktionalität. Dies war vor allem im Bereich der Vorverarbeitung - hauptsächlich bei der Transformation - von Paneldaten der Fall. Im Rahmen dieser Arbeit wurde RapidMiner daher in Form eines Plugins um die benötigten Funktionen erweitert. Die in diesem sogenannten *Panel-Plugin* enthaltenen Operatoren zur Behandlung von Paneldaten werden später im Kontext der Schritte im Datenbearbeitungs- und Analyseprozess (hauptsächlich in der Erläuterung der Vorverarbeitung der Daten in Kapitel 6) beschrieben. Einführende Beschreibungen verwendeter Operatoren werden, wie an dieser Stelle zu sehen, durch ein Symbol am Rand des beschreibenden Textes gekennzeichnet. Darüber hinaus bietet Anhang C eine zusammenfassende Aufstellung der für diese Arbeit implementierten und im Panel-Plugin zusammengefassten Operatoren und erläutert stichpunktartig deren Funktionsweise und Nutzung.

Um die SOEP-Daten in RapidMiner analysieren zu können, bedurfte es zudem einer Lösung, um einzelne Daten aus dem SOEP-Datensatz zu selektieren und zu extrahieren und in einer von RapidMiner lesbaren Form zu speichern. Auch hierfür wurde im Rahmen dieser Arbeit mit dem Programm *PanelX* eine geeignete Lösung implementiert, die zudem leicht für weitere Paneldatensätze oder andere Ausgabeformate und damit Analysewerkzeuge angepasst werden kann. Eine Beschreibung des Programms PanelX erfolgt in Kapitel 5.2 in Bezug auf die Selektion und Extraktion von Daten aus dem SOEP-Paneldatensatz.

1.3 Datenbasis

Diese Arbeit basiert auf den Daten des Sozio-ökonomischen Panels gemäß der Datenlieferung aus dem Jahr 2005, die Daten bis einschließlich Welle U, d.h. bis zum Jahr 2004 enthält. Alle folgenden Ausführungen beschreiben die Daten und das SOEP bis zu diesem Zeitpunkt. Neuere Entwicklungen und Veränderungen des SOEP (beispielsweise hinsichtlich der Biographiedaten zu Kindern unterschiedlichen Alters) und der resultierenden Daten sind nicht einbezogen. Allerdings können alle angewendeten Betrachtungen und Verfahren leicht auf neue Datenlieferungen des SOEP, aber auch auf andere Paneldatensätze übertragen werden.

2 Grundlagen sozio-ökonomischer Empirie

Wie aus Abbildung 1.2 hervorgeht, sind neben der zentralen Komponente Datenanalyse die Datenerhebung sowie die Interpretation der bei der Datenanalyse erstellten Modelle integrale Bestandteile empirischer Forschung. Dieses Kapitel befasst sich grundlegend mit der ersten Komponente empirischer Forschung, nämlich der Erhebung von Daten. Dazu werden einige Aspekte der Erhebung von Daten zunächst allgemein betrachtet, bevor der Fokus auf den Bereich der Datenerhebung im Umfeld sozio-ökonomischer Forschung gelegt wird. Eine weitere Spezialisierung erfolgt anschließend mit der besonderen Berücksichtigung von Panelstudien und Paneldaten.

Bezüglich der Erhebung von Daten muss generell im Wesentlichen die folgende Frage beantwortet werden: *Welche* und *wessen* Daten werden *wann wie* erhoben? Die Beantwortung dieser Fragen ist leider keineswegs trivial. Vielmehr spielen dabei eine Reihe unterschiedlicher, jedoch teilweise interagierender Aspekte eine Rolle. Die Frage, *welche* Daten erhoben werden, ist zum einen natürlich bestimmt durch die inhaltliche und thematische Ausrichtung des Forschungsvorhabens. Zum anderen spielt im Zusammenhang mit dieser Frage auch die Art und Weise der Übersetzung von Eigenschaften empirischer Objekte in Daten eine Rolle. Diese Übersetzung wird als *Messung* bezeichnet, mit der sich die Messtheorie auseinandersetzt. Da die korrekte Messung von Objekteigenschaften großen Einfluss auf die Daten und die Datenanalyse haben kann, erfolgt in Kapitel 2.1 eine Erläuterung der Entstehung von Daten durch Messung von Objekteigenschaften. Außerdem wird einführend erläutert, wie Objekte zur Messung ihrer Eigenschaften ausgewählt werden, und es wird eine allgemeine Definition der aus Messungen resultierenden Daten gegeben. Danach stellt Kapitel 2.2 verschiedene Formen (Designs) von Datenerhebungen, die vorrangig in der sozioökonomischen Forschung Beachtung finden, vor. Hierbei sind vor allem der Erhebungszeitpunkt (*Wann* werden Daten erhoben?) sowie die Menge der für die Erhebung ausgewählten Untersuchungseinheiten (*Wessen* Daten werden erhoben?) relevant. Ebenso wird der Aspekt der Erhebungsmethode (*Wie* werden Daten erhoben?) beleuchtet. Die Erörterung der Auswahl von Objekten zur Erhebung ihrer Daten wird in Bezug auf sozio-ökonomische Datenerhebungen vertieft, und verschiedene Verfahren zur Durchführung einer solchen Auswahl werden erläutert. Anschließend werden in Kapitel 2.3 häufig vorkommende Formen sozio-ökonomischer Daten, unter anderen auch Paneldaten, definiert und der Einfluss des gewählten Erhebungsdesigns auf die Form der Daten beschrieben. Kapitel 2.4 definiert dann zentrale Begriffe im Zusammenhang mit Panelstudien sowie Paneldaten und stellt mögliche Repräsentationsformen von Paneldaten vor. Kapitel 2.5 erläutert Probleme, die bei Panelstudien auftreten, und die teilweise eine Berücksichtigung seitens des Analysten der aus den Studien resultierenden Paneldaten erfordern.

2.1 Entstehung und Form von Daten

Grundsätzlich beschreiben Daten Eigenschaften von Objekten der realen Welt[1]. Diese Objekte werden als *Merkmalsträger* und deren Eigenschaften demzufolge als *Merkmale* bezeichnet. In der Statistik werden Merkmalsträger häufig auch als *Untersuchungseinheit* (*engl.* unit) bezeichnet. Im Bereich des Data Mining sind für Merkmalsträger und Merkmale vor allem die Synonyme *Instanz* und *Attribut* gebräuchlich. Die Merkmale bzw. Attribute können je nach Eigenschaft der zugehörigen Objekte verschiedene Ausprägungen annehmen, die sogenannten *Merkmalsausprägungen*. Für die Datenanalyse ist dabei vor allem von Bedeutung, wie diese Merkmalsausprägungen als Daten repräsentiert werden, d.h. wie Eigenschaften von empirischen Objekten *gemessen* werden können. Dieser Fragestellung widmet sich der folgende Abschnitt.

2.1.1 Messung von Eigenschaften empirischer Objekte

Vor allem aufbauend auf Arbeiten aus dem Bereich der Psychologie formalisiert die *Messtheorie*, wie Eigenschaften von empirischen Objekten und Relationen zwischen diesen Objekten, die in der Realität existieren, durch Daten abgebildet werden können. Stevens (1946) definiert *Messung* als "Zuordnung von Zahlen zu Objekten nach bestimmten Regeln". Dieser Gedanke wurde dahingehend weiterentwickelt, dass Zahlen (oder Symbole) Objekten so zugeordnet werden sollen, dass Relationen, die für die realen Objekte gelten, durch Relationen der Zahlen bzw. Symbole untereinander widergespiegelt werden. Dies soll einführend folgendermaßen formalisiert werden, wobei diese Einführung auf Pfanzagl (1971) und Roberts (1979) basiert. Auf diese sei auch für sehr viel umfassendere Betrachtungen der Messtheorie verwiesen.

Definition 2.1 (Relativ) *Seien A eine Menge von Objekten und $R_1, \ldots, R_p$ (nicht notwendigerweise binäre) Relationen auf der Menge A. Dann wird das $(n+1)$-Tupel*

$$\mathcal{A} = (A, R_1, R_2, \ldots, R_p)$$

als Relativ *bezeichnet. Besteht $\mathcal{A}$ aus einer Menge empirischer Objekte und empirisch beobachteten Relationen zwischen diesen Objekten, so heißt $\mathcal{A}$* empirisches Relativ*. Besteht $\mathcal{A}$ aus einer Menge numerischer Objekte bzw. Symbole und entsprechenden Relationen auf dieser Menge, so wird $\mathcal{A}$* numerisches Relativ *genannt.*

Ein Relativ besteht demnach aus einer Menge und den auf dieser Menge geltenden Relationen. Ein empirischer Relativ besteht speziell aus einer Menge empirischer Objekte, wobei zwischen diesen Objekten Relationen gelten, die durch empirisch beobachtbare Eigenschaften der Objekte definiert werden. Als Beispiel einer solchen Menge empirischer Objekte sei eine Menge von drei Personen, hier als o_1, o_2 und o_3 bezeichnet, gegeben. Das zu untersuchende Merkmal der Personen sei deren Geschlecht. Jede Person kann entweder weiblich oder männlich sein, also genau einer der beiden Klassen Frau bzw. Mann angehören. Bzgl. des Geschlechts können Personen nur hinsichtlich der Zugehörigkeit zu diesen beiden Klassen verglichen werden. Durch das Merkmal Geschlecht wird dementsprechend eine Äquivalenzrelation $\sim$ definiert. Ein Beispiel eines empirischen Relativs ist

[1] Der Begriff *Objekt* ist hier nicht zwangsläufig gegenständlich gemeint. So könnte beispielsweise auch ein Staat als Objekt und dessen Bruttoinlandsprodukt als seine Eigenschaft gemeint sein. Häufig wird daher der Begriff *empirisches Objekt* verwendet.

demnach $(\{o_1, o_2, o_3\}, \sim)$. Ein numerisches Relativ ist beispielsweise die Menge der reellen Zahlen zusammen mit der Relation $>$, also das Tupel $(\mathbb{R}, >)$.

Bei einer Messung von Eigenschaften empirischer Objekte muss nun eine Abbildung gefunden werden, sodass die Relationen eines empirischen Relativs auf ein numerisches Relativ übertragen werden. Dies geschieht mittels einer Abbildung. Dazu sei wie folgt definiert:

Definition 2.2 (Skala) *Seien ein empirisches Relativ $\mathcal{A} = (A, R_1, \ldots, R_p)$ und ein numerisches Relativ $\mathcal{B} = (B, S_1, \ldots, S_p)$ gegeben. Sei ferner $\varphi : A \to B$ eine Abbildung. Definiert φ einen* Homomorphismus *von $\mathcal{A}$ in $\mathcal{B}$, d.h. gilt für alle $(a_1, \ldots, a_k) \in A^k$ und $i = 1, \ldots, p$, dass*

$$(a_1, \ldots, a_k) \in R_i \quad \Leftrightarrow \quad (\varphi(a_1), \ldots, \varphi(a_k)) \in S_i,$$

dann heißt das Tripel $S = (\mathcal{A}, \mathcal{B}, \varphi)$ Skala.

Eine Skala beinhaltet also eine Abbildung φ, die sicherstellt, dass alle Relationen aus dem empirischen Relativ ins numerische Relativ übertragen werden.

Skalen unterscheiden sich häufig durch die Anzahl der im empirischen Relativ enthaltenen Relationen, die durch eine geeignete Abbildung auf das numerische Relativ übertragen werden. Man sagt, sie unterscheiden sich durch ihr sogenanntes *Mess-* oder auch *Skalenniveau*. In Abhängigkeit von den zwischen den ursprünglichen empirischen Objekten geltenden Relationen werden verschiedene Skalenniveaus definiert. Eine klassische Einteilung in Skalenniveaus bzw. Skalentypen geht auf Stevens (1946) zurück. Sie ist in Tabelle 2.1 dargestellt.

Tabelle 2.1: Merkmalsskalen und ihre Eigenschaften

Skala	mögliche Interpretation von Werten	möglicher Mittelwert	zulässige Transformationen	Beispiele
Nominal	Bestimmung der Gleichheit (gleich, ungleich)	Modus	bijektive	Geschlecht, Farben
Ordinal	Bestimmung der Rangfolge (kleiner, größer)	Median	positiv monotone	Schulnoten
Intervall	Vergleich von Intervallen und Differenzen	arithmetischer Mittelwert	positiv lineare ($\varphi' = a\varphi + b$ mit $a > 0$)	Temperatur in $°C$
Verhältnis	Vergleich von Verhältnissen	geometrischer Mittelwert	positiv proportionale ($\varphi' = a\varphi$ mit $a > 0$)	Alter, monetäre Größen
Absolut	*wie Verhältnisskala*		identitätsbewahrende ($\varphi' = \varphi$)	Häufigkeiten, Wahrscheinlichkeiten

Folgende Beobachtungen können hinsichtlich der aufgeführten Skalen gemacht werden. Mit steigendem Skalenniveau nimmt die Anzahl der Relationen, die im empirischen Relativ existieren, zu. Damit nehmen auch die möglichen Interpretationen, die anhand von Skalenwerten gemacht werden können zu. Dies äußert sich auch in der zunehmenden Möglichkeit der Bildung von Mittelwerten. Während bei einem nominalskalierten Merkmal nur die Bildung des Modalwertes Sinn macht, können etwa bei einem intervallskalierten Merkmal auch

Median oder arithmetisches Mittel gebildet und vor allem interpretiert werden. Der Grad der zulässigen Transformationen der bei einer Skala gegebenen Abbildung φ nimmt hingegen bei steigendem Skalenniveau ab.

Laut Diekmann (2007) wird die Einteilung in Skalentypen, obwohl sie weit verbreitet ist und vielfach Verwendung findet, bis heute kontrovers diskutiert. Mögliche Kritikpunkte werden z.B. von Velleman und Wilkinson (1994) formuliert bzw. resümiert. Einer der wichtigen Kritikpunkte lautet, dass die Vorgabe der Skalenniveaus für reale Datenanalyseprobleme zu stringent ist, d.h. dass nicht alle Merkmale zwangsläufig in die aufgeführten Skalentypen eingeordnet werden können, und dass eine Beschränkung der Analyseverfahren (so auch die einfache Bildung von Mittelwerten) auf bestimmte Skalenniveaus zu starr ist. Ein einfaches Beispiel sind Schulnoten, die (wie oben vermerkt) ordinal skaliert sind. Dennoch wird bzgl. Schulnoten beispielsweise häufig die Bildung des arithmetischen Mittels vorgenommen, etwa bei der Berechnung von Durchschnittsnoten. Die angesprochene Diskussion soll hier nicht weiter vertieft werden. Dennoch sollte anhand dieses Abschnittes die Komplexität der Darstellung von Objekteigenschaften durch Daten verdeutlicht worden sein. Gerade bei Datensätzen, die Individuen und deren Merkmale beschreiben, sollte eine genaue Untersuchung der Daten hinsichtlich ihres Aussagegehaltes, d.h. hinsichtlich der mittels der Daten erhaltenen Relationen zwischen den Individuen, erfolgen. Dies ist vor allem deshalb zu propagieren, da gerade bei sozio-ökonomischen Daten für Individuen die erfassten Merkmale meist höchst unterschiedliche Sachverhalte darstellen, wie beispielsweise etwa das Geschlecht, das Nettomonatseinkommen, ein subjektiver Zufriedenheitsindikator, das Geburtsjahr, usw. Diese Unterschiedlichkeit der durch die Merkmale dargestellten Sachverhalte sollte während des gesamten Analyseprozesses, also auch bei der Datenanalyse bedacht werden.

2.1.2 Auswahl zu beobachtender Objekte

Bevor Eigenschaften von empirischen Objekten durch Daten beschrieben werden können, müssen diese Objekte zunächst zur Messung ausgewählt und herangezogen werden. In manchen Fällen ist man gezielt an einem bestimmten Merkmal eines einzelnen empirischen Objekts interessiert. Als Beispiel ist hier das Bruttoinlandsprodukt der Bundesrepublik Deutschland oder die Lufttemperatur in Dortmund zu nennen. In solchen Fällen geschieht die Auswahl der zu beobachtenden Objekte gezielt durch bewusste Auswahl.

Im Gegensatz dazu existiert jedoch häufig eine (typischerweise große) Menge von Merkmalsträgern, über die man Wissen erlangen möchte. Diese Gesamtheit aller möglichen zu beobachtenden Merkmalsträger wird als *Population* bezeichnet. Im Bereich der Statistik wird zudem häufig der Begriff *Grundgesamtheit* verwendet. In vielen Situationen ist es nicht praktikabel, sogar unmöglich oder aber gar nicht erwünscht, alle Merkmalsträger in einer Grundgesamtheit zu beobachten und Daten für diese zu erheben bzw. zu messen. In diesem Fall liegen den später anzuwendenden Data-Mining-Verfahren nicht Beschreibungen aller Merkmalsträger sondern nur die einer Teilmenge vor. Diese Instanzen werden dann auch *Beispiele* genannt. Die Beispiele bilden eine *Stichprobe*, sofern die Instanzen gemäß einer (meist unbekannten) Wahrscheinlichkeitsverteilung zufällig aus der Population gezogen wurden. Eine zufällige Auswahl, also eine Stichprobe, muss jedoch nicht in allen Fällen gegeben sein und hängt stark von der Art der zu bearbeitenden Forschungsaufgabe ab. Gerade wegen dieser sehr unterschiedlichen Herangehensweisen an die Auswahl von Objekten zur Datenerhebung kann hier keine allgemeingültige Darstellung über die unterschiedlichen

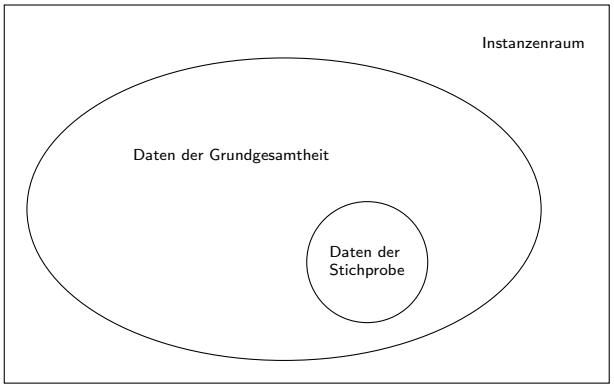


Abbildung 2.1: Instanzenraum, Daten der Grundgesamtheit und Stichprobe

Mechanismen gegeben werden. Stattdessen wird an späterer Stelle (siehe Kapitel 2.2.3) auf die Auswahl von Merkmalsträgern zur Datenerhebung speziell in sozio-ökonomischen Studien, die Daten einzelner Individuen erheben, und das bei dem im Zentrum dieser Arbeit stehenden Datensatz verwendete Auswahlverfahren eingegangen.

2.1.3 Allgemeine Form von Daten

Auf Basis der bisher erfolgten Betrachtungen können Daten zusammengefasst folgendermaßen formal dargestellt werden. Seien m Merkmale von Merkmalsträgern gemessen und A_j die Menge der für das j-te Merkmal möglichen Skalenwerte. Dann ist der *Instanzenraum* X definiert als $X := A_1 \times \cdots \times A_m$. Der Instanzenraum ist also das Kartesische Produkt der Mengen möglicher Skalenwerte der einzelnen Attribute. Die einzelnen Elemente $\mathbf{x} \in X$ entsprechen Merkmalsvektoren aus Skalenwerten, die für einen Merkmalsträger gemessen wurden. Würden bspw. die Merkmale Geschlecht, Alter und Haarfarbe von Personen erhoben, so ist $(\text{männlich}, 31, \text{blond})$ ein möglicher Merkmalsvektor aus dem Instanzenraum. Die für die Datenanalyse verwendeten Daten entsprechen dann einer Menge solcher Merkmalsvektoren, d.h. einer Teilmenge des Instanzenraumes:

$$(\mathbf{x}_i)_{i \in I} \subset X.$$

Im Regelfall ist diese Teilmenge im Vergleich zum gesamten Instanzenraum typischerweise recht klein. Dazu mache man sich klar, dass - vor allem bei Vorliegen einer großen Anzahl gemessener Merkmale sowie einer großen Anzahl möglicher Skalenwerte für die einzelnen Merkmale - selbst die Daten für die Grundgesamtheit, würden sie erhoben, nicht zwangsläufig den gesamten Instanzenraum abdecken müssen, da nicht alle möglichen Merkmalskombinationen auch zwangsläufig in der Realität vorkommen. Außerdem besteht die Möglichkeit, dass die Menge möglicher Skalenwerte - und damit der Instanzenraum - nicht endlich ist. Des Weiteren wird (wie oben erwähnt) häufig nur eine Teilmenge bzw. Stichprobe der Grundgesamtheit durch Daten erfasst. Zur Verdeutlichung ist diese (im Allgemeinen geltende) Teilmengenbeziehung zwischen Instanzenraum, den Daten der Grundgesamtheit und den Daten einer Stichprobe in Abbildung 2.1 dargestellt. Konventionell wird als Indikatorenmenge I für die gegebenen Beispiele die Menge $\{1, \ldots, n\}$ verwendet, wobei n die Anzahl der beobachteten Merkmalsträger ist. Bei zufälliger Auswahl der Merkmalsträger

aus einer Population entspricht dies der Größe der Stichprobe. Die Größe der Grundgesamtheit wird meist mit N bezeichnet.

2.2 Aufbau sozio-ökonomischer Datenerhebungen

Eine Datenerhebung kann im Wesentlichen durch zwei Aspekte charakterisiert werden. Dies sind zum einen das *Erhebungsdesign*, zum anderen die *Erhebungsmethode*. Das Design einer Datenerhebung wird bestimmt durch den Zeitbezug und die Festlegung der Mengen von Untersuchungseinheiten, für die Daten erhoben werden sollen. Die Erhebungsmethode bestimmt, auf welche Art und Weise die Daten von den einzelnen Untersuchungseinheiten schließlich akquiriert werden, also etwa durch direkte Beobachtung oder Befragung.

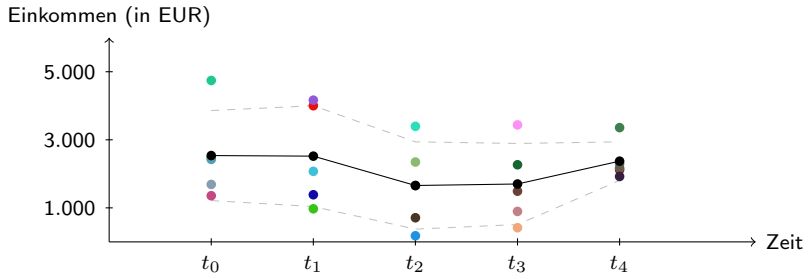
2.2.1 Erhebungsdesign

Grundsätzlich lassen sich Datenerhebungen hinsichtlich ihres Designs anhand ihres Zeitbezugs unterscheiden. Werden Daten lediglich einmalig zu einem Erhebungszeitpunkt (oder in einer vernachlässigbar kleinen Zeitspanne) für eine Menge von Untersuchungseinheiten erhoben, so handelt es sich hierbei um eine *Querschnittstudie*. Im Gegensatz dazu versteht man unter einer *Längsschnittstudie* eine wiederholte Erhebung der Daten einer Menge von Untersuchungseinheiten zu mindestens zwei verschiedenen Zeitpunkten. Die in einer Längsschnittstudie zu den verschiedenen Zeitpunkten betrachteten Mengen von Untersuchungseinheiten müssen dabei nicht zwangsläufig übereinstimmen. Ist jedoch eine statistische Vergleichbarkeit der jeweils verwendeten Mengen von Untersuchungseinheiten gegeben (die sich z.B. aus gleicher Größe der Mengen sowie einem identischen Verfahren zur Auswahl von Untersuchungseinheiten ergibt), so erlauben Längsschnittstudien dynamische Analysen, also die Erfassung bzw. Beobachtung zeitlicher Veränderungen - im Gegensatz zu Querschnittstudien, bei denen nur statische Analysen möglich sind.

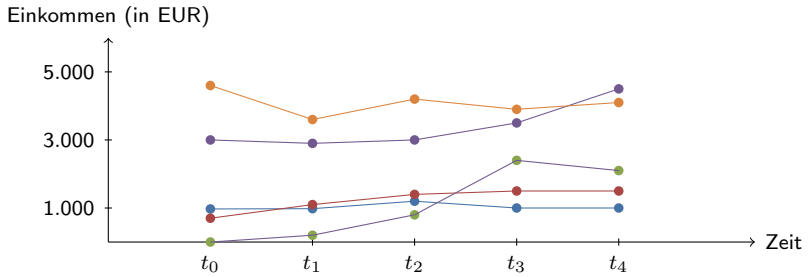
Es existieren verschiedene Formen von Längsschnittstudien, die sich bzgl. der Auswahl der Menge der zu den verschiedenen Zeitpunkten beobachteten Untersuchungseinheiten und der Auswahl der Fragen bzw. Fragestellungen, zu denen Daten erhoben werden, unterscheiden. Die *Panelstudie* ist eine solche spezielle Form der Längsschnittstudie. Besondere Charakteristika der Panelstudie sind, dass (1) die Menge beobachteter Untersuchungseinheiten nur einmalig bestimmt wird und dann zu allen Erhebungszeitpunkten identisch ist, also zu jedem dieser Zeitpunkte die Daten derselben Untersuchungseinheiten erfasst werden, und (2) stets Daten zu denselben Fragen bzw. Fragestellungen erhoben werden. Die statistische Vergleichbarkeit der Mengen von Untersuchungseinheiten ist bei der Panelstudie durch die Verwendung der stets selben Menge trivialerweise implizit gegeben. Bei der *Trendstudie* werden zwar Daten zu denselben Fragen jedoch in jedem Erhebungszeitpunkt von unterschiedlichen Untersuchungseinheiten erhoben. Hansen (1982) listet weitere hiervon differierende Formen wiederholter Datenerhebungen mit Längsschnittcharakter auf.

Zur Verdeutlichung der Unterschiede zwischen den angesprochenen Erhebungsdesigns zeigt Abbildung 2.2 exemplarische Ergebnisse dreier fiktiver Studien zu einer gleichartigen Fragestellung im Querschnitts-, Trend- und Paneldesign. Anhand dieser Abbildung offenbaren sich die analytischen Möglichkeiten hinsichtlich der Auswertung der Ergebnisse von Datenerhebungen mit den vorgestellten Erhebungsdesigns. Durch die einmalige Erhebung von Daten lassen die Ergebnisse einer Querschnittsstudie nur Rückschlüsse auf die zum Zeitpunkt der Erhebung geltende Situation, etwa auf die Einkommensverteilung zum

(a) Querschnittsdesign



(b) Trenddesign. Darstellung des durchschnittlichen Einkommens als schwarze Linie und der Standardabweichung durch gestrichelte Linien.



(c) Paneldesign. Darstellung der individuellen Einkommensverläufe durch Linien.

Abbildung 2.2: Typen von Datenerhebungendesigns (vgl. Diekmann (2007)). Fiktive Studien zur Einkommensverteilung anhand von Stichproben bestehend aus fünf Personen.

Erhebungszeitpunkt, zu. Dynamische Analysen, die Veränderungen der Einkommensverteilung untersuchen, sind nicht durchführbar. Trenddaten lassen eine solche Untersuchung zeitlicher Veränderungen der Einkommensverteilung zu. Allerdings erlauben Trendstudien nicht die Beobachtung von Veränderungen auf Ebene der einzelnen Untersuchungseinheiten. Stattdessen können nur Veränderungen auf aggregierter Ebene, also beispielsweise eine Veränderung der Mittelwerte oder der Streuungsmaße, beobachtet werden. Datenerhebungen mit Paneldesign bieten durch die Beobachtung einer stets identischen Stichprobe zusätzlich auch die Möglichkeit der Analyse von Veränderungen auf Ebene der einzelnen Untersuchungseinheiten.

Der Unterschied zwischen Trend- und Panelstudien, welcher aus der Verwendung unterschiedlicher bzw. einer identischen Stichprobe resultiert, ist hinsichtlich der zu extrahierenden Informationen und damit für die Datenanalyse bedeutsam. Dies zeigt folgendes, ebenfalls fiktives Beispiel der Untersuchung der Entwicklung von Arbeitslosigkeit bzw.

Erwerbstätigkeit. Es seien hierzu zunächst Personen mittels einer Trendstudie mit einer Stichprobengröße von $n = 100$ beobachtet. Der zu beobachtende Arbeitsmarktstatus der einzelnen Personen könne entweder erwerbstätig, arbeitslos oder nicht erwerbstätig sein. Wie bereits ausgeführt, werden in einer solchen Trendstudie keine individuellen Veränderungen, d.h. Arbeitsmarktzustandsübergänge einzelner Personen, erfasst. Eine Auswertung der Trendstudie ist nur bzgl. Aggregaten möglich. Seien als Aggregate etwa die relativen Häufigkeiten der Zustände betrachtet, d.h. die Quoten der einzelnen Arbeitsmarktzustände. Eine Darstellung solcher fiktiver Quoten findet sich in Abbildung 2.3. Auf Aggregatebene

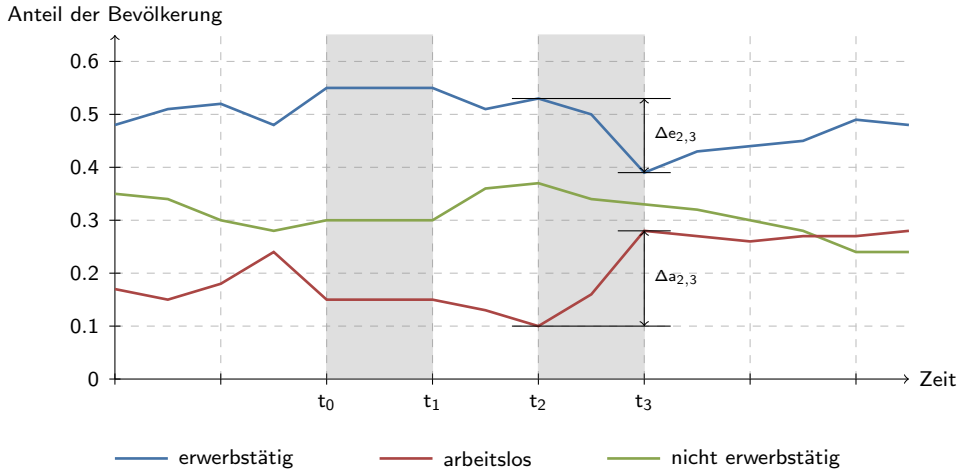


Abbildung 2.3: Fiktive Arbeitsmarktzustandsquoten im Zeitverlauf

lassen sich z.B. im Zeitraum von t_2 nach t_3 eine Abnahme der Quote der Erwerbstätigen um $\Delta e_{2,3}$ und eine Zunahme der Arbeitslosenquote um $\Delta a_{2,3}$ erkennen. Daraus, dass auf Aggregatebene keine Veränderungen zwischen t_0 und t_1 zu erkennen sind, lässt sich jedoch nicht schließen, dass keine Veränderung stattgefunden hat. Selbst wenn in t_0 und t_1 zufälligerweise die gleichen Untersuchungseinheiten befragt wurden, so bedeutet dies nicht, dass keine Zustandsübergänge stattgefunden haben. Möglicherweise existierende Fluktuationen wären jedoch durch die fehlende Zuordnung der individuellen Arbeitsmarktzustände in beiden Zeitpunkten zu den einzelnen Personen nicht erkennbar. Dies soll folgende Betrachtung zweier Fälle verdeutlichen, wobei unterstellt wird, dass zu den Zeitpunkten t_0 und t_1 identische Untersuchungseinheiten beobachtet wurden. Zusammenfassend seien für beide Fälle die Arbeitsmarktzustände zu diesen Zeitpunkten in den Kontingenztabellen aus Abbildung 2.4 dargestellt. Mit Hilfe einer Trendstudie wären nur die Veränderungen auf Aggregatebene erkennbar, also Veränderungen der Häufigkeitsverteilung, die der Randverteilung in den Kontingenztabellen entspricht. Diese Veränderungen heißen auch *Nettoveränderungen*, da sie nur die Differenz aus den Zu- und Abflüssen aus den einzelnen Arbeitsmarktzuständen angeben. In beiden in Abbildung 2.4 dargestellten Beispielen sind die Nettoveränderungen aller drei Zustände gleich null. Allerdings hat nur im Beispiel aus Abbildung 2.4(a) keine Fluktuation zwischen den Arbeitsmarktzuständen stattgefunden. Im Beispiel aus Abbildung 2.4(b) ist dies nicht der Fall. Zu erkennen sind diese sogenannten *Bruttoveränderungen*, die durch die gemeinsame Verteilung der Arbeitsmarktzustände zu den beiden Zeitpunkten gebildet werden, jedoch nur mittels einer Panelstudie, bei der sich die einzelnen Arbeitsmarktzustände einzelnen Personen zuordnen und so Zustandswechsel auf Personenebene

X_{t_0} \ X_{t_1}	arbeitslos	erwerbstätig	nicht erwerbstätig	$h_{t_0\cdot}$
arbeitslos	15	0	0	15
erwerbstätig	0	55	0	55
nicht erwerbstätig	0	0	30	30
$h_{\cdot t_1}$	15	55	30	100

(a) keine Bruttoveränderung der Arbeitsmarktzustände

X_{t_0} \ X_{t_1}	arbeitslos	erwerbstätig	nicht erwerbstätig	$h_{t_0\cdot}$
arbeitslos	0	15	0	15
erwerbstätig	0	40	15	55
nicht erwerbstätig	15	0	15	30
$h_{\cdot t_1}$	15	55	30	100

(b) Bruttoveränderungen der Arbeitsmarktzustände von *arbeitslos* nach *erwerbstätig*, von *erwerbstätig* nach *nicht erwerbstätig* und von *nicht erwerbstätig* nach *arbeitslos*

Abbildung 2.4: Fiktive Kontingenztabellen von Arbeitsmarktzuständen zu zwei Zeitpunkten

beobachten lassen.

2.2.2 Erhebungsmethode

Unabhängig vom jeweiligen Erhebungsdesign können Daten von den Untersuchungseinheiten auf verschiedene Art und Weise erhoben werden. Dies kann zum einen durch schlichte *Beobachtung* eines Sachverhaltes geschehen. Zum zweiten können Daten auch *prozessproduziert* sein, d.h. etwa während eines Prozesses entstehen, der nicht in erster Linie die Datenerhebung zum Ziel hat. Als Beispiel sind hier etwa amtliche Statistiken, z.B. die Zahl der als arbeitslos registrierten Personen, zu nennen. Eine der am häufigsten vor allem in den Sozialwissenschaften verwendeten Datenerhebungsmethoden (siehe Diekmann (2007, S. 434ff.)) ist die Erhebung von Daten durch *Befragung* von Untersuchungseinheiten. Da auch das SOEP befragungsbasiert ist, soll diese Erhebungsmethode hier im Vordergrund stehen und kurz hinsichtlich der Vor- und Nachteile beleuchtet werden.

Ein Vorteil der Befragung als Erhebungsmethode gegenüber anderen Methoden ist, dass mit ihr nicht nur objektive Sachverhalte sondern auch subjektive Indikatoren erhoben werden können. Hierzu zählen bei Personen beispielsweise Meinungen, Einstellungen, Werte oder auch der Grad der Zufriedenheit. Gerade die Erhebung solch subjektiver Merkmale z.B. durch Meinungsumfragen spielen in Gesellschaft, Politik und Wirtschaft eine zunehmend große Rolle zur Erkennung von Stimmungen und Trends. Bezogen auf objektive Merkmale hat die Befragung allerdings den Nachteil, dass auch bei der Beantwortung von Fragen bzgl. solcher Merkmale die Subjektivität der Befragten mit eingeht, und die Antworten daher unter Umständen verfälscht und die Ergebnisse damit verzerrt werden.

Ein weiterer Vorteil der Befragung ist, dass auch Sachverhalte erfragt werden können, die nicht auf den Erhebungszeitpunkt bezogen sind. So können selbst in nur einem einzelnen Erhebungszeitpunkt Daten für einen vergangenen Zeitraum durch retrospektive Fragen erhoben werden und so auch eine quasi-kontinuierliche Erhebung von Daten für einen vergangenen Zeitraum approximiert bzw. simuliert werden. Auch hier wirkt unter Umständen

das subjektive Erinnerungsvermögen jedoch einschränkend auf die Qualität der resultierenden Daten.

Für eine detailliertere Betrachtung verschiedener Erhebungsmethoden sowie der Befragung und mit ihr verbundener Probleme im speziellen, z.B. hinsichtlich der Beeinflussung der Befragten durch die Art der Fragestellung, usw., sei erneut auf Diekmann (2007, Teil C) verwiesen.

2.2.3 Verfahren zur Stichprobenauswahl

Bei Datenerhebungen, die wie Querschnitts-, Trend- oder Panelstudien Daten einer Menge von Untersuchungseinheiten erheben, steht vor jeder Erhebung von Daten die Definition der Grundgesamtheit, die in der jeweiligen Studie betrachtet werden soll. Aus dieser Grundgesamtheit muss dann eine Auswahl von zu beobachtenden Untersuchungseinheiten erfolgen. Es muss also zunächst bestimmt werden, von welchen Merkmalsträgern überhaupt Daten erhoben werden sollen. Werden alle Untersuchungseinheiten der Grundgesamtheit zur Datenerhebung ausgewählt, so handelt es sich bei dieser Erhebung um eine *Voll-* bzw. *Totalerhebung*. Als *Teilerhebung* bezeichnet man eine Erhebung, bei der vor der eigentlichen Erhebung eine explizite Auswahl einer echten - typischerweise im Vergleich zur Grundgesamtheit sehr viel kleineren - Teilmenge der Grundgesamtheit erfolgt. Teilerhebungen bieten sich an, wenn Vollerhebungen aufgrund einer sehr großen oder sogar nicht endlichen Grundgesamtheit nicht möglich oder aber nur mit erheblichem Aufwand zu realisieren sind.

Für die Durchführung der Auswahl von Untersuchungseinheiten aus der Grundgesamtheit existieren diverse Möglichkeiten. Dieser Abschnitt gibt einen Überblick über die wichtigsten dieser alternativen Möglichkeiten zur Auswahl von Untersuchungseinheiten. Der folgende Überblick folgt im wesentlichen der einschlägigen Literatur zu diesem Thema (vgl. z.B. Neubäumer (1982), Kreienbrock (1993) und Althoff (1993)) und stellt die einzelnen Verfahren kurz vor, um ein grundlegendes Verständnis dieser zu schaffen. Die Verfahren der Auswahl von Untersuchungseinheiten aus einer Grundgesamtheit werden in *zufällige* und *nicht zufällige Auswahlverfahren* unterteilt.

2.2.3.1 Zufällige Auswahlverfahren

Bei zufälligen Auswahlverfahren wird eine Stichprobe zufällig aus der Grundgesamtheit gezogen. Dieser Prozess wird auch *engl. Sampling* genannt. Dabei wird vorausgesetzt, dass jedes Element der Grundgesamtheit eine nicht-negative Wahrscheinlichkeit besitzt, Element der Stichprobe zu werden. Gemäß diesen Wahrscheinlichkeiten werden bei der Auswahl Elemente aus der Grundgesamtheit gezogen und zu einer Stichprobe zusammengefasst. Zur Realisierung einer Zufallsstichprobe existieren unterschiedliche Verfahren: Bei der *reinen Zufallsauswahl* werden nacheinander Elemente aus der Grundgesamtheit mit der Wahrscheinlichkeit der einzelnen Elemente (analog zum aus der Statistik bekannten Urnenmodell) gezogen. Bei der *systematischen Zufallsauswahl* wird davon ausgegangen, dass die Elemente der zu untersuchenden Grundgesamtheit in einer Liste vorliegen. Die Stichprobe wird dann folgendermaßen gezogen: Ein Element aus der Liste wird zufällig ausgewählt. Anschließend wird von diesem gewählten Element jedes k-te Element ausgewählt, wobei k gegeben ist durch $\frac{N}{n}$. Hierbei bezeichnet N die Anzahl der Elemente in der Grundgesamtheit und n die gewünschte Stichprobengröße. Ist $\frac{N}{n}$ nicht ganzzahlig, kann nur der ganzzahlige Anteil verwendet werden, und die Auswahl muss bei Erreichen der gewünschten Stichprobengröße

abgebrochen werden. Ein wichtiger Unterschied zwischen der reinen und der systematischen Zufallsauswahl ist, dass bei der systematischen Zufallsauswahl - obschon jedes Element die gleiche Ziehungswahrscheinlichkeit besitzt - im Gegensatz zur reinen Zufallsauswahl nicht alle Teilmengen der Grundgesamtheit die gleiche Ziehungswahrscheinlichkeit besitzen. Ist $k > 1$, so hat eine Teilmenge mit in der Auswahlliste aufeinanderfolgenden Elementen der Grundgesamtheit die Ziehungswahrscheinlichkeit 0 (vgl. Kalton (1983, S. 17)).

Vor allem in von Aufwand und Umfang bedeutenden Studien wie dem SOEP werden Zufallsauswahlen meist aus stichproben- sowie analysetechnischen Gründen mit unterschiedlichen Verfahrensprinzipien kombiniert. So werden Stichproben im Rahmen einer *mehrstufigen Zufallsauswahl* gezogen. Dies bedeutet, dass zunächst eine Stichprobe auf einer der eigentlichen Untersuchungseinheit übergeordneten Einheit gezogen wird. Hierunter fallen typischerweise zum Beispiel Gemeinden oder Wahlkreise. Aus diesen werden dann in einer zweiten Stufe die eigentlichen Untersuchungseinheiten (etwa Haushalte, Personen, Wahlberechtigte, etc.) abermals zufällig ausgewählt.

Eine zweite Möglichkeit der Abwandlung der Ziehung einer Stichprobe ist die *Schichtung* der Zufallsauswahl, das sogenannte stratifizierte Sampling. Bei der Schichtung von Stichproben wird die Kenntnis der Verteilung eines Merkmals in der Grundgesamtheit, anhand dessen die geschichtete Stichprobe gezogen werden soll, vorausgesetzt. Die einzelnen Schichten sind disjunkte Teilmengen der Grundgesamtheit mit unterschiedlichen Ausprägungen dieses Merkmals. Aus den einzelnen Schichten werden dann gesonderte Stichproben gezogen. Die geschichtete Zufallsauswahl kann entweder *proportional* oder *disproportional* erfolgen. Bei der proportionalen Schichtung sind die Ziehungswahrscheinlichkeiten (also die Wahrscheinlichkeiten für Elemente der Grundgesamtheit, Teil der Stichprobe zu sein) identisch. Daraus resultiert, dass bei der proportionalen Schichtung die Stichprobengrößen proportional zur Größe der Schicht in der Grundgesamtheit sind. Bei der disproportionalen Schichtung können einzelne Schichten in der Gesamtstichprobe hingegen über- bzw. unterrepräsentiert sein. Üblicherweise werden spätere Analysen dann ebenfalls nach Schichten gesondert durchgeführt. Soll eine gemeinsame Analyse der Schichten erfolgen, müssen unter Umständen korrigierende Maßnahmen (z.B. durch Gewichtung der Daten) getroffen werden. Dieser Aspekt der wird in Bezug auf das SOEP in Kapitel 3.1.3 erneut aufgegriffen und vertieft.

2.2.3.2 Nicht zufällige Auswahlverfahren

Auch nicht zufällige Auswahlverfahren spielen in der Datenerhebungspraxis eine große Rolle. Ein nicht zufälliges Auswahlverfahren ist die *willkürliche Auswahl*. Bei einer willkürlichen Auswahl wählt ein Interviewer nach eigenem Belieben Untersuchungseinheiten zur Beobachtung bzw. Befragung aus, d.h. es existiert keine explizit festgelegte Strategie zur Auswahl. Ein häufig genanntes Beispiel ist z.B. die Befragung von Menschen auf der Straße, bei dem der Interviewer vorbeigehende Passanten zur Befragen auswählt. Im Gegensatz dazu werden Untersuchungseinheiten bei der *bewussten Auswahl (judgement sampling)* nicht vom Interviewer willkürlich, sondern vom Designer der Studie so ausgewählt, dass diese Auswahl in Bezug auf die Studie sinnvoll erscheint. Ein Spezialfall der bewussten Auswahl ist die *Auswahl typischer Fälle (Monographie)*. Bei dieser sollen aus der Grundgesamtheit genau die Untersuchungseinheiten ausgewählt werden, die typisch für die Grundgesamtheit sind. Fraglich ist dabei allerdings, was als typisch zu erachten ist. Ein weiteres nicht zufälliges Verfahren ist die *Auswahl nach dem Konzentrationsprinzip (cut-off-Verfahren)*.

Hierbei werden nur die Untersuchungseinheiten ausgewählt, die ein bestimmtes Kriterium erfüllen. Ein typisches Beispiel hierfür sind Umfragen bei Firmen, die eine bestimmte Größe haben, also z.B. eine bestimmte Mindestanzahl von Mitarbeitern haben. Bei der *Quotenauswahl* wird die Grundgesamtheit bzgl. bestimmter Eigenschaften in der getroffenen Auswahl nachgebildet. Sollte etwa das Geschlecht einer Person als bestimmende Eigenschaft in einer Quotenauswahl berücksichtigt werden, so müssten Frauen und Männer in der ausgewählten Teilmenge der Grundgesamt im gleichen Verhältnis vorhanden sein, wie sie auch in der Grundgesamtheit vorkommen. Sollen also etwa 100 Personen für die Teilnahme an einer Studie ausgewählt werden, und wären in der Grundgesamtheit 52 Prozent Frauen und 48 Prozent Männer, so müssten 52 Frauen und 48 Männer ausgewählt werden.

2.3 Formen sozio-ökonomischer Daten

Je nach Menge der einbezogenen Merkmalsträger, der zeitlichen Terminierung der Datenerhebung sowie dem Zeitbezug der bei der Erhebung gesammelten Daten ergeben sich verschiedene Formen von Daten. Im Bereich sozio-ökonomischer empirischer Forschung werden vor allem vier Typen von Daten verwendet: (1) *Querschnittsdaten*, (2) *Zeitreihendaten*, (3) *Paneldaten* und (4) *Ereignisdaten*. Daten mit der schon in Kapitel 2.1.3 vorgestellten Form entsprechen Querschnittsdaten. Diese lassen sich wie folgt formal definieren:

Definition 2.3 (Querschnittsdaten) *Seien von n Untersuchungseinheiten Daten erhoben. Sei $\mathbf{x}_i$ der für Untersuchungseinheit i erhobene Merkmalsvektor, wobei $i = 1, \ldots, n$. Dann sind durch*

$$(\mathbf{x}_i)_{i=1,\ldots,n}$$

Querschnittsdaten *gegeben.*

Dies Definition verzichtet bewusst auf eine Einschränkung hinsichtlich des Zeitbezugs wie sie bei Definition der Querschnittstudie erfolgt ist. Die Intention, die in der Definition mit der Abstraktion vom Zeitbezug verfolgt wird, wird an folgendem Beispiel deutlich: Angenommen, man wolle die Lebenserwartung von Personen in Deutschland durch Beobachtung bestimmen. Dies wäre durch eine Querschnittstudie möglich, die etwa das Lebensalter der gestorbenen Personen an einem Tag erfasst. Unter Umständen ist deren Anzahl jedoch relativ gering. Da sich eine bessere Schätzung durch eine größere Stichprobe ergibt, könnte man auch einen längeren Zeitraum im Rahmen einer Längsschnittstudie betrachten. Will man allerdings keine Veränderungen der Lebenserwartung im Zeitverlauf untersuchen, so reicht eine Aufzeichnung des erreichten Lebensalters für die Personen. Diese so aufgezeichneten Daten besitzen dann Querschnittsform, obwohl sie über einen längeren Zeitraum erhoben wurden und sich - genau genommen - auch nicht auf einen Zeitpunkt beziehen.

Eine weitere Form sehr oft verwendeter Daten sind *Zeitreihendaten*. Eine formale Definition kann folgendermaßen geschehen:

Definition 2.4 (Zeitreihe) *Seien von einer Untersuchungseinheit für T Zeitpunkte Daten erhoben. Sei $\mathbf{x}_t$ der für den Zeitpunkt t erhobene Merkmalsvektor, wobei $t = 1, \ldots, T$ ist. Dann heißt die Folge*

$$(\mathbf{x}_t)_{t=1,\ldots,T}$$

Zeitreihe. *Enthält der Vector* $\mathbf{x}_t$ *nur eine Komponente, so heißt die Zeitreihe* univariat, *andernfalls* multivariat.

Im Allgemeinen bezeichnen die Indizes t diskrete Zeitpunkte, die zudem häufig äquidistant sind oder aber als äquidistant angenommen werden.

Zeitreihendaten werden in der ökonomischen Forschung vielfach verwendet, um makroökonomische Aggregate wie z.B. das Bruttoinlandsprodukt, einen Preisindex oder Indikatoren wie das Konsumentenvertrauen über die Zeit zu verfolgen. Solche Daten von Aggregaten werden daher auch als Makrodaten bezeichnet. In der sozio-ökonomischen Forschung werden zudem häufig Daten einzelner Wirtschaftssubjekte (Personen, Haushalte, Firmen, usw.) erhoben. Solche Daten werden als Mikrodaten bezeichnet. Da Daten eines einzelnen, isoliert betrachteten Wirtschaftssubjekts im Regelfall für umfassende Analysen nicht ausreichend sind, betrachtet man meistens Querschnitte solcher Mikroeinheiten. Vielfach soll jedoch nicht nur ein statischer Querschnitt von Untersuchungseinheiten betrachtet werden, sondern es soll eine Vielzahl von Untersuchungseinheiten über die Zeit beobachtet werden. Es bedarf daher einer Kombination aus Querschnitts- und Zeitreihendaten. Dieser Kombination entsprechen *Paneldaten*.

Definition 2.5 (Paneldaten) *Seien von* n *Untersuchungseinheiten für* T *Zeitpunkte Daten erhoben. Sei nun* $\mathbf{x}_{it}$ *der für Untersuchungseinheit* i *für den Zeitpunkt* t *erhobene Merkmalsvektor, wobei* $i = 1, \ldots, n$ *und* $t = 1, \ldots, T$*. Dann sind durch*

$$(\mathbf{x}_{it})_{i=1,\ldots,n,t=1,\ldots,T}$$

Paneldaten *gegeben.*

Auch bei Paneldaten ist die Menge der Zeitpunkte, für die die Daten erhoben sind, meist diskret. Unter Umständen - vor allem aus erhebungstechnischen Gründen, z.B. aufgrund einer aufwendigen und zeitintensiven Erhebung - kann die Zeitdifferenz zwischen zweien solcher Zeitpunkte recht groß sein, sodass evtl. nicht alle Änderungen der Merkmalsausprägungen für Untersuchungseinheiten erfasst werden. Dieser Nachteil wird im Allgemeinen durch *Ereignisdaten* behoben, die Merkmalsausprägungen genau dann erfassen, wenn sich die Ausprägungen von Merkmalen ändern. Unter dieser Voraussetzung lassen sich Ereignisdaten, die auch als *Verlaufsdaten* bezeichnet werden, folgendermaßen definieren:

Definition 2.6 (Ereignisdaten) *Sei für* n *Untersuchungseinheiten jede Änderung der zu den Untersuchungseinheiten gehörigen Merkmalsvektoren* $\mathbf{x}_i$ *mit* $i = 1, \ldots, n$ *erhoben. Sei* $\mathcal{T} = \{1, \ldots, T\}$ *eine Menge von Zeitpunkten zu denen sich der Merkmalsvektor einer Untersuchungseinheit ändern kann. Dann sind durch die Menge*

$$\left\{(t, i, \mathbf{x}_i) \mid \mathbf{x}_{it} \neq \mathbf{x}_{it-1}, i \in \{1, \ldots, n\}, t \in \mathcal{T}\right\}$$

Ereignisdaten *gegeben.*

Zwischen den einzelnen Datenformen bestehen enge Verbindungen. Paneldaten können etwa als Folge von Querschnittsdaten gesehen werden, wobei vorausgesetzt werden muss, dass die Querschnittsdaten stets die gleichen Untersuchungseinheiten betrachten. Prinzipiell ist auch die Betrachtung von Paneldaten als Menge von Zeitreihen für eine Menge von Untersuchungseinheiten möglich. Schließlich sei außerdem erwähnt, dass Paneldaten in Ereignisdaten überführt werden können, sofern die Menge der möglichen Beobachtungszeitpunkte übereinstimmt. Auch die umgekehrte Richtung ist in diesem Fall möglich.

2.4 Panelstudien und Repräsentation von Paneldaten

In Bezug auf Datenerhebungen existieren vielfach spezielle Terminologien. An dieser Stelle sollen in Bezug auf Panelstudien und Paneldaten gebräuchliche Begriffe kurz vorgestellt und erläutert werden. Anschließend wird einführend die Repräsentation von Paneldaten beschrieben.

Die einzelnen Zeitpunkte der Datenerhebungen einer Panelstudie werden *Wellen* (*engl.* waves) genannt. Dies beruht darauf, dass die Daten üblicherweise nur einmal in einer bestimmten Periode (z.B. einmal pro Jahr) - eben *in Wellen* - erhoben werden. Dies ist wiederum unter anderem dadurch begründet, dass die Datenerhebung meist mit einem großen Aufwand verbunden ist. So müssen etwa bei Haushaltspanelstudien wie dem SOEP mehrere tausend Personen befragt werden. Aus diesem Grund ist auch ein exakt einheitlicher Erhebungszeitpunkt häufig nicht realisierbar, stattdessen erstreckt sich die Datenerhebung über einen gewissen Zeitraum innerhalb der jeweiligen Periode. Durch die Definition der in der jeweiligen Periode durchgeführten Erhebung als Welle wird allerdings trotzdem meist ein einheitlicher Erhebungszeitpunkt unterstellt, und eventuelle durch geringfügig unterschiedliche Befragungszeitpunkte verursachte Verzerrungen werden vernachlässigt.

Die in Panelstudien erhobenen Fragestellungen bzw. erfassten Werte(-reihen) werden als *Items* bezeichnet. In Haushaltspanelstudien sind z.B. das Geschlecht, die Religion, der Wohnort oder der Monatsbruttolohn übliche von Personen erfragte Items. Hierbei ist zu beachten, dass ein Item grundsätzlich alle erfassten Werte zu einer Frage bzw. Fragestellung über alle Erhebungszeitpunkte umfasst. So gehören z.B. das Bruttoeinkommen im ersten Jahr der Erhebung und das Bruttoeinkommen im zweiten, dritten und allen weiteren Jahren zum selben Item.

In der üblichen Sicht auf Panelstudien bzw. Paneldaten wird ein Merkmal über die Zeit, d.h. in den einzelnen Wellen, allerdings durch unterschiedliche Variablen erfasst. Dies ist durch die Zeitveränderlichkeit des Merkmals begründet. In der Regel besteht ein Item daher aus mehreren Variablen, die variierende Größen in den einzelnen Wellen erfassen. In obigem Beispiel wäre das Bruttoeinkommen im ersten Jahr daher in einer, das Bruttoeinkommen im zweiten Jahr in einer anderen Variable, usw. erfasst. Zur Verdeutlichung stellt Abbildung 2.5 ein Panel beispielhaft schematisch dar.

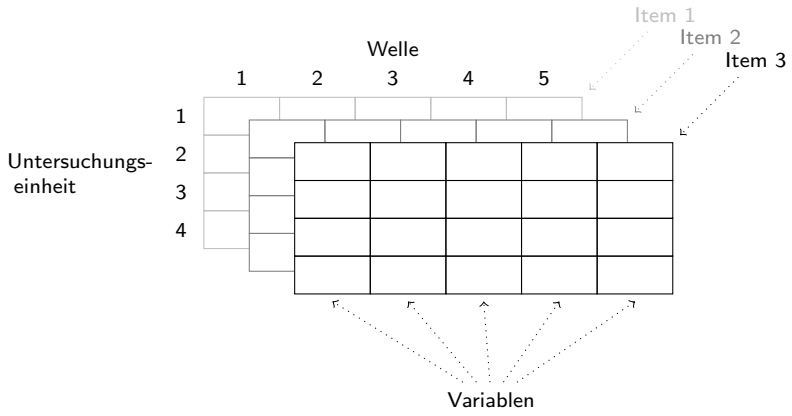
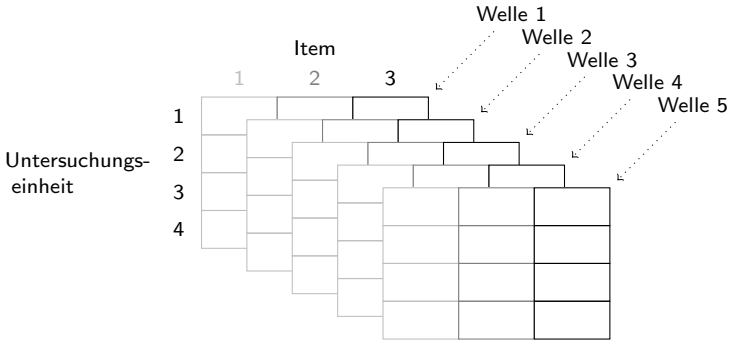
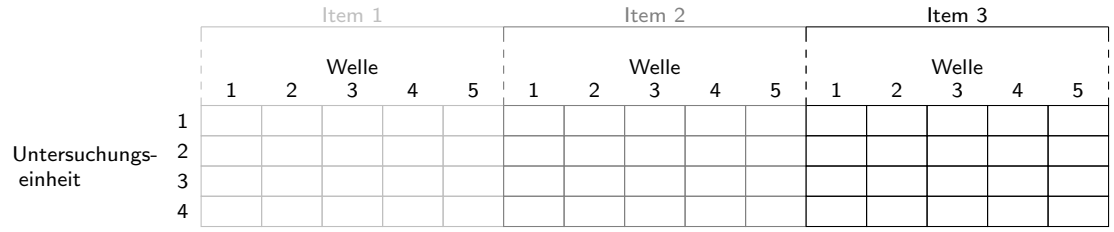


Abbildung 2.5: Schematische Darstellung eines Panels

Abbildung 2.5 zeigt gleichzeitig eine der Möglichkeiten der *Repräsentation* von Paneldaten. Die Repräsentation von Paneldaten meint in diesem Zusammenhang die Anordnung der Dimensionen Untersuchungseinheit, Welle und Item in Tabellen. In Abbildung 2.5 existiert demnach für jedes Item eine Tabelle, die die zugehörigen Werte nach Untersuchungseinheiten (in Zeilen) und Wellen (in Spalten) enthält. Alternativ sind jedoch auch andere - auch in der Praxis häufig auftretende - Formen der Repräsentation denkbar, beispielsweise die in Abbildung 2.6 dargestellten Formen. Die in Abbildung 2.6(a) dargestellte Repräsentations-



(a) Querschnittsrepräsentation



(b) Ein-Tabellen-Repräsentation

Abbildung 2.6: Alternative Repräsentation von Paneldaten

form enthält etwa nicht für jedes Item eine Tabelle, sondern sie beinhaltet eine Tabelle für jede Welle. In jeder dieser Tabellen sind für jede Untersuchungseinheit die Werte erfasst, die für alle Items in der korrespondierenden Welle erhobenen wurden. Dies bedeutet, jede Tabelle bildet einen Querschnitt für die jeweilig zugehörigen Wellen. Diese Repräsentationsform soll daher *Querschnittsrepräsentation* genannt werden. Eine weitere alternative Form ist die in Abbildung 2.6(b) gezeigte, hier sogenannte *Ein-Tabellen-Repräsentation*, die alle zum Panel gehörenden Daten umfasst, indem sie ausgehend von der in Abbildung 2.5 die Tabellen für die einzelnen Items zu einer Tabelle vereint.

Allen bisher vorgestellten Paneldatenrepräsentationen ist gemein, dass Daten für unterschiedliche Wellen immer in unterschiedlichen Spalten erfasst werden. Die Daten befinden sich dann im sogenannten *Wide-Format*. Zur Verdeutlichung des Wide-Formates sei nochmals auf Abbildung 2.6(b) verwiesen. Vom Wide-Format abweichend können Daten alternativ jedoch auch so dargestellt werden, dass für die einzelnen Items nur jeweils eine Spalte bereitgestellt wird, zusätzlich jedoch eine neue Variable hinzugefügt wird, die die Welle erfasst, für die die jeweilig in den Spalten für die einzelnen Items erfassten Daten gültig sind. Abbildung 2.7 stellt die exemplarisch angedeuteten Daten aus Abbildung 2.6(b) in dieser Form da. Die dargestellte Form wird als *Long-Format* bezeichnet. Hierbei sind dann

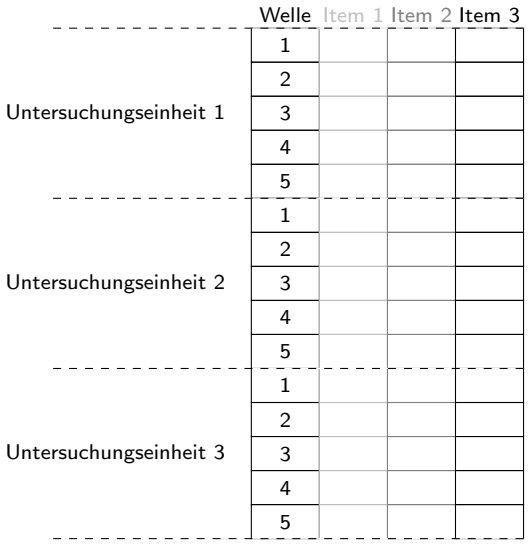


Abbildung 2.7: Long-Format von Paneldaten

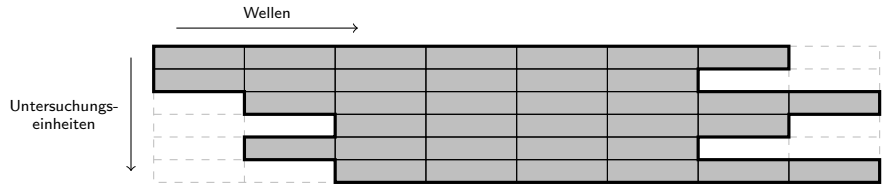
Daten für unterschiedliche Wellen (und unterschiedliche Untersuchungseinheiten) durch unterschiedliche Zeilen der Tabelle gegeben.

Bezüglich aller oben dargestellten Repräsentationsformen ist abschließend erwähnenswert, dass meist Schlüsselvariablen hinzugefügt werden, die eine Zuordnung der Daten zu den Untersuchungseinheiten, zu denen die Daten gehören, erlauben. An Untersuchungseinheiten werden diesbezüglich eindeutig - meist numerische - Identifikationsschlüssel vergeben. Die in den obigen Abbildungen dargestellte implizite Zuordnung bestimmter Tabellenzeilen zu Untersuchungseinheiten erfolgt dann explizit durch eben jene Identifikationsschlüssel.

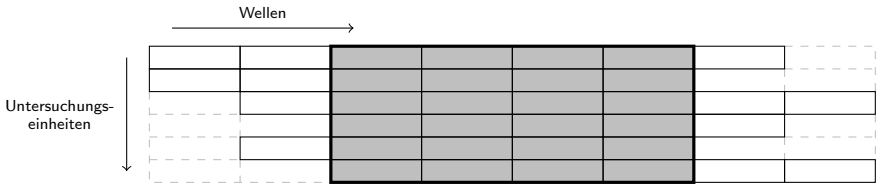
2.5 Probleme bei Panelstudien und Panelanalysen

Dass in jeder Welle die zu erhebenden Daten für alle Items erhoben bzw. beobachtet werden, ist in der Realität häufig nicht erreichbar. Vor allem in Datenerhebungen der Praxis, die - wie z.B. Umfragen oder Befragungen - auf die Interaktion menschlicher Studienteilnehmer angewiesen sind, ist ein idealtypischer Paneldatenaufbau wie in Abbildung 2.5 oder Abbildung 2.6, bei denen für jede Untersuchungseinheit und für jedes Item in jeder Welle das zu beobachtende Datum erhoben werden kann, sehr oft nicht realisierbar. Dies ist im Falle von Umfragen oder Befragungen hauptsächlich darauf zurückzuführen, dass die zur Teilnahme an der Studie ausgewählten Personen Auskünfte teilweise oder auch ganz verweigern. Diesbezüglich benennen Mátyás und Sevestre (1996) mehrere Fälle, die aufgrund ihrer Implikationen für die praktische Durchführung einer Panelstudie und auf ihr basierende Datenanalysen unterschiedlich zu handhaben sind. *Unit Nonresponse* bedeutet, dass eine Untersuchungseinheit die Teilnahme an der Studie grundsätzlich verweigert. In diesem Fall können für diese Untersuchungseinheit gar keine Daten erhoben werden. Vielfach wird Unit Nonresponse innerhalb des Erhebungsdesigns dergestalt berücksichtigt, dass bereits bei der Auswahl von Untersuchungseinheiten zusätzliche Untersuchungseinheiten ausgewählt werden, die als Ersatz für die Teilnahme verweigernde Untersuchungseinheiten dienen. In diesem Fall wird die Datenanalyse nicht von der Problematik der Unit Nonresponse tan-

giert. *Attrition* oder auch *Panelabwanderung* tritt auf, wenn eine Untersuchungseinheit aus dem Panel abwandert. Basiert die Panelstudie auf Befragung als Erhebungsmethode, so ist dies gleichbedeutend dazu, dass anfänglich erfolgreich befragte Studienteilnehmer ab einem späteren Erhebungszeitpunkt die weitere Teilnahme an der Studie verweigern. Dieses Phänomen der Abnahme der Stichprobengröße im Lauf der Zeit ist spezifisch für Panelstudien und wird auch als *Panelmortalität* bezeichnet. Vollständige Daten für alle Wellen existieren dann nur noch für einen Teil der initialen Stichprobe. Verweigert eine Untersuchungseinheit nur in einer Welle die Befragung, nimmt jedoch in der nächsten wieder an der Studie teil, so handelt es sich um *Wave Nonresponse*, welches ebenfalls ein spezielles Problem von Panelstudien ist. *Item Nonresponse* bezeichnet das Fehlen einzelner Werte, welches durch die Nicht-Beantwortung einzelner Fragen durch Studienteilnehmer entsteht. Item Nonresponse kann im Gegensatz zu den beiden zuvor beschriebenen Schwierigkeiten auch in anderen Datenerhebungsformen auftreten. Es findet daher auch in der Literatur große Beachtung. Der Umgang mit der Problematik fehlender Werte, die durch Panelabwanderung, Wave und Item Nonresponse entstehen, ist nicht trivial und hat entscheidenden Einfluss auf die Analyse der Paneldaten. Eine Entscheidung, die diesem Zusammenhang gefällt werden muss, ist etwa die, ob wirklich alle erhobenen Daten in die Analyse mit einbezogen werden sollen, oder nur die Wellen sowie Untersuchungseinheiten, für die vollständige Daten erhoben werden konnten. So könnten beispielsweise nur die Untersuchungseinheiten ausgewählt werden, für die in allen Wellen Daten erhoben werden konnten. Alternativ kann die Anzahl der in der Analyse betrachteten Wellen beschränkt werden, um die Anzahl der Untersuchungseinheiten, für die Daten vollständig erhoben wurden, zu erhöhen. Analysetechnisch gesehen entspricht dies die Entscheidung für ein balanciertes im Gegensatz zu einem unbalancierten Paneldesign bei der Analyse. Wie in Abbildung 2.8 zu erkennen, werden bei dem unbalancierten Paneldesign alle verfügbaren Daten in die Ana-



(a) Unbalanciertes Panel



(b) Balanciertes Panel

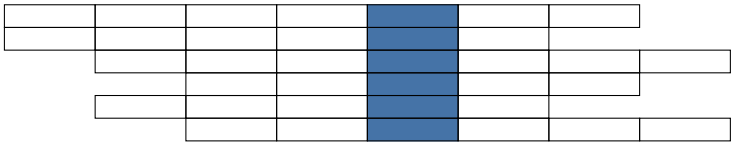
Abbildung 2.8: Panel-Balancierung

lyse einbezogen (grauer Bereich). Dagegen erfolgt bei dem balancierten Paneldesign nur der Einbezug vollständig beobachteter Wellen (oder alternativ - allerdings nicht in der Abbildung dargestellt - vollständig beobachteter Untersuchungseinheiten). Die Reduktion auf ein balanciertes Paneldesign vereinfacht unter Umständen die Analysevorbereitung stark aufgrund der Abstraktion etwa von der Panelabwanderung. Allerdings reduziert dies auch

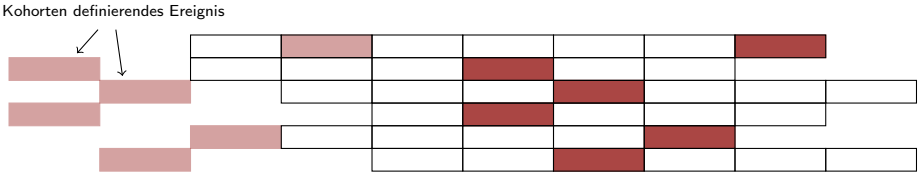
den Umfang der zu verwendenden Daten.

Ein weiteres Problem, welches im Zusammenhang mit Panelstudien, die auf Befragungen als Erhebungsmethode basieren, in der Literatur genannt und daher auch hier kurz erwähnt wird, ist der *Paneleffekt*. Hierunter versteht man die Veränderung des Verhaltens der Studienteilnehmer, welches aus der mehrmaligen Befragung resultiert (siehe Hanefeld (1987)). Allerdings vermutet Hanefeld (1987) eine spürbare Auswirkung eines solchen Effektes lediglich bei Konsumentenpanels. Bei Haushaltspanels, die viele diverse Sachverhalte erfassen, ließe sich der Effekt nicht nachweisen.

Wirken die zuvor genannten Probleme von Panelstudien stark erschwerend auf spätere Analysen, so ist aus analysetechnischen Gesichtspunkten die größte Herausforderung bzw. Schwierigkeit bei der Analyse von Paneldaten die Erfassung der inhärent in den Daten vorhandenen zeitlichen Effekte. Neben der vorhandenen Querschnittsheterogenität, die allein durch die Beobachtung mehrerer Untersuchungseinheiten entsteht, beinhalten Paneldaten zumeist eine Reihe zeitlicher Effekte. Die einfachste Form dieser Effekte ist der *Periodeneffekt*, der alle Untersuchungseinheiten in einer bestimmten Periode (d.h. im Regelfall einer bestimmten Welle) betrifft. Schematisch ist dieser in Abbildung 2.9(a) dargestellt. Ein anderer zeitlicher Effekt ist der *Kohorteneffekt* (vgl. Abbildung 2.9(b)). Dieser bezieht sich



(a) Periodeneffekt



(b) Kohorteneffekt

Abbildung 2.9: Zeitliche Effekte in Paneldaten

jeweils auf ein bestimmtes Ereignis bei Untersuchungseinheiten. Ein Beispiel hierfür ist etwa der Eintritt in die Arbeitslosigkeit oder die Eheschließung. Ein hypothetisches Beispiel für einen Kohorteneffekt wäre beispielsweise eine stark erhöhte Wahrscheinlichkeit von Scheidungen nach beispielsweise sieben Jahren Ehe. Ein Spezialfall von Kohorteneffekten ist der *Alterseffekt*, der sich auf das Ereignis der Geburt der Untersuchungseinheiten bezieht. Ein typischer Alterseffekt ist etwa der Übergang in den Ruhestand, der etwa typscherweise in der ersten Hälfte des siebten Lebensjahrzehnts erfolgt. Die Koexistenz vieler solcher zeitlicher Effekte erschwert vielfach die Datenanalyse bzw. erhöht deren Komplexität. Häufig kann jedoch - zumindest teilweise - von diesen Effekten abstrahiert werden.

3 Das Sozio-oekonomische Panel

Das *Sozio-oekonomische Panel (SOEP)* ist einer der in Deutschland bedeutendsten Haushaltspaneldatensätze, anhand derer sozio-ökonomische Forschung betrieben wird. Es basiert auf einer in Deutschland durchgeführten wissenschaftlichen Studie, bei der in jährlichem Rhythmus Daten von in Deutschland wohnenden Mitgliedern privater Haushalte durch Befragung erhoben werden. In der nicht-wissenschaftlichen Öffentlichkeit ist die Studie unter dem Namen *Leben in Deutschland*[1] bekannt. Das SOEP wurde im Jahr 1983 als Teilbereich des Sonderforschungsbereich 3 "Mikroanalytische Grundlagen der Gesellschaftspolitik" (angesiedelt an den Universitäten Frankfurt am Main und Mannheim) gestartet und wird mittlerweile von einer Abteilung des Deutschen Instituts für Wirtschaftsforschung (DIW Berlin)[2] entwickelt und durchgeführt. Die eigentliche Erhebung der Daten, also die Entwicklung der Fragebögen, die Feldarbeit, d.h. die eigentliche Befragung der Teilnehmer der Studie, und eine Datenprüfung, erfolgt durch das Institut TNS Infratest Sozialforschung, München[3]. Als Serviceeinrichtung der Leibniz-Gemeinschaft[4] stellt die SOEP-Gruppe die erhobenen Daten Wissenschaftlern für Forschung und Lehre zur Verfügung.

Das vorliegende Kapitel gibt einen Überblick über den Aufbau und das Design der SOEP-Studie sowie die Struktur der im Rahmen dieser Studie erhobenen Daten und beleuchtet dabei Aspekte, welche für die spätere Datenanalyse, die im Rahmen dieser Arbeit erfolgt, relevant sind. Aufbauend auf die in Kapitel 2 gegebene Einführung in Datenerhebungen im Allgemeinen und Panelstudien im Speziellen wird in Kapitel 3.1 die Methodik der SOEP-Studie, also deren Konzeption und Durchführung, erläutert. In Kapitel 3.2 folgt anschließend eine Auflistung der Themengebiete, die im Wesentlichen vom SOEP behandelt und durch die Daten abgedeckt werden. Kapitel 3.3 erläutert schließlich den Aufbau des Datensatzes in der Form, wie er Wissenschaftlern zugänglich gemacht wird, und welcher daher den Ausgangspunkt jeglicher empirischer Analysen mit Hilfe des SOEP markiert.

3.1 Konzeption und Durchführung

Das Ziel der Durchführung des SOEP ist nach Hanefeld (1987) die Sammlung repräsentativer Mikrodaten von Personen, Haushalten und Familien in Deutschland, um dynamische Veränderungen von Lebensumständen und -bedingungen in Deutschland zu identifizieren und - daraus abgeleitet - den Grad der Stabilität bzw. der Veränderung der Lebensbedingungen messen zu können. Hierzu werden von in Deutschland lebenden Personen sowohl objektive Indikatoren wie z.B. das Einkommen und der Arbeitsmarktstatus als auch sub-

[1] `http://www.leben-in-deutschland.info`

[2] `http://www.diw.de/deutsch/sop/`

[3] `http://www.tns-infratest-sofo.com`

[4] Die Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL) ist ein Zusammenschluss wissenschaftlich, rechtlich und wirtschaftlich eigenständiger, jedoch durch Bund und Länder kofinanzierter Forschungsinstitute von überregionaler Bedeutung, die meist langfristige Forschungsvorhaben durchführen. Siehe auch `http://www.wgl.de`.

jektive Indikatoren wie beispielsweise der Grad der Zufriedenheit, Werte und Präferenzen, die die individuelle Wahrnehmung der jeweiligen Lebensumstände erfassen, erfragt. Dies geschieht in jährlichem Rhythmus - in Wellen - seit 1984. Die Wellen werden im SOEP durch Buchstaben indiziert, beginnend mit A für die erste Welle. Der im Rahmen dieser Arbeit betrachtete Datensatz umfasst Daten bis Welle U, was der 21. Welle entspricht. Im folgenden werden einige Aspekte der Konzeption und Durchführung der SOEP-Studie im Einzelnen beleuchtet.

3.1.1 Ziehung und Entwicklung der Stichproben

Vor der Befragung von Personen und der eigentlichen Datenerhebung muss die Auswahl der zu befragenden Personen aus einer Grundgesamtheit erfolgen. Sofern diese zufällig erfolgt, handelt es sich um die Ziehung einer Stichprobe. Nach Haisken-DeNew und Frick (2005) besteht die Grundgesamtheit, die durch die SOEP-Studie abgedeckt werden soll und aus der die Stichprobe gezogen wurde, seit 1984 aus der Wohnbevölkerung privater Haushalte in der Bundesrepublik Deutschland (einschließlich West-Berlin) und seit Juni 1990 zusätzlich dazu aus der Wohnbevölkerung der damaligen DDR (einschließlich Ost-Berlin). Die Personen werden aus dieser Grundgesamtheit nicht direkt ausgewählt, vielmehr geschieht eine Auswahl von Haushalten. Das SOEP sieht dann generell vor, dass alle Mitglieder eines zur Befragung ausgewählten Haushaltes beobachtet werden sollen. Selbst befragt werden sollen Personen jedoch erst, nachdem sie das 16. Lebensjahr vollendet haben. Dies bedeutet, dass die für das SOEP ausgewählte Stichprobe von Haushalten direkt auch eine Personenstichprobe impliziert. Diese Vorgehensweise ermöglicht dann sowohl Betrachtungen und Analysen auf der Ebene der Haushalte als auch auf Individualebene.

Die Auswahl der in der SOEP-Studie zu befragenden Haushalte erfolgte prinzipiell zufällig. Eine Besonderheit des SOEP liegt darin, dass nicht eine, sondern initial bereits zwei und im zeitlichen Verlauf der Panelstudie noch weitere Stichproben gezogen wurden. Ein Vorteil dieser Vorgehensweise ist, dass mit den einzelnen Stichproben für einzelne Personengruppen eigene Analysen durchgeführt werden können. Um hierbei einen genügend geringen Stichprobenfehler der einzelnen Stichproben zu garantieren, mussten diese hinreichend groß konzipiert werden. Resultat dessen ist jedoch eine Disproportionalität der einzelnen Stichproben. Dies bedeutet, dass die Stichprobenwahrscheinlichkeiten, also die Wahrscheinlichkeit, dass eine Person der Grundgesamtheit Element der Stichprobe wird, für die einzelnen Stichproben unterschiedlich gewählt wurden. Tabelle 3.1 gibt einen zusammenfassenden Überblick über die einzelnen für das SOEP gezogenen Stichproben, das Jahr der erstmaligen Befragung, ihrer initialen Größen und Stichprobenwahrscheinlichkeiten. Das Design der Stichproben ist im Allgemeinen von vielen Faktoren beeinflusst und daher recht komplex. Daher kann hier keine allumfassende und detaillierte Beschreibung der Stichproben geleistet werden. Es erfolgt jedoch eine kurze, übersichtsartige Darstellung der einzelnen Stichproben und ihrer Eigenschaften. Für eine detailliertere und umfassendere Beschreibung der ersten beiden Stichproben und den Strategien ihrer Ziehung sei bereits an dieser Stelle auf Hanefeld (1987) verwiesen. Für weitere, vertiefende Literaturangaben sei ferner auf Haisken-DeNew und Frick (2005) verwiesen, denen der folgende Überblick im Wesentlichen folgt.

Die initiale Stichprobe der westdeutschen Wohnbevölkerung aus dem Jahr 1984 bestand bereits aus zwei verschiedenen Stichproben: die Stichprobe A enthält Haushalte mit deutschem Haushaltsvorstand[5]. Stichprobe B enthält definitionsgemäß Haushalte mit einem

[5]Der Haushaltsvorstand ist nach Haisken-DeNew und Frick (2005, S. 21) als diejenige Person definiert, die

Tabelle 3.1: Stichproben des SOEP

Stichprobe	Gruppe	Erstmalig Befragung	Initiale Anzahl von Haushalten	Stichproben-wahrschein-lichkeit
A	Deutsche (West)	1984	4.528	0,0002
B	Ausländer (West)	1984	1.393	0,0008
C	Deutsche (Ost)	1990	2.179	0,0004
D	Einwanderer	1994/95	522	0,0002
E	Ergänzung	1998	1.067	0,00003
F	Innovation	2000	6.052	0,00028/0,0005
G	Hocheinkommensbezieher	2002	1.224	

türkischen, griechischen, jugoslawischen, spanischen oder italienischen Haushaltsvorstand. Haushalte mit einem ausländischen Haushaltsvorstand anderer Nationalität wurden aufgrund ihres geringen Anteils an der gesamten Wohnbevölkerung proportional in der Stichprobe der deutschen Haushalte mitberücksichtigt. Stichprobe C, die bereits im Jahr 1990 gestartet wurde, deckt die Wohnbevölkerung der damaligen DDR ab, wobei der Haushaltsvorstand Einwohner der DDR sein musste. Stichprobe D wurde in den Jahren 1994 bzw. 1995 zum SOEP hinzugefügt und beinhaltet Haushalte, in denen mindestens ein Mitglied lebt, das zwischen 1984 und 1993 nach Deutschland immigriert ist. Stichprobe E wurde im Jahr 1998 initiiert und unabhängig von den ersten vier Stichproben definiert. Sie deckt Haushalte unabhängig von der Nationalität des Haushaltsvorstandes gemäß des Vorkommens in der Grundgesamtheit ab. Die Abdeckung der Stichprobe F ist ähnlich zu Stichprobe E. Haushalte mit einem Haushaltsmitglied, welches keine deutsche Nationalität hat, erhielten jedoch eine höhere Auswahlwahrscheinlichkeit. Stichprobe G konzentriert sich schließlich auf Hocheinkommensbezieher derart, dass nur Haushalte, deren monatliches Nettohaushaltseinkommen nicht kleiner als 3.835 Euro war, berücksichtigt wurden[6].

Allen Vorgehensweisen der Ziehung der einzelnen Stichproben gemein ist die Mehrstufigkeit und Schichtung des Stichprobenverfahrens. Zunächst werden geographische Bereiche in Deutschland zufällig ausgewählt. Innerhalb dieser Bereiche werden danach ebenso zufällig Haushalte zur Befragung ausgewählt.

Die Strategie zur Ziehung der Stichprobe von Deutschen (Stichprobe A) basiert auf dem Stichprobenverfahren der Arbeitsgemeinschaft Deutscher Marktforschungsinstitute (ADM). Basierend auf Wahlbezirksstatistiken des Statistischen Bundesamtes werden bei diesem Verfahren Stimmbezirke, sogenannte Sample-Points, zufällig ausgewählt. Anschließend mussten aus den für das SOEP ermittelten 584 Sample-Points Haushalte ausgewählt werden. Die Auswahl wurde nach dem Random-Route-Verfahren ebenfalls zufällig durchgeführt. Dazu erhielten die Interviewer, die schließlich die Befragungen durchführen sollten, eine zufällig vorgegebene Adresse innerhalb des Stimmbezirks. Von dieser ausgehend mussten sie nach bestimmten Regeln den Bezirk durchlaufen und fortlaufend Haushaltsadressen notieren. Aus diesen Adressen wurde schließlich jede siebte ausgewählt und anschließend die zugehörigen Haushalte kontaktiert.

Im Vergleich zur Stichprobe der Haushalte mit deutschem Haushaltsvorstand sind beim

den besten Überblick über die generellen Lebensumstände des Haushalts hat.

[6]Seit 2003 beträgt die Einkommensschwelle 4.500 Euro. Haushalte mit niedrigerem monatlichem Nettohaushaltseinkommen werden nicht weiter befragt.

Sampling der Stichprobe der Ausländer vor allem zwei abweichende Aspekte zu nennen. Zum einen wurden in der ersten Stufe als Sample-Points nicht Stimmbezirke sondern Kreise bzw. kreisfreie Städte zufällig auf Basis einer Auswertung des beim Bundesverwaltungsamt in Köln geführten Ausländerzentralregisters[7] ausgewählt. Nach einer vorgegebenen Methode wurden dann ebenfalls anhand der Ausländerkarteien dieser Kreise die ausländischen Haushalte zufällig ausgewählt. Ein weiterer, nennenswerter Aspekt ist, dass innerhalb der Stichprobe der Ausländer die Haushalte mit Haushaltsvorständen der einzelnen Nationalitäten nicht proportional zu ihrem Vorkommen in der Grundgesamtheit gesamplet wurden, sondern auch in der Stichprobe die einzelnen Nationalitäten disproportional vertreten sind.

Die Ziehung der Stichprobe C erfolgte auf Basis des Zentralregisters der DDR. Für Details zur Konstruktion der Stichprobe und Literaturverweise zur Konstruktion von Stichprobe D siehe Haisken-DeNew und Frick (2005).

Die Prozesse zur Ziehung von Stichprobe E und Stichprobe F ähneln im wesentlichen dem von Stichprobe A. Stichprobe G basiert schließlich auf dem Infratest-Telefon-Master-Sample (ITMS). Das ITMS ist eine mehrstufig und stratifiziert gezogene Stichprobe von Haushalten. Dabei werden die Haushalte gemäß einer telefonischen Anwahl mit zufällig erzeugten Telefonnummern kontaktiert.

Problematisch bei der Ziehung aller Stichproben ist vor allem die Verweigerung der Teilnahme an der Studie durch die kontaktierten Personen. Diesem Problem wird zumeist dadurch begegnet, dass ersatzweise weitere Adressen (oder Telefonnummern) gesammelt und bereitgestellt werden, die bei Bedarf, also beim Nicht-Erreichen einer gewünschten Stichprobengröße durch eine zu hohe Quote an Verweigerern zur Kontaktaufnahme genutzt werden können.

Die Stichproben werden, nachdem sie einmal gezogen worden sind, Teil des SOEP, und ihre Mitglieder werden jährlich befragt. Die Stichproben, d.h. die Menge potentiell zu befragender Studienteilnehmern ist jedoch im Allgemeinen nicht statisch. Verschiedene Faktoren führen vielmehr zu einem Absinken der Größe der Stichproben. Dieser Sachverhalt, das Ausscheiden von Studienteilnehmern aus dem Panel, wird mit dem Begriff *Panelmortalität* bezeichnet. Die Ursachen für Panelmortalität sind vielfältig. Zum einen scheiden Personen durch Tod aus dem Panel aus. Zum anderen sind Personen, die ihren Wohnsitz ins Ausland verlegen, definitionsgemäß nicht mehr Bestandteil der vom SOEP betrachteten Grundgesamtheit. Weiterhin besteht die Möglichkeit, dass Personen innerhalb Deutschlands umziehen, diese Umzüge jedoch von den Interviewern nicht nachvollzogen und die Personen daher nicht erneut kontaktiert werden können. Als letzten aber vielleicht bedeutsamsten Punkt ist schließlich zu nennen, dass sich Personen im Laufe der Zeit entscheiden, nicht weiter an der Studie teilzunehmen und infolgedessen die Befragung verweigern. Diese Ursache der Abnahme der Stichprobengröße wird als *Panelabwanderung* (*engl.* panel attrition) bezeichnet. Zur Illustration dieser Phänomene sind in Abbildung 3.1 die Größen der einzelnen Stichproben nach Personen im Zeitverlauf abgebildet. Darin sind neben den tatsächlich befragten Personen auch die in den Stichprobenhaushalten lebenden Kinder enthalten. Unschwer zu erkennen ist die vorherrschende kontinuierliche Abnahme der einzelnen Stichprobengrößen durch die oben beschriebenen Effekte. Eine detaillierte Analyse der Größenabnahme der Stichproben und deren Gründe findet sich in Kroh und Spieß (2006).

[7] In diesem sollten alle in der Bundesrepublik Deutschland lebenden Ausländer namentlich erfasst sein.

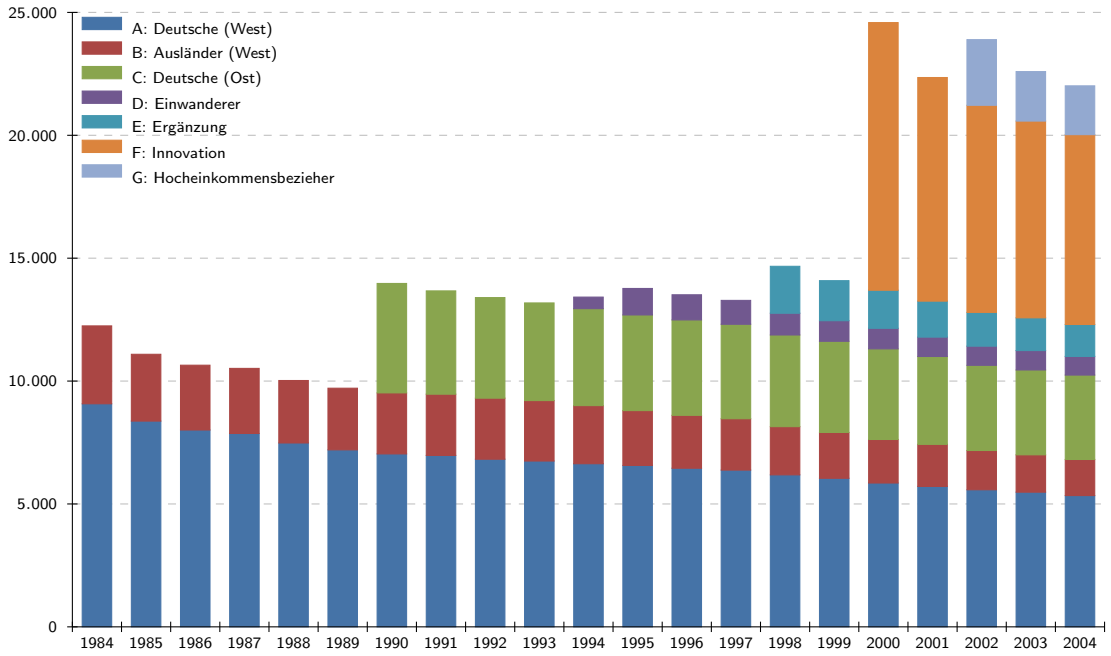


Abbildung 3.1: Stichprobengrößen nach Personen im Zeitverlauf

3.1.2 Panelpflege und Follow-Up-Konzept

Im Gegensatz zum Tod von Studienteilnehmern bzw. zur Emigration ins Ausland ist die Stichprobenabnahme durch Verlust von Kontaktdaten aufgrund des Umzugs der Haushalte ein weitestgehend vermeidbares Problem. Um Umzüge von ganzen Haushalten nachvollziehen zu können und keine Kontaktadressen zu verlieren, werden Adresslokalisierungen etwa anhand von Nachsendeanträgen oder bei Einwohnermeldeämtern durchgeführt. Auszüge von einzelnen Haushaltsmitgliedern werden häufig erst bei Durchführung der tatsächlichen Befragungen bekannt. Dies bedeutet jedoch lediglich eine spätere Befragung der ausgezogenen Personen, wie später in diesem Abschnitt deutlich wird. Im SOEP werden die Recherchen, die im Zusammenhang mit der Mobilität der Haushalte oder einzelner Mitglieder betrieben werden, unter dem Begriff Adresspflege zusammengefasst.

Um der Panelabwanderung durch Verweigerung einer weiteren Teilnahme an der Studie entgegenzuwirken, existieren zudem verschiedene Maßnahmen, die sich mit dem Begriff *Panelpflege* konnotieren lassen (vgl. (Hanefeld, 1987, Seite 263f.)). So erhält jeder Befragte als Anreiz zur Teilnahme ein kleines Geschenk. Darüber hinaus zählen die Information der Teilnehmer bzgl. des Ablaufs der Studie oder des Datenschutzes zu den Schwerpunkten der Panelpflege. Eine vollständige Auflistung der durchgeführten Maßnahmen sowohl der Adress- als auch der Panelpflege findet sich in den Methodenberichten der SOEP-Gruppe der TNS Infratest Sozialforschung (siehe bspw. von Rosenbladt (2004)).

Aus Sicht des Datenanalysten sind vor allem jedoch die folgenden Maßnahmen zur Vermeidung einer Abnahme der Stichprobengrößen relevant. Zum ersten versucht man, das vollständige Ausscheiden von Personen aus dem SOEP, sofern sie die Befragung verweigert haben, dadurch zu verhindern, dass Personen nicht direkt beim erstmaligen Ablehnen der Befragung aus der Studie ausscheiden. Wird etwa die Befragung durch eine Person erst-

malig abgelehnt, so bleibt diese Person dennoch datenmäßig registriert und erfasst. In der auf die erstmalige Ablehnung folgenden Welle wird dann versucht, die Person erneut zu befragen. Erst wenn sie die Befragung abermals ablehnt, wird eine endgültige Verweigerung unterstellt, und die Person scheidet aus der Panelstudie aus. Ansonsten entsteht lediglich eine Lücke innerhalb der Daten, die im Vergleich zum Ausscheiden der Person aus dem Panel jedoch vorzuziehen ist.

Die zweite und vielleicht wichtigere Maßnahme, um ein Absinken der Stichprobengröße zu verhindern, ist das sogenannte *Follow-Up-Konzept*. Das Follow-Up-Konzept beschreibt, welche Personen und Haushalte über die Zeit verfolgt werden. Grundlegendes Prinzip dabei ist, dass jeder, der einmal Teil eines vom SOEP befragten Haushaltes war, verfolgt wird. Dies bedeutet zum einen, dass alle Personen, die initial in einer Stichprobe waren, auch dann nachverfolgt werden, wenn sie aus diesem Haushalt ausziehen und einen neuen gründen oder in einen bereits bestehenden Haushalt einziehen. Zum anderen hat dieses Prinzip zur Folge, dass auch Personen, die nicht initial zu einer Stichprobe gehören, später Teil einer Stichprobe im SOEP sein können. Dies geschieht beispielsweise, wenn eine vorher nicht vom SOEP befragte Person in einen SOEP-Haushalt zieht und damit Teil einer SOEP-Stichprobe wird. Ebenso werden Mitglieder eines Haushaltes, in den eine vormalig vom SOEP befragte Person zieht, Teil einer SOEP-Stichprobe, und der jeweilige Haushalt wird ebenso als neuer Haushalt in das SOEP aufgenommen. Somit können sowohl neue Personen zu den Stichproben hinzukommen als auch neue SOEP-Haushalte entstehen. Voraussetzung für eine Teilnahme neu zu befragender Personen am SOEP ist dabei natürlich immer das Einverständnis der Befragten. Haisken-DeNew und Frick (2005) schildern diesbezüglich allerdings eine relativ niedrige Bereitschaft zur Teilnahme von Personen, die nicht auch initial in einer Stichprobe waren.

Wie bereits oben erwähnt, werden nur Personen eines Haushaltes befragt, die bereits 16 Jahre alt sind. Im Zusammenhang mit dem Follow-Up-Konzept folgt daraus, dass auch in einem SOEP-Haushalt lebende Kinder, sobald sie das 16. Lebensjahr vollendet haben, mit in die Menge der vom SOEP befragten Personen aufgenommen werden.

Grundsätzlich kann die Zahl der vom SOEP betrachteten Personen sowie Haushalte durch das vom SOEP angewandte Nachverfolgungskonzept auch steigen, es also zu sogenanntem *Panelzuwachs* kommen. Die hierfür in Frage kommenden, möglichen Fälle des Auftretens neuer Haushalte sind in Anlehnung an Haisken-DeNew und Frick (2005) zusammenfassend in Tabelle 3.2 dargestellt. Nichtsdestotrotz ist die Panelmortalität im Zeitverlauf im

Tabelle 3.2: Fälle von Panelzuwachs gemäß Follow-Up-Konzept

Personen	Haushalte	
	alt	neu
alt	keine Veränderung, Umzug des gesamten Haushaltes	Auszug aus SOEP-Haushalt
neu	Geburt, Einzug in SOEP-Haushalt	Auszug aus SOEP-Haushalt und Einzug in Nicht-SOEP-Haushalt

allgemeinen jedoch größer als der Panelzuwachs und netto ist meist eine Abnahme der Stichprobengrößen zu verzeichnen. Diese Nettoabnahme bzw. -zunahme wird relativ durch die Kennzahl *Panelstabilität* als dem Quotienten der Stichprobengröße aus einer Welle und

der Stichprobengröße aus der vorherigen Welle (vgl. von Rosenbladt (2004)) erfasst. Für das Jahr 2004 ergibt sich auf Personenebene beispielsweise bei einer Gesamtstichprobengröße von 22.019 Befragungspersonen in 2004 gegenüber 22.611 Befragungspersonen in 2003 eine Panelstabilität von 97,38%.

3.1.3 Gewichtung

Wie in Kapitel 3.1.1 erwähnt, verfolgt das SOEP einen disproportionalen Stichprobenansatz, d.h. die Stichprobenwahrscheinlichkeiten der einzelnen Stichproben, aus der sich die Gesamtstichprobe zusammensetzt, sind unterschiedlich. Darüber hinaus sind selbst die Ziehungswahrscheinlichkeiten in manchen Stichproben (etwa Stichprobe B) unterschiedlich. Zudem ändert sich die Zusammensetzung der Stichproben im Lauf der Zeit etwa durch Panelabwanderung oder durch Panelzuwachs (vgl. Kapitel 3.1.2). Diese Gründe machen eine Gewichtung der Untersuchungseinheiten, d.h. der Personen und Haushalte, bei Analysen auf Basis des SOEP notwendig. Das SOEP stellt daher Gewichte sowohl für Quer- als auch für Längsschnittanalysen für beide Analyseeinheiten bereit. Die Berechnung solcher Gewichte ist relativ komplex und soll hier nicht erläutert werden. Für eine grundlegende Erläuterung der Erlangung von Querschnittsgewichten sei beispielsweise auf Haisken-DeNew und Frick (2005), hinsichtlich der Thematik der Längsschnittgewichtung auf Galler (1987) verwiesen. Festzuhalten ist an dieser Stelle lediglich, dass die Gewichte bei der Durchführung von Analysen zu berücksichtigen sind. Dies erfolgt im Regelfall durch Einbeziehung der Gewichte in den zu analysierenden Datensatz und spezielle Kennzeichnung der Variable, welche die Gewichte enthält, innerhalb der Analysesoftware.

3.1.4 Erhebungsinstrumente und Interviewmethodik

Die eigentliche Erhebung von Daten der ausgewählten Personen und Haushalte ist Aufgabe der Feldarbeit. Dazu werden die Haushalte von den eingesetzten Interviewern innerhalb der Feldzeit kontaktiert und - wenn möglich - befragt. Bei dieser Erhebung von Daten werden verschiedene *Erhebungsinstrumente* genutzt. Als erstes ist diesbezüglich das Adressenprotokoll zu nennen, welches vom Interviewer ausgefüllt wird. In dem Protokoll werden grundlegende Daten wie Vorname, Geschlecht, Geburtsjahr und Stellung zum Haushaltsvorstand der einzelnen Personen im Haushalt erfasst werden. Des Weiteren werden in dem Adressprotokoll auch die durchgeführten Kontakte zum Haushalt protokolliert und Ausfälle (durch Verweigerung der Befragung usw.) dokumentiert. Zweck dieses Adressprotokolls ist die erhebungstechnische Steuerung des Panels. Auf Basis des Adressprotokolls werden so etwa numerische Schlüssel vergeben, die später der Identifikation der Personen in den Daten dienen.

Zusätzlich zum Adressprotokoll existieren als Erhebungsinstrumente außerdem einige Fragebögen. Die wichtigsten dieser Fragebögen sind zum einen der Personenfragebogen und zum anderen der Haushaltsfragebogen. Der Personenfragebogen soll von jeder Befragungsperson in einem ausgewählten Haushalt, d.h. jeder Person in dem Haushalt, die bereits 16 Jahre alt ist, beantwortet werden. Seit 1996 ist dieser Personenfragebogen für alle Personen einheitlich. Davor gab es für die Personen in Stichprobe B und D eigene Fragebögen, die auch immigrationsspezifische Fragen enthielten. Des Weiteren existierte in den Jahren 1990 und 1991 ein spezieller Fragebogen für die Personen aus Stichprobe C, um an diese Fragen bzgl. ihrer Situation während der Phase der deutsch-deutschen Wiedervereinigung stellen

zu können.

In Ergänzung des Personenfragebogens existiert ein spezieller Fragebogen zur Nacherhebung von Daten der jeweils vorherigen Welle im Fall des temporären Ausfalls von Personen in dieser Welle.

Der Haushaltsfragebogen soll vom jeweiligen Haushaltsvorstand beantwortet werden und enthält Fragen zur Gesamtsituation des Haushaltes wie etwa zur Wohnsituation. Zusätzlich enthält der Haushaltsfragebogen wenige Fragen zu sich im Haushalt befindlichen Kindern unter 16 Jahren und deren Lebenssituation.

Sowohl Personen- als auch Haushaltsfragebogen werden jährlich angepasst und verändern sich daher von Welle zu Welle. Demgegenüber existieren diverse Fragebögen zur Erhebung der Biographie von Personen. Diese sollen über die Zeit identisch bleiben und werden den Personen zu unterschiedlichen Zeitpunkten bzw. Ereignissen übergeben. Der Lebenslauffragebogen wird etwa Personen vorgelegt, die neu zum Panel hinzustoßen. Er enthält Fragen zur bisherigen Biographie der neu zu befragenden Personen. Den Jugendfragebogen sollen Jugendliche beantworten, die seit der letzten Befragungswelle 16 Jahre alt geworden sind und daher erstmalig selbst befragt werden. Er enthält Fragen zur Situation der Jugendlichen und zur subjektiven Beurteilung dieser Situation. Der seit 2003 existierende Mutter-Kind-Fragebogen schließlich ist an Mütter mit Kindern gerichtet, die nach Anfang des letzten Jahres (für die Befragung im Jahr 2004 beispielsweise nach dem 1.1.2003) geboren wurden.

Zur eigentlichen Erhebung der Daten werden die Personen direkt befragt, d.h. es werden keine Proxy-Interviews, also Erhebungen von Daten einer Person durch Befragung einer dritten Person als sogenanntem Proxy, durchgeführt. Bei der Befragung wird ein Mixed-Mode-Ansatz verfolgt (vgl. von Rosenbladt (2004)). Damit ist gemeint, dass zur Erhebung der Daten, also zur Befragung der Personen, verschiedene Interviewmethoden zum Einsatz kommen. Die zur Befragung präferierte Methode ist das persönliche, mündliche Interview (*engl.* face-to-face interview), bei dem der Interviewer den zu befragenden Personen die Fragen des Fragebogens stellt und den Fragebogen ausfüllt. Diese Befragung wird entweder unter Verwendung eines Papierfragebogens (*engl.* paper-and-pencil interview) oder mittlerweile auch mit Computerunterstützung (*engl.* computer-assisted personal interview) durchgeführt. Alternativ besteht die Option, die zu befragenden Personen den Fragebogen selbst ausfüllen zu lassen (*engl.* self administered interview). Dies kann entweder im Beisein des Interviewers erfolgen oder aber nach Zustellung der Fragebögen per Post.

Die Diversifizierung von Methoden wird durch die Hoffnung begründet, durch die individuelle Auswahl der jeweiligen Interviewmethode möglichst viele Personen zur Teilnahme an der Befragung zu bewegen und eine Verweigerung dieser zu verhindern. Eine ausführliche Beschreibung der der einzelnen Methoden und ihrer Vor- und Nachteile findet sich z.B. ansatzweise in Hanefeld (1987) oder in von Rosenbladt (2004).

3.2 Themen und Zeitdimensionen der Inhalte

Das SOEP behandelt ein breites Spektrum an Themen und deckt viele Bereiche des täglichen Lebens der Studienteilnehmer durch Fragen und somit erhobene Daten ab. Das SOEP bietet daher eine Fülle von Analysemöglichkeiten zu den unterschiedlichsten Fragestellungen. Bzgl. der Menge der im SOEP gestellten Fragen erfolgt eine Zweiteilung hinsichtlich des Befragungsrhythmus. Mit einer Menge im Wesentlichen gleich bleibender Kernfragen werden jährlich Daten zu den wichtigsten Themenkomplexen erhoben. Hierzu zählen Fragen

zu Persönlichkeitsmerkmalen, Familienbiographien, der Zusammensetzung der Haushalte, Wohnsituation, Mobilität, Erwerbstätigkeit, Einkommen und sozialer Absicherung. Neben diesen objektiven Daten über Personen und Haushalte werden zusätzlich subjektive Daten der Personen in Fragen zu persönlichen Neigungen, zu Werten und zur Zufriedenheit mit verschiedenen Aspekten des Lebens erhoben.

In Ergänzung dieser kontinuierlichen Daten erfolgt in jedem Jahr modulartig die Konzentration auf eines der oben genannten Themen als Schwerpunkt. In lockerem Rhythmus erfolgt so beispielsweise eine Vertiefung der jeweilig erhobenen Daten zu Themen wie Nachbarschaft, sozialen Netzwerken, sozialer Absicherung sowie Weiterbildung und Qualifikationen. Eine vollständige Liste der jeweiligen Themenmodule in den einzelnen Wellen findet sich in (Haisken-DeNew und Frick, 2005, S. 17).

Die Erhebung von Daten zu den oben angegebenen Themenbereichen geschieht mit dem Ziel, Veränderungen aber auch Stabilität der Lebenssituation der Studienteilnehmer zu beobachten bzw. zu messen. Wie bereits in Kapitel 2 ausgeführt, bieten auf Befragung basierende Panels nur eine zeitlich punktuelle, also diskrete Datenerfassung zu den jeweiligen Erhebungszeitpunkten. Quasi-kontinuierliche Daten liefern dagegen nur prozessbezogene Ereignisdaten mit hinreichend kleiner Granularität. Bei Datenerhebungen, die wie das SOEP auf Befragungen basieren, lässt sich dieser Nachteil von Paneldaten jedoch zum Teil beheben, sodass Ereignisdaten künstlich erzeugt werden können. Dies resultiert aus der Möglichkeit, in den Befragungen nicht nur Fragen bzgl. des gegenwärtigen Zeitpunkts und somit der aktuellen Situation sondern auch Fragen zu vergangenen Zeiträumen bzw. zu vergangenen Ereignissen stellen zu können. Das SOEP nutzt diese Möglichkeit auf verschiedene Weise. Zum einen werden in den in Kapitel 3.1.4 angesprochenen Biographiefragebögen vergangenheitsbezogene Daten zur Biographie der Studienteilnehmer erhoben. Zum anderen werden auch in den jährlich ausgegebenen Personenfragebögen retrospektive Fragen gestellt. So enthalten die Personen- und Haushaltsfragebögen eine Reihe von Fragen zu vergangenen Zeiträumen, sodass Ereignisse und Veränderungen innerhalb dieser Zeiträume erfasst werden. Die im SOEP verwendeten Konzepte hinsichtlich des Zeitbezugs von Fragen sind in Abbildung 3.2 exemplarisch für die Wellen 19 bis 21 dargestellt. Das zunächst einfachste und intuitivste Zeitkonzept realisieren Fragen zum gegenwärtigen Zeitpunkt bzw. Fragen zur gegenwärtigen Situation als Ist-Zustand. Das SOEP enthält gegenwartsbezogene Fragen beispielsweise zum aktuellen Erwerbsstatus, dem aktuellen Erwerbs- und Haushaltseinkommen, der aktuell höchsten erworbenen Ausbildung, dem aktuellen Gesundheitszustand, der Zeitverwendung und zu subjektiven Indikatoren wie der aktuellen Meinung und Zufriedenheit. Durch Wiederholung solcher gegenwartsbezogenen Fragen lassen sich Zustandsänderungen von einem Erhebungszeitpunkt auf den nächsten erkennen. Nicht möglich ist jedoch die Erkennung mehrerer Zustandswechsel in dieser Periode. Dieses Problem wird durch retrospektive Fragen zu Ereignissen oder Veränderungen in vergangenen Zeiträumen zum Teil gelöst. Das SOEP verwendet retrospektive Fragen zu unterschiedlichen Zeiträumen. So erfolgt die Fragenstellung hinsichtlich beruflicher und familiärer Veränderungen, des Abschlusses einer Aus- oder Weiterbildungsmaßnahme oder der Modernisierung der Wohnung bzgl. des Zeitraumes zwischen dem Beginn des letzten Kalenderjahres und dem aktuellen Befragungszeitpunkt. Daneben existieren sogenannte Kalenderdaten. Die Fragen, mittels derer diese Daten erhoben werden, beziehen sich dabei auf das gesamte letzte Kalenderjahr vor dem Jahr der aktuellen Erhebung. Im Gegensatz zu den vorherig genannten retrospektiven Daten sind diese weniger bezogen auf ein Ereignis als vielmehr auf die einzelnen Monate als Periode und erfragen z.B. das Einkommen, Transferzahlungen oder die Erwerbsbeteili-

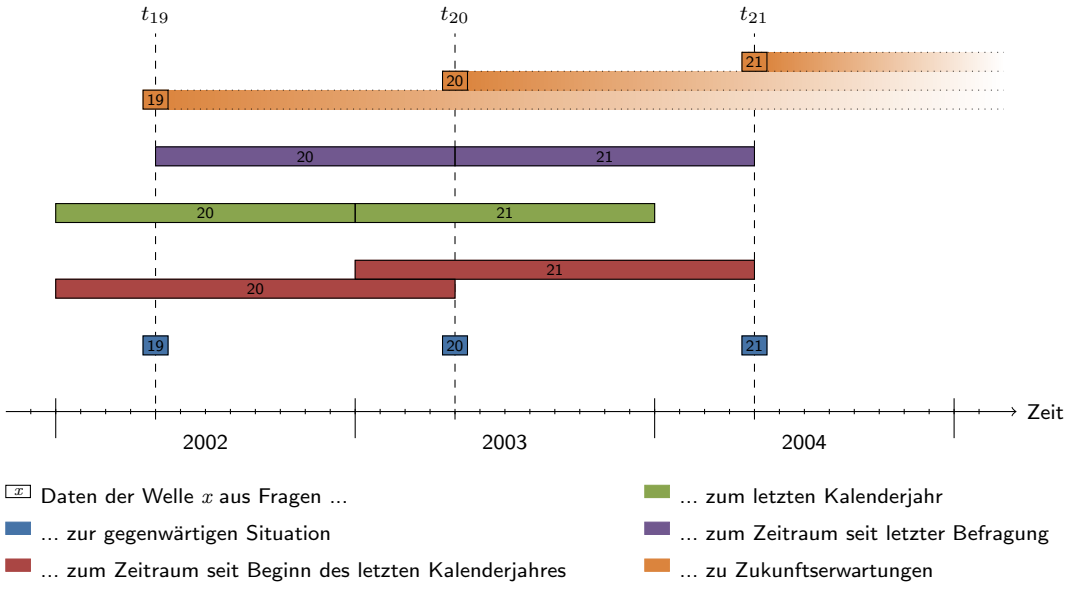


Abbildung 3.2: Zeitbezug von Fragen im SOEP (nach Hanefeld (1987))

gung in den einzelnen Monaten des spezifizierten Kalenderjahres. Schließlich beziehen sich einige Fragen auf die Zeit zwischen dem Erhebungszeitpunkt der letzten Welle und dem aktuellen Erhebungszeitpunkt. Hierzu gehören Fragen zu Wohnungswechseln sowie Fragen zur Veränderung der Haushaltszusammensetzung. Des Weiteren existieren neben den gegenwartsbezogenen und retrospektiven Fragen im SOEP Fragen zu Zukunftserwartungen oder Plänen für die Zukunft. Diese Fragen können einerseits als Sonderfall der gegenwartsbezogenen Fragen angesehen werden, da sie einerseits aktuelle Erwartungen oder Pläne erfassen. Andererseits können sie potentiell zukünftige Entwicklungen prognostizieren oder vorwegnehmen und sind damit gleichermaßen zukunftsbezogen. Zu nennen sind hier z.B. Fragen zum Plan der zukünftigen Aufnahme einer Erwerbstätigkeit.

3.3 Aufbau und Struktur des Datensatzes

Nach Abschluss der Feldarbeit durch Infratest Sozialforschung werden die erhobenen Daten anonymisiert an die SOEP-Gruppe am DIW weitergegeben. Dort wird aus diesen Rohdaten der Datensatz erzeugt, der Wissenschaftlern zur Forschung zugänglich gemacht wird.

Der Datensatz besteht nicht aus einer einzigen Datei, sondern enthält eine Vielzahl einzelner Dateien. So existieren etwa für jede Welle und jede Analyseeinheit (Personen oder Haushalte) separate Dateien. Eine grobe Einteilung der gesamten Dateien kann aufgrund des Zeitbezugs der Daten in den Dateien erfolgen. Einige Dateien enthalten nur Daten aus einer Welle und bilden somit einen Querschnitt (von Personen oder Haushalten) ab. Weiter existieren eine Reihe von Dateien, die wellenübergreifende Daten, also Längsschnittdaten, beinhalten. Bevor erstere in Abschnitt 3.3.2 und letztere in Abschnitt 3.3.3 vorgestellt werden, erfolgt in Abschnitt 3.3.1 eine kurze Erklärung der grundlegenden Form der Dateien im SOEP-Datensatz unter technischen Gesichtspunkten. Abschnitt 3.3.4 erläutert dann die Instrumente und hinzugefügte Metadaten, die eine Verknüpfung der Daten aus den einzelnen

Dateien erlauben und damit verschiedenste weitreichende Analysen ermöglichen. Obschon beide im SOEP betrachteten Analyseebenen (sowohl Personen als auch Haushalte) bei der Darstellung berücksichtigt werden, konzentrieren sich die folgenden Ausführungen teilweise auf Personen als Analyseeinheit. Dies ist durch die später in dieser Arbeit fokussierten Analysen der Arbeitslosigkeit, die gänzlich auf Personenebene durchgeführt werden, begründet. In Abschnitt 3.3.5 wird abschließend auf die Kodierung fehlender Werte im SOEP eingegangen.

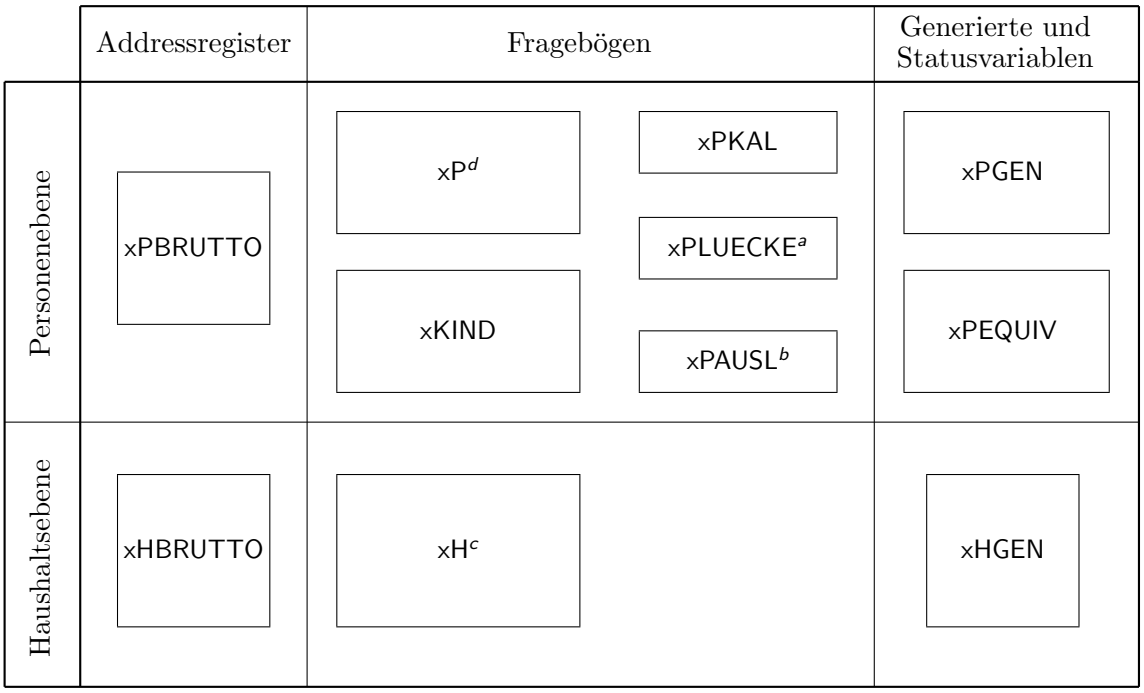
3.3.1 Technische Aspekte

Der SOEP-Datensatz wird in verschiedenen Formaten geliefert, zum einen als ASCII-Daten im CSV-Format, zum anderen auch als Dateien in den Formaten der kommerziellen Statistiksoftwarepakete SPSS und Stata. Per Format-Definition erfolgt in SPSS- und Stata-Dateien eine Speicherung aller Werte (d.h. auch der nominalen) als numerische Werte. Die Datei-Header der Dateien erhalten für nominale Attribute jedoch zusätzlich die Zuordnung von den numerischen Werten zu den als Zeichenketten gespeicherten Bedeutungen, den sogenannten Wertelabeln. Diese enge Verbindung von den eigentlichen, nominalen Werten (repräsentiert als Zeichenkette durch die Wertelabel) und den numerischen Werten, durch die sie kodiert sind, ist im SOEP instrumentalisiert. So ist Ausprägungen aller nominaler SOEP-Variablen stets ein numerischer Repräsentant als Kodierung eindeutig und persistent zugeordnet. So werden auch in den angesprochenen CSV-Daten die nominalen Werte nicht durch ihre Wertelabel, sondern mittels ihrer numerischen Repräsentanten dargestellt. Da das CSV-Format im Gegensatz zu den Dateiformaten der Statistiksoftwarepakete keine Datei-Header vorsieht, erfolgt eine Zuordnung von Wertelabeln und numerischen Kodierungen in zusätzlichen ASCII-Dateien.

3.3.2 Querschnittsdateien

Die Struktur des SOEP-Datensatzes ist im Wesentlichen querschnittsartig. Dies bedeutet, dass für jede Analyseebene (also Personen und Haushalte) für jede Welle eine Datei existiert, die jeweils einen Querschnitt der für die jeweilige Welle befragten Personen respektive Haushalte darstellt. Somit liegen die Daten in der in Abbildung 2.6(a) dargestellten Querschnittsrepräsentation vor. Dadurch entsteht eine recht enge Verbindung zwischen den jährlich ausgegebenen Fragebögen, also den Personenfragebögen und den Haushaltsfragebögen, und den sie repräsentierenden Daten. So werden auf Personenebene alle Antworten aus den Personenfragebögen in den Dateien AP, BP, CP bis UP[8] gespeichert. Auf Haushaltsebene umfassen analog die Dateien AH bis UH die Daten der Haushaltsfragebögen der jeweiligen Wellen. Daneben existieren eine Reihe weiterer Dateien auf Querschnittsbasis als Ergänzung der vorgenannten Dateien xP und xH. Eine Aufstellung der vorhandenen Querschnittsdateien zeigt Abbildung 3.3. Die Dateien xPKAL enthalten sogenannte Kalenderdaten. Diese Kalenderdaten erfassen für jeden Monat des jeweils letzten Kalenderjahres (vgl. hierzu auch Kapitel 3.2) etwa den Erwerbsstatus bzw. die vorrangig ausgeübte Tätigkeit, erhaltene Transferzahlungen sowie das Einkommen. Diese Daten korrespondieren ebenfalls mit Fragen aus den Personenfragebögen, sind jedoch nicht in xP enthalten, sondern in die ge-

[8] Der erste Buchstabe benennt die jeweilige Welle, der zweite steht für die Analyseebene. Sind alle diese Dateien für eine Analyseebene gemeint, so schreibt man häufig xP bzw. xH, wobei x ein Platzhalter für die Buchstabenbezeichnung der Wellen ist.



[a] Alle Wellen ausser der ersten (Welle A) und der letzten
[b] nur Wellen A bis L
[c] GHOST enthält zusätzlich die Haushaltsdaten von Haushalten in der DDR (Stichprobe C) für das Jahr 1990.
[d] GPOST und HPOST enthalten zusätzlich die Personendaten von Bewohnern der DDR (Stichprobe C) für die Jahre 1990 und 1991.

Abbildung 3.3: Querschnittsdateien des SOEP-Datensatzes (vgl. Haisken-DeNew und Frick (2005))

nannten Dateien xPKAL ausgelagert. Analog zu den Dateien xP liegen für die Stichprobe C gesonderte Dateien GPKALOST und HPKALOST vor[9].

Neben den bereits genannten Dateien existieren für einzelne Subgruppen der vom SOEP betrachteten Personen weitere Dateien. Erstens sind in den Dateien xKIND Daten über die in den SOEP-Haushalten lebenden Kinder gespeichert. Da Kinder im SOEP prinzipiell nicht befragt werden, entstammen diese Daten den Haushaltsfragebogen, bei denen der Haushaltsvorstand als Proxy für das Kind Fragen zur Lebenssituation der Kinder beantwortet. Zweitens beinhalten die Dateien APAUSL bis LPAUSL separate Daten, die in den Wellen A bis L von Ausländern zusätzlich erfragt wurden. Nach dieser Welle, also ab 1996, erhielten Ausländer stets den gleichen Fragebogen wie Befragungsteilnehmer deutscher Nationalität. Ab Welle M entfallen daher diese zusätzlichen Dateien für Ausländer. Drittens wurden, wie oben bereits erwähnt, Bewohnern der DDR, die in Stichprobe C neu erfasst wurden, in den Jahren 1990 und 1991 zusätzliche Fragen zu ihrer Situation in Bezug auf die Wiedervereinigung gestellt. Daher befinden sich die Daten zu den Personen aus Stichprobe C nicht in den Personendateien GP und HP, sondern in den Dateien GPOST und HPOST.

[9] Diese enthalten ausnahmsweise jedoch nicht Daten für Januar bis Dezember 1989 bzw. 1990, sondern für Juli 1989 bis Juni 1990 und von Juli 1990 bis März 1991.

Viertens enthalten die Dateien BPLUECKE bis TPLUECKE in der darauffolgenden Welle nacherhobene Daten für Personen, die in einer einzelnen Welle an der Befragung nicht teilgenommen haben. Logischerweise existieren keine Dateien für die erste Erhebungswelle, da nur Personen initial ins Panel aufgenommen wurden, die die Befragung nicht verweigerten, sowie die jeweils letzte im Datensatz erfasste Erhebungswelle, da die Daten erst in der nächsten Welle nacherhoben werden.

Neben den oben beschriebenen Daten, die mit den jeweiligen Fragebögen korrespondieren, werden auf Basis der erhobenen Daten bestimmte Daten extrahiert und neu zusammengefasst sowie zusätzliche Daten generiert und dem Datensatz hinzugefügt. Auf Personenebene finden sich solche Daten primär in den Dateien xPGEN. Diese Dateien beinhalten einerseits Statusvariablen, die die in den einzelnen Wellen eingenommen Stati der Personen hinsichtlich bestimmter Sachverhalten erfassen. Beispiel hierfür sind etwa der gegenwärtig höchste erreichte Bildungsgrad oder die Nationalität. Der Grund für die Extraktion dieser Daten aus den beschriebenen Dateien und deren Zusammenfassung in den Dateien xPGEN liegt darin, dass die Daten häufig in unterschiedlichen Zusammenhängen, also etwa während des jährlichen Personeninterviews, im Rahmen der Nacherhebung einer Lücke oder auch retrospektiv in Biographiefragebögen, erhoben werden. Somit reduziert die Bereitstellung solcher Statusvariablen den Aufwand bei der Extraktion von Daten aus dem Datensatz erheblich. Auf der anderen Seite enthalten die Dateien xPGEN Variablen, die auf Basis der erhobenen Daten generiert werden, also nicht direktes Resultat der Erhebung per Fragebogen sind. Beispielhaft zu nennen sind hier die Stellung im Beruf oder die Autonomie beruflichen Handelns (vgl. SOEP-Gruppe (2004)). Im Gegensatz zu den Inhalten der Dateien xP, die durch sich über die Jahre ändernde Fragebögen nicht zwangsläufig kontinuierlich die gleichen Sachverhalte erfassen und dementsprechend gleiche Variablen enthalten, sind die Dateien xPGEN auf Kontinuität ausgelegt. So werden diese Dateien und ihre Inhalte so zusammengestellt, dass sie über die Wellen hinweg gleiche Variablen mit möglichst gleichen Domänenwerten enthalten, sodass die betrachteten Sachverhalte über alle Wellen hinweg gleichartig erfasst werden.

Auch die Dateien xPEQUIV korrespondieren nicht direkt mit den Fragebögen des SOEP. Vielmehr stellen sie eine Teilmenge der mittels Erhebung gewonnenen Daten dar, wobei die Daten in der Struktur des Cross-National Equivalent File[10] vorliegen.

Die Dateien xPBRUTTO beinhalten Metadaten, die im wesentlichen zur Nachverfolgung der Personen dienen und daher als Adressprotokoll bezeichnet werden.

Die Dateistruktur auf Haushaltsebene ähnelt im Wesentlichen der soeben für Personen als Analyseeinheit beschriebenen. So existieren analog zur Personenebene neben den Dateien xH auch Dateien xHGEN, die sowohl generierte als auch Statusvariablen enthalten. Zudem existieren ebenfalls Adressprotokolldateien xHBRUTTO. Unterschiede ergeben sich dahingehend, dass keine zusätzlichen Dateien für Haushalte mit ausländischen Mitgliedern existieren, da die Haushaltsfragebögen in einer Welle nicht zwischen solchen für ausländische und deutsche Haushalte differierten. Zudem gibt es für die Haushalte aus Stichprobe C nur für das Jahr 1990 eine separate Datei GHOST. Für das Jahr 1991 sind die Daten für diese Haushalte mit in die Datei HH einbezogen.

[10] Das *Cross-National Equivalent File (CNEF)* ist eine Verbund aus Daten von Panelstudien verschiedener Länder. Hierzu gehören neben Daten des SOEP Daten der US-amerikanischen Panel Study of Income Dynamics (PSID), der British Household Panel Study (BHPS), von Household Income and Labour Dynamics in Australia (HILDA) und des Canadian Survey of Labour and Income Dynamics (SLID). Interessierte Leser seien für weitere Information beispielsweise auf Burkhauser et al. (2000) verwiesen.

3.3.2.1 Variablenbenennung

Ein für den Datenanalysten wichtiger Aspekt bei der Analyse von SOEP-Daten bzw. bei der Extraktion von Daten aus dem Datensatz ist die im SOEP benutzte Methode zur Benennung der Variablen. Obschon die Extraktion von Daten auch automatisiert erfolgen kann (siehe auch Kapitel 5), ist die Kenntnis dieser Methode sinnvoll, um eine Verbindung zwischen den Daten und den zugehörigen Fragen in den Fragebögen herstellen zu können. Der Einbezug des Fragebogens ist in vielen Fällen erforderlich, um eine genaue Kenntnis der mittels der Daten erfassten Sachverhalte zu erlangen.

Die im SOEP-Datensatz vorherrschende, enge Verbindung zwischen den Personen- und Haushaltsfragebögen und den mit diesen korrespondierenden Dateien (xP und xH) im Datensatz wird durch die Methode zur Variablenbenennung weiter verstärkt. Das grundlegende Prinzip zur Benennung von Variablen, die Antworten aus diesen Fragebögen erfassen, ist in Abbildung 3.4 dargestellt. Die erste Stelle des Variablennamens ist ein Buchstabe, der

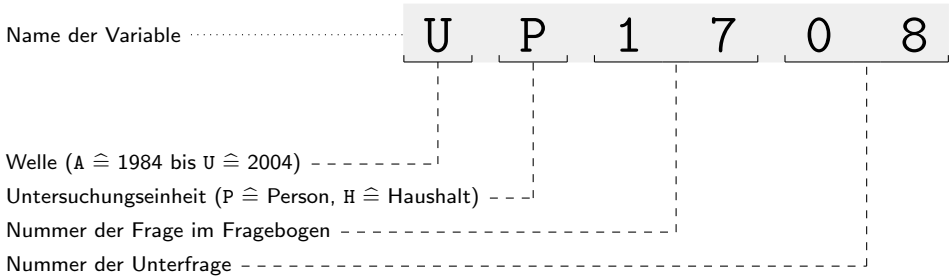


Abbildung 3.4: Benennung von Variablen aus Personen- und Haushaltsfragebögen

die Welle benennt. Der nachfolgende Buchstabe bezeichnet die Analyseeinheit Person oder Haushalt. Die nachfolgenden zwei bis drei Ziffern bilden die Nummer der zugehörigen Frage im Fragebogen. Die darauffolgenden zwei Ziffern nummerieren die Unterpunkte der Frage. Variablen für Fragen aus den Fragebögen für Ausländer aus den Wellen bis 1995 sowie für Ostdeutsche aus den speziellen Fragebögen für Stichprobe C wurden an Stelle fünf bzw. sechs (also zwischen der Nummer der Frage und des Unterpunkts) mit `A` bzw. `O` gekennzeichnet. Die Variablennamen in den Kalenderdateien xPKAL werden unter Einbezug des jeweilig betrachteten Erwerbsstatus bzw. der Einkommensart definiert. Die exakte Definition dieser Variablennamen kann Haisken-DeNew und Frick (2005, S. 87ff.) entnommen werden.

Die Benennung von Variablen bei generierten und in den Dateien xPGEN und xHGEN gespeicherten Daten geschieht entweder durch Kombination des eine Welle bezeichnenden Buchstabens an erster Stelle mit einem Stammkürzel für das jeweilige Item oder durch Kombination dieses Kürzels mit den letzten beiden Ziffern der Jahreszahl der Welle an letzter Position. Beispiele hierfür sind `UBILZEIT` für die Dauer der Ausbildung einer Person in Jahren bzw. `STIB04` für deren Stellung im Beruf, erfragt im Jahr 2004 in Welle U.

3.3.2.2 Item-Correspondence-Tabelle

Um die Veränderung und Dynamik bestimmter Sachverhalte in Längsschnittanalysen erkennen zu können, bedarf es einer Extraktion der Daten, die diesen Sachverhalt abbilden. Dabei müssen die Variablen verknüpft werden, die den gleichen Sachverhalt erfragen, die

also zum gleichen Item gehören. Durch die oben beschriebene Variablenbenennung ist eine Erkennung und Verknüpfung generierter Variablen für dasselbe Item recht einfach, da sich von Welle zu Welle nur die Wellenbezeichnung jedoch nicht das Bezeichnungskürzel ändert. Dies ist bei Variablen für Daten des Personen- bzw. Haushaltsfragebogens nicht notwendigerweise gegeben, da sich sowohl die Position als auch die Konzeption einzelner Fragen ändern kann. Aus diesem Grund wird zusätzlich zum eigentlichen Datensatz eine sogenannte *Item-Correspondance-Tabelle* bereitgestellt. In dieser Tabelle sind für jedes Item die zugehörigen Variablen für jede Welle aufgelistet. Über die Wellen korrespondierende Variablen eines Items sind so leicht identifizierbar.

3.3.3 Längsschnittdateien

Zusätzlich zu den Querschnittsdateien umfasst der SOEP-Datensatz einige Dateien, die wellenübergreifende Daten, d.h. direkt Längsschnittdaten enthalten. Zum einen sind dies Dateien, die Daten enthalten, die mittels der beiden Biographiefragebögen erhoben wurden. Neu zum SOEP hinzu kommende Personen werden zum Zeitpunkt des Eintritts in das SOEP befragt. Diese Daten sind in der Datei BIOSOC erfasst. Generell umfassen diese Daten grundlegende Angaben zur Biographie der Personen, hauptsächlich in Bezug auf die schulische und berufliche Ausbildung. Weitere Biographiedaten werden von Jugendlichen, die bereits vom SOEP erfasst werden, bei der der erstmaligen Berücksichtigung als Befragungsperson, d.h. nach Vollendung des 16. Lebensjahres, erhoben. Diese Daten sind in der Datei BIOYOUTH zu finden. Thematisch sind auch hier Daten zur bislang erfolgten schulischen Ausbildung erfasst, allerdings auch viele subjektive Indikatoren, die sich auf das Umfeld, die Werte und Erwartungen der Jugendlichen beziehen.

Zum zweiten umfasst der SOEP-Datensatz Dateien mit sogenannten Spelldaten. Spelldaten sind Ereignisdaten, d.h. die in den Dateien vorliegenden Untersuchungseinheiten sind Zustände, die Personen in einem Zeitraum eingenommen haben. Eine Beobachtung enthält demnach den Typ, den Anfangs- und den Endzeitpunkt des Zustands. Des Weiteren ist die Beobachtung einer Person zugeordnet. Im SOEP existieren eine Reihe von Dateien dieser Form. Die wichtigsten sind die Dateien ARTKALEN und EINKALEN, die Angaben zur beruflichen Tätigkeit bzw. zu Einkünften für Personen erfassen. Zusätzlich beinhalten die Dateien BIOMARSM und BIOMARSY Angaben zum Familienstand von Personen auf monatlicher bzw. jährlicher Ebene.

Drittens enthalten eine Reihe weiterer Längsschnittdateien Daten zu speziellen Themenbereichen, etwa die Immigration und die Pflege, die im Rahmen dieser Arbeit im Wesentlichen keine Rolle spielen und deswegen nicht näher erläutert werden.

3.3.4 Metadaten

Neben den eigentlich erhobenen sowie den generierten Daten sind auch Daten, die zur Verknüpfung der Daten aus den einzelnen Dateien oder zur adäquaten Durchführung von Analysen benötigt werden, von Bedeutung. Die dem SOEP hinzugefügten Metadaten, die diesen Zweck erfüllen, werden im Folgenden beschrieben.

3.3.4.1 Identifikationsschlüssel

Wie bereits in Kapitel 3.1.4 angedeutet erhalten die einzelnen am SOEP teilnehmenden Personen und Haushalte Identifikationsnummern als Schlüssel. Die Bereitstellung von Iden-

tifikationsschlüsseln geschieht unter anderem mit der für den Datenanalysten relevanten Absicht, die anonyme Identifikation von Untersuchungseinheiten, d.h. Personen und Haushalten, in dem Datensatz sowie die Zuordnung der sich in den einzelnen Dateien befindenden Daten zu den einzelnen Untersuchungseinheiten zu ermöglichen. Daten aus mehreren Dateien, die zu denselben Personen gehören, können so für die spätere Analyse verknüpft werden. Eine Betrachtung der technischen Durchführung dieser Extraktion und Verknüpfung von Daten erfolgt später in Kapitel 5.

Für Personen und Haushalte werden verschiedene Schlüssel verwendet. Personen werden über einen unveränderlichen Schlüssel identifiziert, der in allen personenbezogenen Dateien durch die Variable `PERSNR` dargestellt wird. Auf Haushaltsebene erfolgt die Identifikation über einen Haushaltsschlüssel. Allerdings wird bei diesem keine Unveränderlichkeit verlangt, sodass die Identifikation über die jeweils in einer Welle aktuelle Haushaltsnummer, üblicherweise erfasst in der Variable `HHNRAKT`, erfolgt. Darüber hinaus besitzt jeder existierende Haushalt eine Ursprungshaushaltsnummer, die den Haushalt identifiziert, aus dem andere Haushalte - beispielsweise durch Gründung eines neuen Haushaltes nach Auszug einer Person aus einem bestehenden Haushalt - hervorgegangen sind. Die Nummer des Ursprungshaushaltes wird stets in der Variablen `HHNR` gespeichert.

3.3.4.2 Indexdateien PPFAD und HPFAD

Mit Hilfe der vorgestellten Identifikationsschlüsseln beinhaltet der SOEP-Datensatz (1) die Zuordnung von Personen zu Haushalten, in denen die Personen leben, sowie (2) die Möglichkeit der Nachverfolgung neu entstandener Haushalte. Erstere geschieht in der Datei PPFAD, die für jede am SOEP teilnehmende Person sowohl die Personennummer als auch die für jede Welle die jeweilige Nummer des Haushalts, in der die Person lebte, enthält. Die Nachverfolgungsmöglichkeit neu entstehender Haushalte durch Auszug wird dadurch gewährleistet, dass die Datei HPFAD für jeden jemals existierenden und vom SOEP betrachteten Haushalt die aktuelle Haushaltsnummer der letzten Erhebungswelle (`HHNRAKT`), die jeweilige Nummer des Haushalts für jede Welle (`AHHNR` bis `UHHNR`) sowie die Ursprungshaushaltsnummer (`HHNR`) enthält.

Die Dateien PPFAD und HPFAD bieten jedoch nicht nur soeben genannte Zuordnungsmöglichkeiten, sondern auch Variablen zur Erfassung des Befragungsstatus, und sie dienen daher generell als Indexdateien für Personen bzw. Haushalte im SOEP. Dies bedeutet, dass die Datei PPFAD für jede jemals im SOEP betrachtete Person den Befragungsstatus in jeder Erhebungswelle beinhaltet. Dies gilt analog für Haushalte in der Datei HPFAD.

Die den Befragungsstatus von Personen in den einzelnen Wellen angebenden Variablen heißen `ANETTO` bis `UNETTO`. In Tabelle 3.3 sind die möglichen Werte dieser Variablen dargestellt. Die Variablen `xNETTO` dienen daher dazu, die tatsächlich befragten Personen in der Bruttostichprobe zu erkennen. Für diese sind demnach Daten in den Dateien xP sowie xPKAL und xPKAL vorhanden. Darüber hinaus lassen sich vom SOEP erfasste Kinder, die aufgrund ihres Alters (noch) nicht selbst Befragungsperson sind, und zugehörige Daten in den Dateien xKIND identifizieren. Verweigern Personen erstmalig die Befragung, so erscheinen dennoch Daten für sie in den Adressprotokolldateien xPBRUTTO. Nehmen die Personen in der darauffolgenden Welle wieder an der Befragung teil, so werden Daten nacherhoben, in den Dateien xPLUECKE gespeichert und der Status in der zugehörigen Variable `xNETTO` entsprechend geändert. Verweigern Personen erneut die Teilnahme, so liegt - wie in Kapitel 3.1.2 beschrieben - eine Abwanderung dieser Personen aus dem Panel vor, d.h. es handelt

Tabelle 3.3: Befragungsstatus gemäß der Variablen xNETTO

Kodierung	Bedeutung
0	Personenausfall
1	Befragungsperson
2	Kind
3	nur Adressprotokoll
4	nacherhobene Lücke

sich um sogenannte Personenausfälle. Diese werden kumulativ (d.h. für alle Wellen mit dem jeweiligen Jahr des Ausfalls) in der Datei YPBRUTTO erfasst. Für Personen, die sich nicht in der Bruttostichprobe einer Welle befinden, sind die Werte der korrespondierenden Variable xNETTO gemäß der SOEP-Konvention (siehe zu fehlenden Werten im SOEP auch Kapitel 3.3.5) als fehlend gekennzeichnet.

Analog enthält die Datei HPFAD Variablen xHNETTO, die zwischen Brutto- und Nettostichprobe unterscheiden und somit Haushalte mit erfolgreich durchgeführtem Haushaltsinterview erkennen lassen.

3.3.4.3 Gewichtedateien PHRF, HHRF, PBLEIB und HBLEIB

Wie bereits in Kapitel 3.1.3 erwähnt wurde, stellt das SOEP Gewichte bereit, die in die Analyse einzubeziehen sind. Diese Bereitstellung geschieht in den Dateien PHRF, HHRF, PBLEIB und HBLEIB. Generell werden für entsprechende Analysen unterschiedliche Gewichte für Personen und Haushalte bereitgestellt. In den Dateien PHRF und HHRF werden für Personen bzw. Haushalte Querschnittsgewichte für jede Welle angegeben. So umfasst etwa die Datei PHRF im wesentlichen die Variablen APHRF bis UPHRF, die Querschnittsgewichte für Personen in den Wellen A bis U darstellen. Analog umfassen die Variablen AHHRF bis UHHRF Querschnittsgewichte für Haushalte für die einzelnen Wellen. Darüber hinaus stellen die Dateien PBLEIB sowie HBLEIB Daten in Form von Bleibewahrscheinlichkeiten bereit, anhand derer durch Kombination mit Querschnittsgewichten Gewichte für Längsschnittanalysen explizit berechnet werden können. Da diese Form der Gewichtung im Rahmen dieser Arbeit nicht verwendet wurde, sei hier lediglich auf die Erläuterung der Längsschnittgewichtung in Haisken-DeNew und Frick (2005) verwiesen.

3.3.5 Fehlende Werte

Wie bereits in Kapitel 2.5 beschrieben, existieren mehrere Gründe für das Auftreten fehlender Werte, etwa Panelabwanderung, Wave Nonresponse oder auch Item Nonresponse. Durch die querschnittsartige Speicherung der Erhebungsdaten in Dateien sind bei Panelabwanderung sowie Wave Nonresponse die Daten in den Querschnittsdateien sowie in Längsschnittdateien im Long-Format einfach nicht vorhanden. Bei Dateien im Wide-Format, welches auch die Indexdateien PPFAD und HPFAD aufweisen, müssen nicht zu erhebende Daten jedoch als fehlend gekennzeichnet werden. Auch bei Item Nonresponse fehlen einzelne Werte, die besonders gekennzeichnet werden müssen.

Fehlende Werte sind im SOEP generell gemäß der Zuordnung in Tabelle 3.4 kodiert. Hierbei ist jedoch zu beachten, dass *trifft nicht zu* keinen fehlenden Wert im eigentlichen

Tabelle 3.4: Kodierung fehlender Werte im SOEP

Kodierung	Bedeutung
-1	keine Angabe
-2	trifft nicht zu
-3	nicht valide

Sinn bezeichnet. Vielmehr ist diese Bedeutung als eigener Domänenwert einer Variable zu verstehen, der angibt, dass keine der anderen Antwortmöglichkeiten zutrifft. Pyle (1999, S. 63f.) bezeichnet Domänenwerte solcher Art als leere Werte (*engl.* empty values). Demgegenüber können die anderen möglichen Werte aus Tabelle 3.4 als fehlend (*engl.* missing values) im eigentlichen Sinn betrachtet werden. Bei der späteren Verarbeitung der Daten zur Vorbereitung auf durchzuführende Analysen ist diese Unterscheidung von hoher Bedeutung und sollte daher entsprechende Beachtung erfahren. Die Information, dass entweder keine Angabe gemacht wurde, also keine Antwort gegeben wurde, oder aber eine nicht valide Antwort gegeben wurde, ist allerdings sicherlich in den meisten Analyseaufgaben - sofern diese nicht gezielt das Antwortverhalten von Befragten analysieren - zu vernachlässigen.

4 Definition der Lernaufgabe

Im Mittelpunkt dieser Arbeit steht die Durchführung einer Analyseaufgabe zum Themenbereich *Arbeitslosigkeit* mit Hilfe von Methoden des Data Mining bzw. des maschinellen Lernens. Das SOEP bietet diesbezüglich umfangreiche Daten zur Erwerbssituation der befragten Personen. Aufgrunddessen ist das SOEP Basis vieler in diesem Themenbereich durchgeführter Untersuchungen. Kluve et al. (2005) befassen sich beispielsweise mit der Identifikation von individuellen und makroökonomischen Einflussfaktoren von Arbeitsmarktzustandsübergängen. Diese Fragestellung wurde auch für diese Arbeit zur Analyse ausgewählt.

Dieses Kapitel stellt die Analyseaufgabe dieser Arbeit zum Themenbereich Arbeitslosigkeit vor. Einführend wird dazu in Kapitel 4.1 zunächst die Arbeitslosigkeit als volkswirtschaftliches sowie individuelles Problem vorgestellt und ihre statistische Erfassung kurz beleuchtet. Kapitel 4.2 motiviert die Analyse von Arbeitsmarktdynamiken und erläutert das dieser Arbeit zugrunde liegende Modell von Arbeitsmarktzuständen und Übergängen zwischen diesen. Kapitel 4.3 benennt Größen, die einen potentiellen Einfluss auf die zu analysierenden Arbeitsmarktdynamiken haben könnten. Hierzu werden zum einen potentielle mikroökonomische Faktoren (d.h. Variablen auf Personenebene) identifiziert, die in die Analyse einfließen sollen. Zusätzlich hierzu wird die Hinzunahme makroökonomischer Daten als Hintergrundinformation für die zu analysierenden SOEP-Daten motiviert.

4.1 Arbeitslosigkeit und ihre Erfassung

Hohe Arbeitslosigkeit ist vor allem in den westlichen Wirtschaftsnationen ein vielbeachtetes Problem. Die Arbeitslosenquote ist neben anderen Größen wie beispielsweise das Wachstum des Bruttoinlandsprodukts oder die Inflation eine der wichtigsten Kennzahlen moderner Volkswirtschaften. Dies ist zum einen der volkswirtschaftlichen Relevanz geschuldet, zum anderen spielt die Arbeitslosigkeit auch in der Gesellschaft eine große Rolle. So ist sie auf der einen Seite ein volkswirtschaftliches Problem mit hohen Kosten (z.B. durch Transferzahlungen wie Arbeitslosengeld) für den Staat und damit indirekt auch für die Bürger als Steuerzahler. Auf der anderen Seite spüren von Arbeitslosigkeit betroffene Bürger direkt materielle, psychologische und soziale Folgen. Damit ist Arbeitslosigkeit für viele Personen direkt greifbar. Sie spielt daher - im Gegensatz zu eher abstrakten Größen - auch in der nicht fachspezifisch akademischen Gesellschaft eine große Rolle, welche sich in der hohen Relevanz beispielsweise in der politischen Diskussion zeigt. Die Verringerung der Arbeitslosigkeit ist aus den dargelegten Gründen ein wichtiges Ziel zur Verbesserung der gesamtwirtschaftlichen Situation als auch der individuellen Situation vieler Menschen. Auch aus diesem Grund ist die Arbeitslosigkeit Thema vieler empirischer Untersuchungen wie der in dieser Arbeit durchgeführten, die Einflussfaktoren der Arbeitslosigkeit identifizieren sollen, um daraus Handlungsempfehlungen etwa für politische Maßnahmen abzuleiten.

Obwohl bei vielen eine intuitive, allgemeine Auffassung der Definition von Arbeitslosigkeit besteht, wird diese zur definitorischen Klärung im folgenden konkretisiert, indem

Arbeitslosigkeit definiert und ihre übliche statistische Erfassung kurz erläutert wird. Allgemein bezeichnet man Personen als *arbeitslos*, wenn sie weder abhängig beschäftigt noch selbständig tätig sind, obwohl sie gerne einer solchen Beschäftigung nachgehen würden, also Anbieter ihrer Arbeitskraft sind (vgl. Sauermann (2005)). Diese recht intuitive Definition wird gesetzlich durch die Definition der *registrierten Arbeitslosigkeit* konkretisiert. Nach deutschem Recht[1] sind Personen registriert arbeitslos, die (1) sich bei der Agentur für Arbeit als arbeitslos gemeldet haben, (2) in keinem Beschäftigungsverhältnis stehen oder nur einer geringfügigen Beschäftigung (mit weniger als 15 Stunden pro Woche) nachgehen, (3) eine Beschäftigung (von 15 Stunden oder mehr) suchen und (4) Bemühungen zur Vermittlung einer Stelle zur Verfügung stehen. Bestimmte Personengruppen sind per Definition ausgeschlossen: so etwa Schüler, Studenten, Rentner und Erwerbsunfähige.

Erfasst wird die Arbeitslosigkeit in Deutschland vor allem durch zwei Kennzahlen: zum einen der Zahl der registrierten Arbeitslosen, zum anderen der Arbeitslosenquote. Erstere erfasst alle bei der Agentur für Arbeit (bzw. früher bei Arbeitsämtern) als arbeitslos gemeldete Personen. Zur Berechnung der Arbeitslosenquote existieren mehrere, unterschiedliche Konzepte. Allen gemein ist, dass die Zahl der Arbeitslosen in Bezug gesetzt wird zur Anzahl der Erwerbspersonen, wobei sich diese ergibt als Summe aus der Zahl der Erwerbstätigen und der Zahl der Arbeitslosen, d.h. aller Personen, die Arbeit anbieten. Somit spiegelt die Arbeitslosenquote relativ das nicht genutzte Arbeitsangebotspotential wider. Die von der Bundesagentur für Arbeit berechnete Arbeitslosenquote ergibt sich konkret als Quotient aus der Anzahl der registrierten Arbeitslosen und der Anzahl der zivilen Erwerbstätigen bzw. der zivilen abhängig Erwerbstätigen. Daneben existieren weitere Verfahren zur Berechnung der Arbeitslosenquote. Das Statistische Bundesamt etwa veröffentlicht die Arbeitslosenquote auf Basis des Labour-Force-Konzeptes der International Labour Organization (ILO), das sich unter anderem durch die Definition der Arbeitslosigkeit bzw. Erwerbstätigkeit sowie durch die Art der Erhebung der zugehörigen Größen abhebt. Interessierte Leser finden eine detaillierte Abgrenzung der beiden erwähnten Konzepte der Bundesagentur für Arbeit und des Statistischen Bundesamtes in Sauermann (2005). Zur Illustration der beschriebenen Kennzahlen sind in Abbildung 4.1 die Zahl der registrierten Arbeitslosen sowie die Arbeitslosenquote als Jahresdurchschnitte für die Jahre 1984 bis 2004 abgebildet.

4.2 Analyse von Arbeitsmarktdynamiken

Die im Rahmen dieser Arbeit durchzuführende Analyseaufgabe soll Einflussfaktoren der Arbeitslosigkeit erkennen. Solche Einflussfaktoren können vielfältig sein, da generell viele Gründe für Arbeitslosigkeit verantwortlich sein können. So werden bereits klassisch verschiedene Formen der Arbeitslosigkeit in Abhängigkeit vom jeweilig zutreffenden Grund unterschieden. Baßeler et al. (2002) unterscheidet beispielsweise die *friktionelle*, *strukturelle* und *saisonale* Arbeitslosigkeit. Friktionelle Arbeitslosigkeit ist kurzfristige Arbeitslosigkeit, die durch den Wechsel von Personen zwischen Arbeitsstellen entsteht. Die strukturelle Arbeitslosigkeit resultiert aus wirtschaftsstrukturellen Gegebenheiten. Beispielsweise erhöhte der im Ruhrgebiet in der zweiten Hälfte des 20. Jahrhunderts vollzogene Stellenabbau in Bergbau und Stahlindustrie die strukturelle Arbeitslosigkeit. Saisonale Arbeitslosigkeit meint die saisonale Schwankung der Arbeitslosigkeit aufgrund vermehrter bzw. verringerter Tätigkeiten in bestimmten Jahreszeiten, z.B. in der Baubranche. Der getroffenen Unter-

[1] Sozialgesetzbuch (SGB), Drittes Buch, §§ 16 und 119

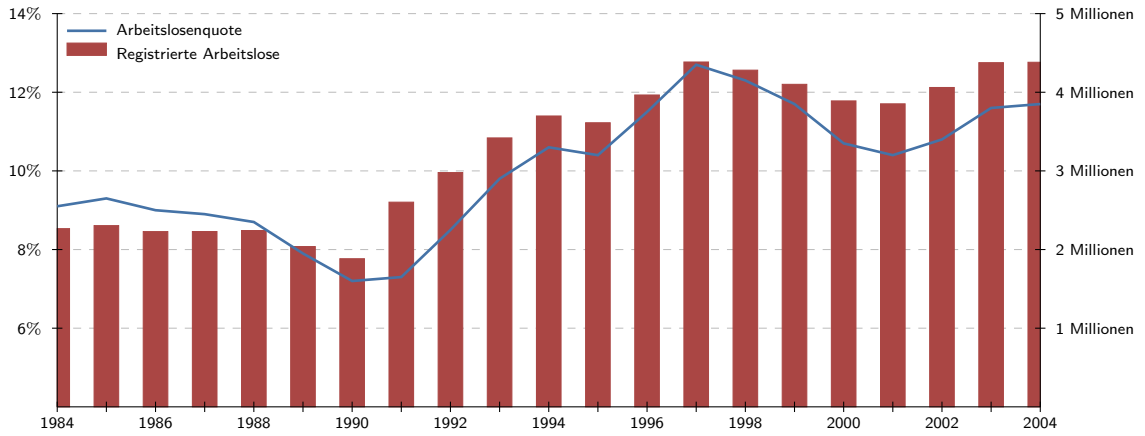


Abbildung 4.1: Registrierte Arbeitslosen und Arbeitslosenquote (1984-2004), Quelle: Bundesagentur für Arbeit

scheidung ist zudem noch die *konjunkturelle* Arbeitslosigkeit hinzuzufügen. Konjunkturelle Arbeitslosigkeit bezeichnet die durch konjunkturelle Faktoren hervorgerufene Komponente der Arbeitslosigkeit. Sie entsteht beispielsweise durch Entlassungen als Folge eines konjunkturell bedingten Konsumrückgangs und daraus resultierender Absatzschwierigkeiten für Unternehmen.

Aus diesen Betrachtungen dieser Typen geht hervor, dass auch zeitliche, dynamische Phänomene eine große Rolle bei der Erklärung der Arbeitslosigkeit spielen und damit auch die Arbeitslosigkeit selbst ein höchst dynamisches Phänomen ist. Die friktionelle Arbeitslosigkeit etwa resultiert aus auftretenden Ereignissen (dem Wechsel von Arbeitsplätzen), die saisonale Arbeitslosigkeit geht aus jahreszeitlichen Mustern wirtschaftlicher Aktivität hervor, die konjunkturelle Arbeitslosigkeit wird forciert durch Schwankungen wirtschaftlicher Aktivität über einen längeren Zeitraum (siehe auch Kapitel 4.3.1).

Aus diesen Überlegungen ist abzuleiten, dass eine rein statische Analyse der Arbeitslosigkeit vermutlich nicht alle Aspekte, die die Arbeitslosigkeit beeinflussen, erkennt. Stattdessen sollte eine dynamische Analyse erfolgen. Der in dieser Arbeit verfolgte und im folgenden beschriebene Ansatz berücksichtigt dynamische Aspekte derart, dass der Einfluss nicht auf die Arbeitslosigkeit als Bestandsgröße, sondern auf Ströme in bzw. aus der Arbeitslosigkeit analysiert werden. Auf der betrachteten Mikroebene, d.h. auf Ebene der einzelnen Personen bedeutet dies, dass nicht tatsächliche Zustände, d.h. arbeitslos gegenüber anderen Arbeitsmarktzuständen (allen voran: erwerbstätig), als Indikator herangezogen werden, sondern Zustandsübergänge in den Arbeitsmarktzustand Arbeitslosigkeit bzw. aus dem Zustand Arbeitslosigkeit heraus.

Prinzipiell werden auch in dieser Arbeit Übergänge zwischen Arbeitsmarktzuständen betrachtet. Allerdings werden nicht nur die Zustandsübergänge zwischen den Zuständen arbeitslos und erwerbstätig (also abhängig beschäftigt) betrachtet. Zustandsübergänge müssen aus den Daten aus Beobachtungen der Zustände zu verschiedenen Zeitpunkten abgeleitet werden. Somit sind zunächst die betrachteten Zustände einzuführen und zu definieren sowie zu erklären, wie sie in den SOEP-Daten bereitgestellt werden. Die in dieser Arbeit verwende-

Tabelle 4.1: Zuordnung von Arbeitsmarktzuständen aus den Variablen `LFSxx`

Wert aus `LFSxx`	Arbeitsmarktzustand	Kürzel
Nichterwerbstätig o.w. Info	*Not-working*	*N*
Nichterwerbstätig und älter als 65 Jahre	*Retired*	*R*
Nichterwerbstätig in Ausbildung	*Training*	*T*
Nichterwerbstätig in Erziehungsurlaub	*Parenthood*	*P*
Nichterwerbstätig in Wehr-Zivildienst	*Military*	*M*
Nichterwerbstätig und arbeitslos gemeldet	*Unemployed*	*U*
Nichterw. und manchmal nebenerwerbstätig	*Jobbing*	*J*
Nichterw. aber letzte 7 Tagen erwerbstätig	*Jobbing*	*J*
Nichterw. und regelmäßig nebenerwerbstätig	*Jobbing*	*J*
Erwerbstätig	*Working*	*W*
Erwerbstätig aber letzte 7 Tage nichterw.	*Working*	*W*

ten Arbeitsmarktzustände basieren auf den in den Variablen `LFSxx`[2] (d.h. `LFS84` bis `LFS04`) im SOEP erfassten Werten. Diese Variablen finden sich in den Dateien xPGEN und erfassen den zum Zeitpunkt der Befragung aktuellen Arbeitsmarktzustand der befragten Personen. Der Fakt, dass die Variablen `LFSxx` nicht direkt Antworten aus Fragebögen widerspiegeln, sondern aus anderen Variablen generiert wurden, spielt hierbei keine Rolle. Um die Komplexität etwas zu reduzieren wurden die möglichen Werte der SOEP-Variablen `LFSxx` gemäß der in Tabelle 4.1 dargestellten Wertezuordnung auf die Arbeitsmarktzustände *Working*, *Unemployed*, *Not-working*, *Jobbing*, *Parenthood*, *Retired* und *Military* abgebildet. Der klassische Zustand der Arbeitslosigkeit wird damit beispielsweise unter dem Zustand *Unemployed* erfasst. Aufbauend auf dieser Definition der Zustände werden innerhalb dieser Arbeit Übergänge zwischen eben diesen Zuständen betrachtet. Eine Betrachtung aller möglichen Zustandsübergänge ist allerdings nicht nötig und daher auch nicht sinnvoll, denn ein Übergang von *Retired* nach *Parenthood* oder nach *Training* ist beispielsweise im Regelfall nicht möglich. Auch andere Zustandsübergänge sind eher ungewöhnlich. Für diese Arbeit wurden daher lediglich zehn Übergänge zwischen Arbeitsmarktzuständen für die Analyse ausgewählt, die zum einen hinsichtlich der Analyse als besonders interessant erscheinen, zum anderen als relativ häufig vorkommend vermutet werden. Diese Übergänge sind aus dem Übergangsmodell ersichtlich, welches in Abbildung 4.2 dargestellt ist. Für den Zustand der Arbeitslosigkeit werden also sowohl die einfließenden Ströme aus den Zuständen *Working* bzw. *Jobbing* betrachtet, als auch die ausfließenden in diese Zustände. Des Weiteren werden für die Zustände *Working* sowie *Jobbing* einfließende Ströme aus den Zuständen *Training* und *Parenthood* betrachtet. Zusätzlich hierzu erfolgt eine Betrachtung der Übergänge aus den Zuständen *Working* und *Jobbing* in den Zustand *Not-working*.

Die obige Definition der Arbeitsmarktzustände und der Übergänge zwischen ihnen impliziert einen wichtigen Punkt: Die zugrunde liegenden Daten aus den Variablen `LFSxx` liegen nur auf Jahresbasis vor, da sie genau die Arbeitsmarktzustände der Personen erfassen, welche die Personen zum Zeitpunkt der Befragung inne hatten. Dies hat zur Konsequenz, dass erstens die Arbeitsmarktzustandsübergänge nur mit der Granularität von Jahren zu erfassen sind. Folglich können saisonale Komponenten der Arbeitslosigkeit nicht erkannt werden.

[2] `xx` ist hier Platzhalter für die letzten beiden Ziffern der Jahreszahl der zugehörigen Welle

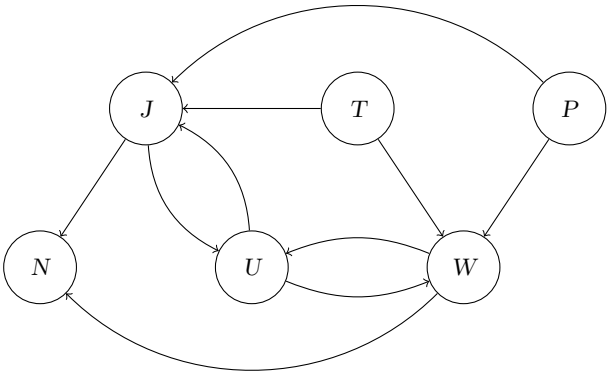


Abbildung 4.2: Betrachtete Übergänge von Arbeitsmarktzuständen

Allerdings muss dies nicht zwangsläufig ein Nachteil sein. Es wird vielmehr die saisonale Komponente von Arbeitslosigkeit häufig explizit eliminiert, um auch in der kurzen Frist, d.h. etwa auf Monatsbasis, ein nicht saisonal beeinflusstes Wachstum der Arbeitslosigkeit bestimmen zu können. Dies geschieht vor allem auch im Kontext mittel- bis langfristig orientierter, makroökonomischer Empirie, bei der Zeitreihen - auch zur Arbeitslosigkeit - in additive oder multiplikative Komponenten zerlegt und Trend, Konjunkturschwankung sowie saisonale Ausschläge als Komponenten isoliert werden. Die zweite Implikation ist, dass nur der Anfangs- sowie der Endzustand innerhalb der Zeitspanne zwischen den Erhebungszeitpunkten berücksichtigt wird. Eventuell auftretende, weitere Zustandswechsel zwischen den Befragungen werden vernachlässigt. Dies betrifft in besonderem Maß kurzzeitige (wie etwa friktionelle) Arbeitslosigkeit. Diese wird nur erfasst, wenn der entsprechende Wechsel in bzw. aus der Arbeitslosigkeit kurz vor der Befragung geschieht.

Zwar propagieren beide genannten Punkte die Nutzung von Daten mit feinerer Granularität, die im SOEP auf monatlicher Basis in Form von Spelldaten in der Datei ARTKALEN tatsächlich vorliegen. Allerdings stehen dem zwei Gegenargumente gegenüber, die für diese Arbeit schließlich ausschlaggebend waren: Zum einen sind die monatlichen Daten retrospektiv erfragt und basieren somit auf Erinnerungen der Befragten. Sie weisen daher vermutlich eine höhere Verzerrung auf als die verwendeten Daten, die genau den Zustand zum Zeitpunkt der Befragung erfassen. Das gewichtigere Argument ist jedoch, dass der Großteil der im SOEP verfügbaren Daten, die als mögliche Einflussfaktoren in Frage kommen, ebenfalls nur auf jährlicher Basis vorhanden ist. Ob eine eklatante Verbesserung der Analyseergebnisse durch eine Analyse auf Monatsbasis erzielt werden könnte, die die sehr viel komplexere Aufbereitung der Daten zur Änderung der Granularität von Jahres- auf Monatsbasis und eine vermutlich einhergende, immense Vergrößerung der Analyselaufzeit aufwiegt, ist fraglich. Aus diesem Grund werden in dieser Arbeit eventuelle, oben beschriebene Mängel der Analyse hinsichtlich saisonaler Faktoren bzw. hinsichtlich der Berücksichtigung von Kurzzeitarbeitslosigkeit hingenommen und alle Analysen generell auf Jahresbasis durchgeführt.

4.3 Potentielle Einflussgrößen

Die im letzten Abschnitt beschriebenen Zustandsübergänge stellen die abhängigen Zielvariablen dar. Aus den obigen Ausführungen ist zu entnehmen, dass diese Übergänge von einer Reihe potentieller Einflussfaktoren abhängen. Ein Ziel dieser Arbeit ist, solche Einflussfaktoren und die entsprechenden Wirkzusammenhänge zu finden und zu benennen.

Die meisten Analyseverfahren sind entweder hinsichtlich ihrer Skalierbarkeit beschränkt, oder aber die Laufzeit wächst bei Hinzunahme vieler Variablen exorbitant. Daher kann allein aus praktischen Gründen nicht jedes im SOEP-Datensatz verfügbare Datum als möglicher Faktor in die Analyse eingehen. Außerdem ist das SOEP hinsichtlich seines Umfangs derart groß und detailliert, dass prinzipiell nicht jede verfügbare Variable als mögliche Einflussgröße in Frage kommen kann. Somit muss eine Vorauswahl erfolgen, die aus den verfügbaren Daten potentiell wichtige Variablen zur Aufnahme in die Analyse auswählt. Im Folgenden erfolgt daher eine thematische Vorauswahl möglicher Variablen, die in die Analyse einbezogen werden sollen.

Wie bereits in Kapitel 3.2 umrissen, beinhaltet das SOEP detaillierte Daten zu vielen Themenbereichen. Als wichtig für die Analyse von Einflüssen auf Arbeitsmarktzustandsübergänge erachtet werden hier zum einen vor allem persönliche Eigenschaften. Die wichtigsten persönlichen Merkmale sind diesbezüglich das Geschlecht und das Alter. Des Weiteren werden Daten zur Herkunft der Personen einbezogen. Dies umfasst erstens das Herkunftsland, womit vor allem Immigrationsaspekte berücksichtigt werden sollen. Zweitens sollen außerdem auch regionale Indikatoren beachtet werden. Als weiterer wichtiger Punkt, der häufig und gerade in der letzten Zeit in der gesellschaftlichen Diskussion erneut eine große Rolle spielt, muss der Aspekt Bildung berücksichtigt werden. Bildung umfasst dabei sowohl die Schulbildung, die Berufsausbildung als auch die Hochschulbildung. Ebenfalls in die Analyse einbezogen werden zudem Daten zur familiären Situation, vor allem zu Kindern. Da das SOEP gerade auch Daten zur Arbeitssituation von Erwerbstätigen kontinuierlich erhebt, sollen solche Daten auch im Rahmen der Analysen berücksichtigt werden. Dies ist allerdings nur bei Analysen hinsichtlich der Zustandsübergänge aus der Erwerbstätigkeit bzw. Nebenerwerbstätigkeit heraus möglich. Eine genaue Beschreibung der aus dem SOEP extrahierten Variablen und der aus diesen zusätzlich generierten, neuen Variablen, die in der Summe die soeben beschriebenen Sachverhalte erfassen, erfolgt in den Kapiteln 5 und 6.

4.3.1 Makroökonomische Daten

Es steht zu vermuten, dass Übergänge zwischen Arbeitsmarktzuständen bzw. die Wahrscheinlichkeit solcher Übergänge nicht allein durch mikroökonomische Faktoren, also Eigenschaften der Personen als zugrunde liegenden Untersuchungseinheiten, beeinflusst werden. Vielmehr kommen auch gesamtwirtschaftliche Größen als Determinanten hierfür in Betracht und sollten daher in die Analyse einbezogen werden. Aufgrunddessen werden auch in dieser Arbeit makroökonomische Daten als Hintergrundinformation einbezogen. Als Indikator für die gesamtwirtschaftliche Tätigkeit kann hier eine Klassifikation der jeweiligen Phase des Konjunkturzyklus Berücksichtigung finden. Der Konjunkturzyklus bezeichnet die alternierende, schwingungsförmige Zu- und Abnahme gesamtwirtschaftlicher Aktivität und ist schematisch in Abbildung 4.3 dargestellt. Je nach Definition und Modellierung wird der Konjunkturzyklus in zwei Phasen Aufschwung und Abschwung oder aber in vier

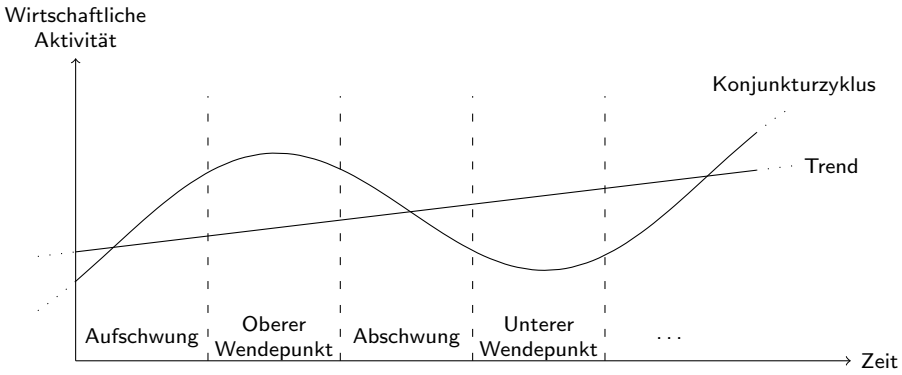


Abbildung 4.3: Schematische Darstellung des Konjunkturzyklus

Phasen, die außer Auf- und Abschwung auch noch Phasen für die Wendepunkte von Auf- zu Abschwung und umgekehrt umfassen, unterteilt. Zeitliche Terminierungen der Wechsel zwischen den Konjunkturzyklusphasen in einem Zwei-Phasen-Modell werden beispielsweise vom Economic Cycle Research Institute[3] auch für Deutschland veröffentlicht. Heilemann und Münch (2007) terminieren Phasenwechsel in einem Vier-Phasen-Modell. Zusätzlich bzw. alternativ zur Konjunkturzyklusklassifikation können und sollen weitere, tiefer gegliederte Daten der Gesamtwirtschaft einbezogen werden. Hierunter fallen etwa das gemeinhin als Wirtschaftswachstum bezeichnete Wachstum des Bruttoinlandsprodukts oder die Inflation, also die Steigerung der Verbraucherpreise. Basierend auf Daten der Organization for Economic Cooperation and Development (OECD) sind diese für Deutschland im durch die SOEP-Daten abgedeckten Zeitraum in Abbildung 4.4 dargestellt.

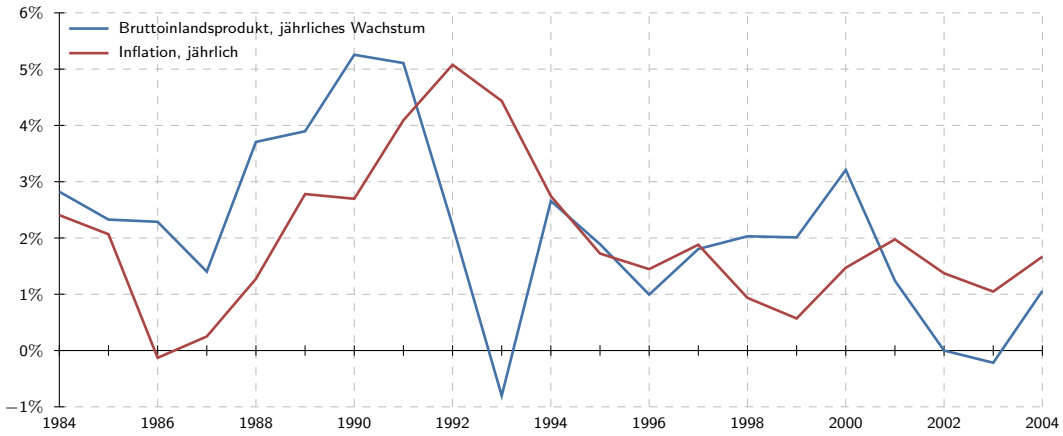


Abbildung 4.4: Bruttoinlandsprodukt und Inflation (1984-2004), Quelle: OECD

Daneben ist es interessant, weitere - weitestgehend exogene - Indikatoren einzubeziehen. Hier wäre beispielsweise die in den jeweiligen Wellen gültige Legislaturperiode zu nennen oder ein Indikator, der Zeiträume *vor* und *nach* der Wiedervereinigung von einander abgrenzt. Die zugehörigen Daten sind zusammen mit einem Indikator, der die jeweilige

[3]Das Economic Cycle Research Institute (ECRI) ist ein unabhängiges Institut, welches sich mit der Erforschung und Vorhersage von Konjunkturzyklen befasst. Siehe auch `http://www.businesscycle.com`.

Konjunkturzyklusphase angibt, in Tabelle 4.2 aufgelistet. Die Konjunkturzyklusphase der einzelnen Jahre basiert auf der oben erwähnten Einteilung des Konjunkturzyklus in zwei Phasen durch das ECRI. Diese vom ECRI auf Monatsbasis herausgegebene Einteilung wird dergestalt in eine jährliche transformiert, dass Jahre, die ganz einer der Phasen Auf- bzw. Abschwung zugeordnet werden können, durch diese Phase gekennzeichnet werden. Jahre, in denen ein Wechsel von Auf- zu Abschwung oder umgekehrt stattfindet, werden als Wendepunkt gekennzeichnet. Die Legislaturperiode ergibt sich aus den Wahlperioden des Bundestages der Bundesrepublik Deutschland. Wahljahre, in denen sich die Legislaturperiode ändert, werden dabei noch der Periode vor dem Wechsel zugerechnet. Der Indikator Wiedervereinigung schließlich ist bis einschließlich Jahr 1990, in dem die Wiedervereinigung offiziell in Kraft trat, als noch nicht eingetreten gekennzeichnet, ab 1991 dann als eingetreten. Die Zurechnung der Jahre, in denen ein Wechsel der genannten Indikatoren Legislaturperiode bzw. Wiedervereinigung stattfand, zu dem vorherigen Status kann dadurch begründet werden, dass mögliche Einflüsse, die dieser Wechsel auf den Arbeitsmarkt und auf entsprechende Übergänge von Arbeitsmarktzuständen hat, vermutlich nur langsam und verzögert auftreten.

Tabelle 4.2: Gesamtwirtschaftliche und politische Indikatordaten (1984-2004)

Jahr	Zyklus	Legislaturperiode	Wiedervereinigung
1984	Aufschwung	10	Nein
1985	Aufschwung	10	Nein
1986	Aufschwung	10	Nein
1987	Aufschwung	11	Nein
1988	Aufschwung	11	Nein
1989	Aufschwung	11	Nein
1990	Wendepunkt	11	Nein
1991	Abschwung	12	Ja
1992	Abschwung	12	Ja
1993	Abschwung	12	Ja
1994	Wendepunkt	12	Ja
1995	Aufschwung	13	Ja
1996	Aufschwung	13	Ja
1997	Aufschwung	13	Ja
1998	Aufschwung	13	Ja
1999	Aufschwung	14	Ja
2000	Aufschwung	14	Ja
2001	Wendepunkt	14	Ja
2002	Abschwung	14	Ja
2003	Wendepunkt	15	Ja
2004	Aufschwung	15	Ja

5 Extraktion von Paneldaten aus dem SOEP

Wie im letzten Kapitel beschrieben liegen die Daten des SOEP aufgrund des großen Datenumfangs in einer Reihe von einzelnen Dateien vor. Um Analysen auf Daten des SOEP durchführen zu können, bedarf es daher in der Regel einer Extraktion einzelner Daten aus diesen Dateien und einer anschließenden Verknüpfung dieser Daten. Vor der technischen Durchführung dieser Extraktion und Verknüpfung steht jedoch immer die Auswahl der Daten, die in die spätere Analyse einbezogen werden sollen. Grundsätzlich besteht die eigentliche Akquirierung von Daten daher aus zwei Komponenten: (1) der Selektion von Daten gemäß der Analysevorgaben und (2) der technischen Extraktion der Daten aus den SOEP-Dateien und Verknüpfung dieser Daten zu einem einzelnen Datensatz für die spätere Analyse. Aufgrund der Fülle von Variablen im SOEP und der eher kryptischen Benennung dieser Variablen, ist eine rein manuelle Selektion der Variablen ineffizient und nahezu unmöglich. Vielmehr ist das Vorliegen einer Item-Correspondance-Tabelle notwendig. Daher existieren bereits zwei informationssystemartige Lösungen, die die zum SOEP bereitgestellte Item-Correspondance-Tabelle als Hintergrundinformation nutzen, um die Selektion von Variablen zu unterstützen. Diese bereits bestehenden Lösungen sowie die implizierten Möglichkeiten zur technischen Extraktion und Verknüpfung der Daten werden in Kapitel 5.1 dargestellt. Da die bestehenden Lösungen diverse Nachteile aufweisen, wurde im Rahmen dieser Arbeit eine eigene Lösung implementiert, die sowohl die Selektion der Daten unterstützt als auch die technische Extraktion und Verknüpfung der Daten eigenständig löst. Das resultierende Produkt wird in Kapitel 5.2 vorgestellt. Anschließend beschreibt Kapitel 5.3 kurz die Daten, die unter Berücksichtigung der potentiell wichtigen, in Kapitel 4.3 benannten Einflussgrößen der Arbeitslosigkeit mittels des entwickelten Programms aus dem SOEP-Datensatz extrahiert wurden.

5.1 Bestehende Lösungen

Die unterstützte Selektion von Variablen zur späteren Analyse konnte bislang vor allem mit Hilfe von zwei bestehenden Lösungen durchgeführt werden. Dies ist zum einen das web-basierte Informationssystem *SOEPinfo*[1], welches von der SOEP-Gruppe am Deutschen Institut für Wirtschaftsforschung zur Verfügung gestellt wird. Des Weiteren existiert ein Add-On zur Statistiksoftware *Stata* namens *PanelWhiz*[2]. Beide werden im Folgenden kurz beschrieben.

5.1.1 SOEPinfo

Das web-basierte Informationssystem *SOEPinfo* (vgl. Abbildung 5.1) unterstützt die Auswahl von Variablen im SOEP zur späteren Analyse durch Abbildung unterschiedlicher Korrespondenzen zwischen den Variablen und darauf basierender Suchmöglichkeiten. Zum

[1] `http://panel.gsoep.de/soepinfo/`
[2] `http://www.panelwhiz.eu`

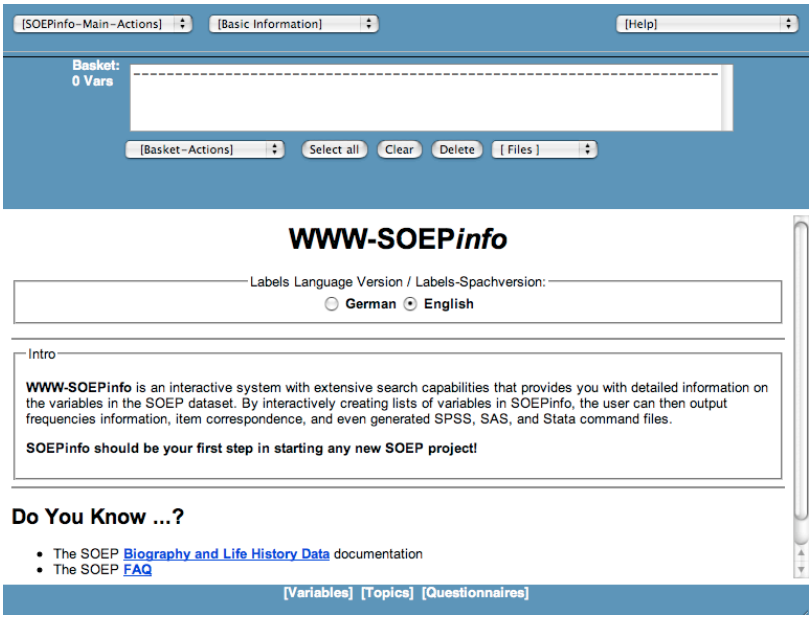


Abbildung 5.1: Web-basierte Oberfläche von SOEPinfo

einen berücksichtigt SOEPinfo die zum SOEP bereit gestellt Item-Correspondance-Tabelle und erlaubt damit die Zuordnung von Variablen zu Items und umgekehrt. Dies bedeutet, es können sowohl zu einer Variable das zugehörige Item als auch zu einem Item die zugehörigen Variablen identifiziert werden und somit auch zu einer Variable die weiteren Variablen desselben Items gefunden werden. Des Weiteren enthält SOEPinfo die in der Item-Correspondance-Tabelle enthaltene Themenstruktur der Items, sodass zu einem Thema im SOEP vorhandene Variablen identifiziert werden können. Zusätzlich dazu bietet SOEPinfo eine Ansicht der jeweils verwendeten Fragebögen, in der direkte Verknüpfungen zu den die Fragen repräsentierenden Variablen vorhanden sind. Es ist ebenfalls möglich, die in den einzelnen Dateien des SOEP-Datensatzes befindlichen Variablen aufzulisten. Die hier aufgelisteten, umfangreichen Optionen, um Variablen zu suchen, werden ergänzt durch die Möglichkeit, die Werte der einzelnen Variablen sowie die absoluten und relativen Häufigkeiten ihres Auftretens anzuzeigen.

Zur eigentlichen Selektion der Variablen existiert eine Liste (genannt *Basket*), in der ausgewählte Variablen gespeichert werden können. Baskets können zur Sicherung lokal abgespeichert werden sowie für spätere Änderungen in SOEPinfo hochgeladen werden. Eine Möglichkeit der direkten Extraktion und Verknüpfung der ausgewählten Daten ist mittels SOEPinfo nicht möglich. Es existiert allerdings die Möglichkeit, nach Auswahl der Variablen Programmcode für verschiedene Statistiksoftwarepakete zu erzeugen, bei dessen Anwendung innerhalb dieser Statistiksoftwarepakete die Extraktion automatisch durchführt wird. Um eine solche Extraktion durchführen zu können und SOEPinfo damit sinnvoll und umfassend nutzen zu können, muss somit die Verfügbarkeit eines der unterstützten Statistikprogramme (SPSS, Stata oder SAS) gegeben sein. Keines dieser drei Programmpakete ist jedoch frei verfügbar.

Zusammengefasst bietet SOEPinfo zwar den Vorzug vielfältiger Suchmöglichkeiten bei der Auswahl von Variablen, die sich aus der Kombination der Metadaten aus der Item-

Correspondance-Tabelle und den Fragebögen ergeben. Nachteilig ist jedoch, dass SOEPinfo erstens nur online zugreifbar und die Nutzung des Informationssystems zur Selektion von Variablen nur online durchführbar ist. Zweitens existiert keine Unterstützung für frei verfügbare Systeme zur technischen Extraktion und Verknüpfung der Daten.

5.1.2 PanelWhiz

Ähnliche Unterstützung bei der Selektion von Variablen wie SOEPinfo bietet *PanelWhiz*. PanelWhiz ist ein modulares Interface für das Statistikprogramm Stata. Modular bedeutet in diesem Fall, dass es neben der Selektion und Extraktion von Daten aus dem SOEP auch für einige weitere Paneldatensätze geeignet ist. Bezüglich des SOEP inkorporiert PanelWhiz ebenfalls die verfügbaren Metadaten, d.h. die Korrespondenz der Variablen und Items untereinander sowie die thematische Struktur der Items. Somit ergeben sich die gleichen Möglichkeiten der Variablensuche wie bei SOEPinfo. Sich der Funktionalität von Stata bedienend bietet PanelWhiz zudem die Möglichkeit, die selektierten Variablen tatsächlich aus den Dateien zu extrahieren und zu verknüpfen. Ein Vorteil von PanelWhiz in Kombination mit Stata sind dabei die umfangreichen, in PanelWhiz implementierten Möglichkeiten der automatisierten Datenbereinigung (beispielsweise hinsichtlich Inkonsistenzen in den möglichen Variablenwerten über die Wellen hinweg) und Datentransformation. Ein weiterer unbestrittener Vorteil ist die Möglichkeit, einmal gespeicherte Variablenselektionen bei der Veröffentlichung neuer Versionen der Paneldatensätze automatisch auf diese anzupassen.

Der Nachteil von PanelWhiz, der im Rahmen dieser Arbeit die Anwendung von PanelWhiz ausschließt, ist wie bei SOEPinfo, dass die Verfügbarkeit eines nicht freien Softwarepaketes verlangt wird. Für eine detailliertere Beschreibung der Funktionen von PanelWhiz sei an dieser Stelle auf die Dokumentation durch Haisken-DeNew und Hahn (2006) verwiesen.

5.2 PanelX

Aufgrund der angesprochenen Nachteile wurde im Rahmen dieser Arbeit eine eigene Lösung namens *PanelX* (für *engl.* PanelExtraction) implementiert, die beide oben angesprochenen Elemente, d.h. die unterstützte Selektion von Variablen und die eigentliche Extraktion und Verknüpfung der zugehörigen Paneldaten, erlaubt. Neben diesen primären Aufgaben und Anwendungszwecken der implementierten Lösung waren zusätzliche Anforderungen bei der Entwicklung des resultierenden Programms maßgeblich. Zum einen sollte die Lösung natürlich auf Daten des SOEP anwendbar sein. Zudem sollte eine Ausgabe der extrahierten Daten in einer für die zu verwendende Analysesoftware lesbaren Form garantiert werden. Gleichzeitig sollte jedoch ein möglichst modularer Aufbau eine Erweiterbarkeit und Wiederverwendbarkeit hinsichtlich neuer Versionen des SOEP-Datensatzes aber auch anderer Paneldatensätze sowie anderer Dateiformate für Ein- und Ausgabe gewährleisten. Im Folgenden wird das entwickelte Programm kurz hinsichtlich seines Aufbaus und seiner Funktionalität beschrieben. Da vor allem die spätere Anwendung des Programms zur tatsächlichen Auswahl und Extraktion von Daten aus dem SOEP-Datensatz relevant ist, wird dabei auf die Beschreibung von Implementierungsdetails im Wesentlichen verzichtet.

Der Aufbau des entwickelten Programms ist in Abbildung 5.2 rudimentär[3] dargestellt. Kern des Programms sind die Klassen **Panel**, **Topic**, **Item** und **Variable** sowie **Panel-**

[3]Es werden nur die wichtigsten, d.h. die für das Verständnis des Aufbaus notwendigen Klassen abgebildet.

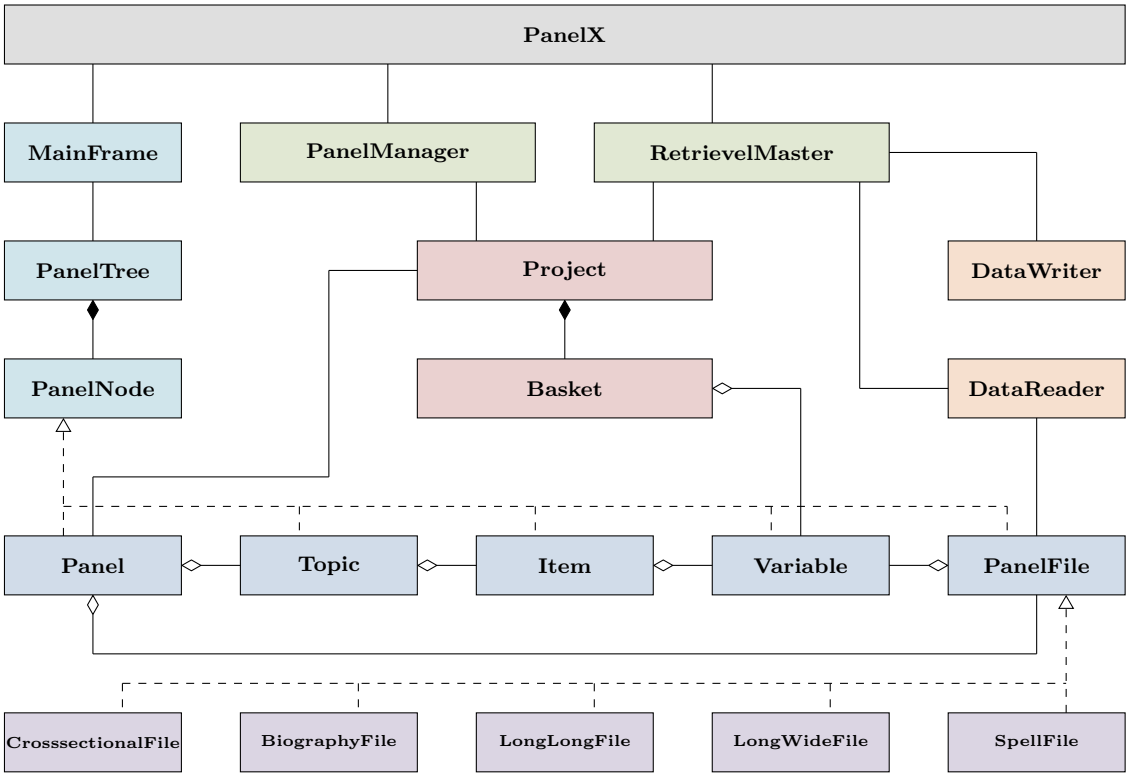


Abbildung 5.2: Rudimentäres Klassendiagramm von PanelX

File (und deren Subklassen). Sie erlauben zum einen die Abbildung einer hierarchischen thematischen Struktur, wie sie im SOEP durch die Bereitstellung von Themen in der Item-Correspondance-Tabelle vorhanden ist. Diese thematische Struktur wird mit Hilfe der Klasse **PanelManager** durch Auswertung einer bereitgestellten Item-Correspondance-Tabelle aufgebaut. Zum zweiten erlauben die Kernklassen die Abbildung der Struktur eines Paneldatensatzes, in dem die Daten zu Variablen in verschiedenen Dateien vorhanden sind. Hierbei werden verschiedene Typen von Dateien unterstützt, die bei der Extraktion und Verknüpfung der Daten beachtet und unterschiedlich behandelt werden müssen (siehe hierzu auch Kapitel 5.2.3). Eine grafische Darstellung der thematischen Struktur sowie der Datei-Struktur erfolgt durch die Klasse **PanelTree** und die soeben aufgelisteten Kernklassen selbst, die als Subklassen von **PanelNode** ihre grafische Darstellung unterstützen. Zur Unterstützung der Extraktion stehen die Klassen **Project** und **Basket** bereit. Die Klasse **Basket** verwaltet dabei eine Liste von durch den Benutzer hinzugefügter Variablen, die später aus dem Datensatz extrahiert werden sollen. Die Klasse **MainFrame** stellt eine umfassende Benutzerschnittstelle zur Verfügung, innerhalb derer die angesprochene grafische Darstellung der Panelansichten erfolgt, sowie Benutzeraktionen (wie etwa das Hinzufügen von Variablen zum Basket) durchgeführt werden können. Die eigentliche Extraktion und Verknüpfung der Daten (von den sich im Basket befindlichen Variablen) erfolgt durch die Klasse **RetrievalMaster** in Verbindung mit den Klassen **DataReader**, die Daten aus dem Paneldatensatz einliest, und **DataWriter**, die die resultierenden, extrahierten Daten abspeichert.

5.2.1 Benötigte Metainformationen

Zum Aufbau der inneren Datenstrukturen sowie zur späteren Extraktion und Verknüpfung benötigt PanelX (bzw. die Klasse **PanelManager** als die Datenstrukturen verwaltende Klasse) einige Metainformationen zum Paneldatensatz, der betrachtet bzw. aus dem Daten extrahiert werden sollen. Dies umfasst vor allem eine Auflistung der Dateien, aus denen der Datensatz besteht sowie der Art dieser Dateien bzw. deren Form (siehe hierzu auch Kapitel 5.2.3). Außerdem muss für jede Datei die Schlüsselvariable angegeben werden, anhand derer die Daten für eine in dieser Datei betrachtete Untersuchungseinheit identifiziert werden können. Des Weiteren wird eine Angabe der einzelnen Wellen benötigt. Für Dateien, die Variablen enthalten, die zu unterschiedlichen Wellen gehören, muss zudem eine Angabe der Zuordnung von Variablen zu Wellen erfolgen. Da im Rahmen dieser Arbeit nur das SOEP als Paneldatensatz (und des Weiteren nur eine Version dieses Datensatzes) betrachtet wird, enthält die Implementierung von PanelX bislang nur eine statische Angabe der oben genannten Sachverhalte. PanelX kann diesbezüglich jedoch leicht um die Option einer einfachen Konfigurierbarkeit (beispielsweise durch eine XML-Datei) hinsichtlich dieser Sachverhalte ergänzt werden und so generisch - auch auf anderen Paneldatensätzen - eingesetzt werden.

5.2.2 Selektion von Variablen

PanelX unterstützt die Selektion von Variablen durch eine grafische Darstellung der im Panel enthaltenen Variablen auf der einen Seite (Panelansicht) und der zu einem Basket hinzugefügten Variablen auf der anderen Seite (Basketansicht). Die Panelansicht bietet mehrere Optionen zur Darstellung der im Panel enthaltenen Variablen. Wie bereits erwähnt, bildet PanelX so erstens die vom SOEP zur Verfügung gestellte Themenstruktur mittels entsprechender Klassen ab. Aus der in der Item-Correspondance-Tabelle enthaltenen Information über Themen und ihre hierarchische Anordnung kann so eine Baumstruktur aus Panel (als Wurzelknoten), Themen, Items und Variablen erzeugt werden. Diese Baumstruktur bietet für den Benutzer die Möglichkeit, bequem nach Variablen hinsichtlich eines Themas für durchzuführende Analysen zu suchen. Zur Verdeutlichung zeigt Abbildung 5.3 die PanelX-Oberfläche und einen exemplarischen Ausschnitt der hierarchischen Baumstruktur der Themen des SOEP und zugehöriger Items und Variablen. Zur Unterstützung der Variablenselektion werden zu jedem Knoten des Baumes zugehörige Informationen angezeigt. Für Variablen werden beispielsweise die zugehörige Welle (sofern die Variablen explizit Daten für eine Welle erfassen) und die möglichen Werte der Variablen angezeigt.

Neben der Ansicht der Themenstruktur bietet PanelX dem Nutzer zusätzlich eine Ansicht der Dateistruktur des Datensatzes. Ebenfalls baumartig stellt diese die zum Panel (als Wurzelknoten) gehörenden Dateien und die in diesen Dateien enthaltenen Variablen dar. Dem Benutzer wird damit die Möglichkeit eröffnet, den Datensatz auch auf Dateiebene hinsichtlich in die Analyse einzubeziehender Variablen zu durchsuchen. Dies erleichtert zum einen das gezielte Auffinden von Variablen, die beispielsweise zuvor durch Konsultation der SOEP-Fragebögen zur Selektion identifiziert wurden, und bietet zum anderen (zumindest bei der Struktur des SOEP-Datensatzes) einen direkten Überblick, welche Variablen etwa in bestimmten Wellen oder in bestimmten Fragebögen (z.B. dem Biographiefragebogen) erfasst wurden.

Eine dritte Darstellungsmöglichkeit in PanelX erlaubt eine Volltextsuche in den Themen,

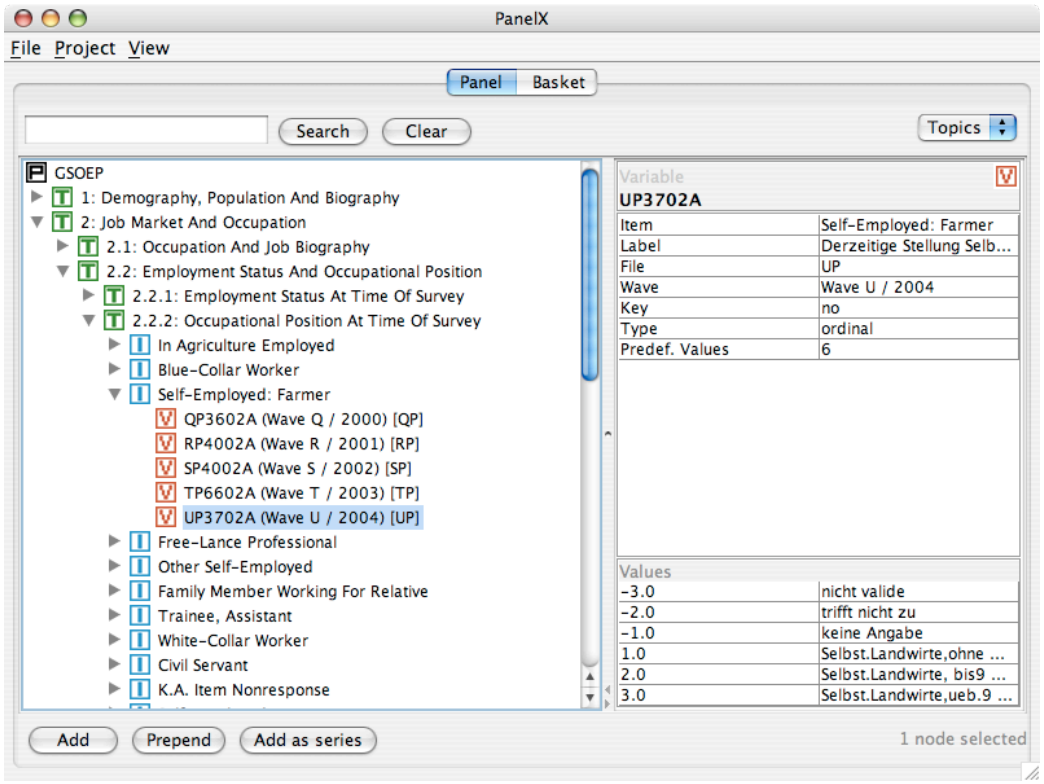


Abbildung 5.3: Oberfläche von PanelX: Darstellung der hierarchischen Themenstruktur des SOEP

Items und Variablennamen. Die Darstellung des Ergebnisses einer solchen Suche erfolgt wiederum baumartig. So werden etwa Themen, die den gesuchten Text in ihrer Beschreibung beinhalten, mit allen Unterthemen, den enthaltenen Items und Variablen dargestellt.

In allen drei beschriebenen Ansichten können Variablen markiert und über den Button *Add* zum Basket hinzugefügt werden. Der Button *Add* bewirkt dabei ein Anfügen der Variablen am Ende der im Basket enthaltenen Variablenliste. Der Button *Prepend* fügt die Variablen am Anfang der Liste ein.

Die Basketansicht zeigt die so hinzugefügten Variablen auf eine den bisherigen Darstellungen ähnelnden Weise. Sie bietet ebenfalls die Anzeige der bereits angesprochenen Informationen über die einzelnen, hinzugefügten Variablen. Funktional bietet sich in dieser Ansicht die Möglichkeit, die sich im Basket befindlichen Variablen aus diesem zu löschen sowie ihre Reihenfolge zu ändern. Baskets mit selektierten Variablen können zur späteren Veränderung bzw. Erweiterung gespeichert sowie natürlich auch geladen werden. Die persistente Speicherung erfolgt in einem relativ einfachen XML-Format, welches in Abbildung 5.4 anhand eines einfachen Beispiels nachvollzogen werden kann.

5.2.3 Extraktion und Verknüpfung der Variablen

Neben der grafisch unterstützten Selektion von Variablen aus einem Paneldatensatz ermöglicht PanelX auch die tatsächliche Extraktion der zu diesen Variablen gehörenden Daten und ihre Verknüpfung. Die Extraktion und Verknüpfung wird hauptsächlich innerhalb der

```
<?xml version="1.0"?>
<project>
    <panel/>
    <basket>
        <variable name="BP5201" file="BP"/>
        <variable name="JP4401" file="JP"/>
        <variable name="JP4401" file="JPLUECKE"/>
        <variable name="LP7201" file="LP"/>
        <variable name="LP7201" file="LPLUECKE"/>
    </basket>
</project>
```

Abbildung 5.4: Exemplarischer XML-Code für ein PanelX-Projekt

Klasse **RetrievalMaster** durchgeführt und geschieht auf Basis der (bislang statisch im Programmcode) bereitgestellten Metainformationen zur Form des Datensatzes. Hierbei spielen vor allem die Form der einzelnen Dateien sowie die als Schlüssel identifizierten Variablen in den einzelnen Dateien eine Rolle. Hinsichtlich der Form werden durch die bereits in Abbildung 5.2 dargestellten Subklassen der Klasse **PanelFile** fünf verschiedene Dateitypen unterstützt. Diese sind in Tabelle 5.1 mit den sie bestimmenden Eigenschaften aufgelistet.

Tabelle 5.1: Von PanelX unterstützte Dateitypen

Typ	Beschreibung	Beispiele
CrosssectionalFile	enthält Daten für eine bestimmte Welle	xP, xHGEN, ...
BiographyFile	enthält nur einmalig erhobene Daten	BIOSOC, BIOYOUTH
LongWideFile	enthält Paneldaten für mehrere Wellen in Ein-Tabellen-Form	PPFAD
LongLongFile	enthält Paneldaten im Long-Format	BIOIMMIG
SpellFile	enthält Daten im Spell-Format	ARTKALEN

Die Klasse **CrosssectionalFile** repräsentiert Dateien, die nur Daten für eine bestimmte Welle (die für jede solcher Dateien bekannt sein muss) enthalten. Die Daten aus einem **BiographyFile** beziehen sich auf keinen speziellen Zeitpunkt, sondern sind eher an der Biographie der Untersuchungseinheiten ausgerichtet. Insofern kann der Erhebungszeitpunkt im Regelfall vernachlässigt werden. Dateien vom Typ **LongWideFile** enthalten Daten im Ein-Tabellen-Format (siehe Abbildung 2.6(b)). Sie beinhalten somit Variablen für Daten aus unterschiedlichen Wellen. Für sie muss eine Zuordnung (sofern sie nicht bereits durch die Item-Correspondance-Tabelle klar wird) zwischen Variablen und Wellen bekannt sein. Für alle drei bisher vorgestellten Dateitypen genügt ein Schlüssel, der die jeweiligen erfassten Untersuchungseinheiten identifiziert, um Daten aus Dateien dieser Typen zu verknüpfen. Dies gilt jedoch nicht für Dateien des Typs **LongLongFile**. Diese liegen im Long-Format (siehe Abbildung 2.7) vor, bei denen der Zeitbezug explizit durch eine gesonderte Variable berücksichtig wird. Diese Variable muss bei der Verknüpfung von Daten dieses Formats mit anderen Daten einbezogen werden. Ähnliches gilt für Dateien des Typs **SpellFile**. Sie beinhalten Spells, die einen Start- und Endzeitpunkt haben, die ebenfalls bei der Extraktion und Verknüpfung der Daten mit anderen Daten berücksichtigt werden müssen.

Neben den Typen der einzelnen Dateien sind die Identifikationsschlüssel, die Analyseein-

heiten innerhalb der Dateien identifizieren lassen, von hoher Bedeutung. Auch sie müssen für jede Datei explizit angegeben werden. Da sich die im Rahmen dieser Arbeit durchgeführte Analyseaufgabe lediglich auf Personen als Untersuchungseinheiten bezieht, unterstützt PanelX zunächst nur Extraktionen von Personendaten aus dem SOEP. Die folgende Beschreibung bezieht sich daher auch nur auf diesen Fall. Eine Extraktion auf Haushaltsebene verliefe jedoch vollkommen analog; PanelX kann daher auch leicht um die Unterstützung solcher Extraktionen erweitert werden.

Zu Beginn einer Extraktion von Paneldaten aus dem Datensatz müssen zu den sich im Basket befindlichen Variablen die zugehörigen Schlüssel gefunden werden. Anschließend wird die dem Panel zugrundeliegende Indexdatei (hier: PPFAD) bzw. die vorher identifizierten Schlüssel aus dieser Datei gelesen. Diese Daten bilden die initiale Datentabelle, die um weitere Daten ergänzt wird. Hierzu werden für jede Variable aus dem Basket die Daten für diese Variable sowie die Identifikationsschlüssel aus der zugehörigen Datei (in der sich die Variable befindet) gelesen und so eine zweite Datentabelle erzeugt. In den Fällen, in denen die Datei ein **CrosssectionalFile**, **BiographyFile** oder **LongWideFile** ist, genügt dann ein einfaches Matching der zuvor aus der Indexdatei gelesenen Datentabelle mit der neu gelesenen Datentabelle über den jeweiligen Schlüssel. Technisch betrachtet werden dazu zunächst beide Datentabellen nach dem Identifikationsschlüssel aufsteigend sortiert. Anschließend werden beide Dateien abwechselnd bis zur nächsten Übereinstimmung der Schlüssel durchlaufen. Im Fall einer Übereinstimmung werden die neu gelesenen Daten zugeordnet in die neue Datentabelle eingefügt. In Zeilen der Indexdatentabelle, für die die zweite Datentabelle keine Werte enthält (etwa aufgrund von Panelabwanderung oder Wave Nonresponse der betroffenen Person), werden die Daten bzgl. der neu gelesenen Variable als fehlend gekennzeichnet. Für Variablen, die aus Dateien der Typen **LongLongFile** und **SpellFile** gelesen werden, muss vor dem oben beschriebenen Matching zunächst eine Transformation der gelesenen Daten erfolgen. Dazu werden die Daten in eine Ein-Tabellen-Form überführt, bei der eine implizite Berücksichtigung der Zeit durch Einführung mehrerer Variablen, die Beobachtungen zu einzelnen Zeitpunkten darstellen, erfolgt. Dies geschieht auf Basis der ebenfalls aus den Dateien gelesenen Erhebungszeitpunkte bzw. der Anfangs- und Endzeitpunkte der Spells.

Das eigentliche Auslesen der Daten aus den Dateien des Datensatzes geschieht mittels der Klasse **DataReader** bzw. ihrer Subklassen. Im Rahmen dieser Arbeit wurde der SOEP-Datensatz im SPSS-Format verwendet. Daher wurde als bislang einzige Subklasse die Klasse **SPSSDataReader** implementiert. Für andere Eingabeformate kann PanelX leicht um weitere Subklassen der Klasse **DataReader** ergänzt werden. Die unter Anwendung des oben beschriebenen Verfahrens aus der Extraktion und Verknüpfung resultierenden Daten können mittels der Klasse **DataWriter** gespeichert werden. Auch hier wurde nur eine Subklasse **ArffDataWriter**, die die Daten im für die späteren Analysen passenden ARFF-Format abspeichert, implementiert; die Implementierung weiterer Subklassen ist jedoch ebenso leicht möglich.

Im Gegensatz zum Eingabeformat, d.h. zum SPSS-Format, in dem die einzelnen Dateien des SOEP-Datensatzes in PanelX eingelesen werden (siehe hierzu auch Kapitel 3.3.1), unterstützt das ARFF-Format keine Kodierung nominaler Werte durch numerische Werte. Aus diesem Grund ersetzt PanelX bei der Speicherung von Daten, die aus dem SOEP-Datensatz extrahiert wurden, alle numerischen Repräsentanten nominaler Werte durch die eigentlichen, als Zeichenketten gegebenen Label der nominalen Werte. Werte, die gemäß der oben beschriebenen Vorgehensweise bei der Verknüpfung der Daten als fehlend gekenn-

zeichnet werden, weil sie im SOEP nicht vorhanden sind, werden gemäß dem ARFF-Format als Fragezeichen (*?*) gespeichert.

5.3 Extrahierte Daten

Gemäß der Betrachtungen aus Kapitel 4.3, in denen potentielle Einflussfaktoren der Arbeitslosigkeit genannt wurden, wurden in einem ersten Schritt Variablen im SOEP identifiziert, die diese Faktoren möglichst zutreffend widerspiegeln. Die zu diesen Variablen gehörigen Daten wurden mit Hilfe von PanelX aus dem SOEP-Datensatz extrahiert. Hierzu wurden die in Tabelle 5.2 aufgelisteten Variablen selektiert. Die ersten beiden Gruppen von Variablen (`xNETTO` und `xPHRF`) dienen später eher technischen Aspekten. Dies ist zum einen die Identifikation der tatsächlich erfolgreich befragten Personen im SOEP und damit der Ausschluss von Personen, die die Befragung verweigern bzw. von nicht mit in die spätere Analyse einzubeziehenden Kindern, die im SOEP und daher in der Index-Datei `PPFAD` ebenfalls erfasst sind. Zum anderen werden damit Querschnittsgewichte, also Gewichte für die Personen in den einzelnen Wellen bereitgestellt, die bei der späteren eigentlichen Datenanalyse berücksichtigt werden müssen. Die folgenden Variablen `LFSxx` sind die schon in Kapitel 4.2 beschriebenen Arbeitsmarktzustände, aus denen Arbeitsmarktzustandsübergänge als Zielattribute konstruiert werden sollen. Die im darauffolgenden Teil der Tabelle aufgelisteten Variablen sollen die in Kapitel 4.3 identifizierten Faktoren, die potentiell einen Einfluss auf die betrachteten Übergänge zwischen Arbeitsmarktzuständen haben, erfassen. Hierbei ist zu beachten, dass für nicht statische Items, d.h. Items, die sich über die Zeit ändern können, stets alle Variablen - d.h. die Variablen für alle Wellen - selektiert und extrahiert wurden. Die Auswahl der Items und damit der genannten Variablen ist hierbei ein Kompromiss zwischen einer Auswahl möglichst vieler Faktoren zur Entdeckung möglicherweise unerwarteter Zusammenhänge auf der einen Seite und einer noch handhabbaren Datensatzgröße (in Bezug sowohl auf Komplexität als auch auf technische Aspekte) auf der anderen Seite. Die Auswahl ist daher eher als initiale Grundlage eines iterativen Analyseprozesses aufzufassen, die bei Bedarf erweitert oder verändert werden kann. Bezüglich der Auswahl der Variablen im Einzelnen ist anzumerken, dass beispielsweise das Geburtsjahr der Person bzw. die Geburtsjahre derer Kinder nicht zwangsläufig direkt in die spätere Analyse einfließen sollen. Vielmehr wurden diese Variablen einbezogen, um eine Möglichkeit bereitzustellen, etwa das Alter oder die Anzahl der Kinder einer Person zu einem bestimmten Zeitpunkt berechnen zu können. Diese Attribute werden im SOEP nicht explizit bereitgestellt, können jedoch aus den angegebenen Variablen für die gewünschten Zeitpunkte berechnet werden. Dies ist Teil der nachfolgenden Datenvorbereitung (siehe hierzu Kapitel 6). Die im unteren Teil der Tabelle genannten Variablen sollen additiv zu den generell einzubeziehenden Variablen Aspekte der Erwerbstätigkeit von Personen für die Lernaufgaben, die Zustandsübergänge aus der Erwerbstätigkeit heraus betrachten, erfassen. Neben den explizit selektierten Variablen umfassen die extrahierten Daten implizit zusätzlich die bei der Extraktion zur Verknüpfung der Daten benötigten Identifikationsschlüssel. Dies sind erstens der eindeutige Personenschlüssel `PERSNR` und zweitens die - für das Hinzufügen der Variablen auf Haushaltsebene (`xTYPHH` und `xHHGR`) benötigten - Haushaltsschlüssel `HHNR` sowie `AHHNR` bis `UHHNR`. Diese können später dazu dienen, weitere Transformationen der Daten durchzuführen, dürfen jedoch nicht in die eigentliche Datenanalyse mit einbezogen werden.

Tabelle 5.2: Zur Extraktion selektierte Variablen. Die Platzhalter `x` bzw. `xx` sind dabei durch Buchstaben `A` bis `U` bzw. - im zweiten Fall - Ziffern `84` bis `04`, die die einzelnen Wellen indizieren, zu ersetzen. Dementsprechend sind hier immer alle Variablen eines Items - d.h. die Variablen für alle Wellen gemeint.

Variablenname(n)	Dateien	Item-Beschreibung
xNETTO	PPFAD	Befragungsstatus
xPHRF	PHRF	Hochrechnungsfaktor/Gewicht
LFSxx	xPGEN	Arbeitsmarktzustand
SEX	PPFAD	Geschlecht
GEBJAHR	PPFAD	Geburtsjahr
GEBMONAT	PPFAD	Geburtsmonat
GERMBORN	PPFAD	in Deutschland geboren
CORIGIN	PPFAD	Herkunftsland
IMMIYEAR	PPFAD	Jahr der Immigration
ORTKINDH	PPFAD	Ort der Kindheit
ORTKIND1	PPFAD	lebt noch am Ort der Kindheit?
LOC1989	PPFAD	Aufenthalt im Jahr 1989
xBULA	xPGEN	Bundesland
xFAMSTD	xPGEN	Familienstand
xTYPHH2	xHGEN	Haushaltstyp
xHHGR	xHBRUTTO	Haushaltsgröße
AH06 ...UH53	xH	Kinder unter 16 Jahren im Haushalt?
KIDGEB01...15	BIOBIRTH/BIOBRTHM	Geburtsjahre der Kinder
xPSBIL	xPGEN	Schulbildung
xPBBIL01	xPGEN	Beruflicher Bildungsabschluss
xPBBIL02	xPGEN	Hochschulabschluss
xBILZEIT	xPGEN	Dauer der Ausbildung (in Jahren)
xNACE	xPGEN	Wirtschaftszweig/Branche
IS88xx	xPGEN	Berufsklassifikation nach ISCO88
OEFFDxx	xPGEN	im öffentlichen Dienst
ERLJOBxx	xPGEN	Tätigkeit im erlernten Beruf
AUSBxx	xPGEN	für Tätigkeit erforderliche Ausbildung
ERWZEITxx	xPGEN	Dauer der Betriebszugehörigkeit
AUTONOxx	xPGEN	berufliche Autonomie
BETRxx	xPGEN	Unternehmensgröße

Wie bereits beschrieben, speichert PanelX die Daten in der Ein-Tabellen-Form aus Abbildung 2.6(b). Somit liegen die extrahierten Daten in einer Datentabelle vor, deren Spaltenanzahl mit der Anzahl der Variablen übereinstimmt, die aus dem SOEP-Datensatz extrahiert wurden. Zu beachten ist hierbei, dass für jedes nicht statische Item jeweils 21 Variablen extrahiert wurden. Die Anzahl der Zeilen der Datentabelle entspricht der Anzahl aller jemals vom SOEP betrachteten und damit in der Datei `PPFAD` erfassten Personen. Basierend auf der oben aufgelisteten Variablenauswahl ergibt sich eine Tabellengröße von 466 Spalten (d.h. Variablen) und 56.150 Zeilen (d.h. Untersuchungseinheiten). Dieser Datensatz ist initiale Grundlage der in dieser Arbeit durchgeführten Analyse. Allerdings können Datenanalyseverfahren selbst noch nicht direkt auf dem Datensatz angewandt werden. Vielmehr bilden die extrahierten Daten die Eingabe für die nachfolgende, dem eigentlichen Datenanalyseschritt vorgelagerte Datenvorverarbeitung, welche in Kapitel 6 erläutert wird.

Da sich bei Analysevorhaben wie dem in dieser Arbeit beschriebenen meist iterative Vor-

gehensweisen nicht vermeiden lassen, etwa weil die Ergebnisse nicht eine erwartete Güte haben, können in weiteren Iterationsschritten des gesamten Analyseprozesses im Extraktionsschritt, der Thema dieses Kapitels war, weitere (zumeist alternative) Variablen hinzugefügt werden. Aufgrund der Unterstützung der persistenten Speicherung der Baskets in PanelX, die Listen zu extrahierender Variablen (wie etwa die in Tabelle 5.2) enthalten, und der einfachen Veränderbarkeit und Erweiterbarkeit dieser Listen, ist auch eine Extraktion eines vergrößerten oder modifizierten Datensatzes in analoger Vorgehensweise zu der oben beschriebenen vollkommen unproblematisch und somit leicht möglich.

6 Vorverarbeitung

Vor der Analyse bzw. der Anwendung von Verfahren zur Auswertung eines Datensatzes steht im Allgemeinen dessen *Vorverarbeitung* (*engl.* preprocessing). Darüber hinaus äußern Adriaans und Zantinge (1996) die weitestgehend als Konsens akzeptierte These, dass häufig bis zu 80 Prozent des Aufwandes bei der Analyse realer, d.h. nicht synthetisch erzeugter Daten - und damit der größte Teil - auf die Vor- und Aufbereitung der Daten entfällt. Die Gründe für die Notwendigkeit einer Bearbeitung eines Datensatzes vor dessen Analyse sind dabei mannigfaltig. Dazu zählen vor allem Eigenheiten (bzw. Unsauberkeiten) des Datensatzes, die aus der Datenaufzeichnung bzw. -erhebung resultieren. Hier sind etwa nicht einheitliche Wertekodierungen und die Existenz von fehlenden Werten zu nennen. Solche Störungen sind bei der Anwendung von Data-Mining-Verfahren höchst problematisch und verfälschen unter Umständen die resultierenden Ergebnisse. Aufgrunddessen muss eine Berücksichtigung bzw. Korrektur dieser Artefakte erfolgen. Doch die Vorverarbeitung eines Datensatzes besteht nicht nur aus Korrekturmaßnahmen. Vielmehr zählen etwa auch das Hinzufügen von Hintergrunddaten oder das Erzeugen neuer Attribute auf Basis der bereits vorhandenen zu den Maßnahmen, die während der Vorverarbeitung durchgeführt werden müssen.

Auch im Rahmen dieser Arbeit ist die Vorverarbeitung unverzichtbarer und absolut notwendiger Bestandteil des Analyseprozesses. Vielmehr macht die Vorverarbeitung den größten Teil des Aufwands bei der in dieser Arbeit durchgeführten Analyse aus und stützt damit die oben genannte Ansicht. Die Notwendigkeit der Vorverarbeitung extrahierter Daten folgt zum einen aus Eigenheiten des SOEP-Datensatzes, etwa bzgl. der Kodierung von fehlenden Werten oder inkonsistenter Wertekodierungen in Variablen eines Items in verschiedenen Wellen. Zum zweiten folgt die Notwendigkeit jedoch auch daraus, dass die aus dem Paneldatensatz extrahierten Daten in eine für das später anzuwendende Datenanalyseverfahren sinnvolle Form gebracht werden mussten. Im Folgenden werden die im Rahmen dieser Arbeit durchgeführten Schritte der Datenvorverarbeitung motiviert und erläutert. Im Einzelnen werden dabei folgende Komponenten erklärt: Kapitel 6.1 erläutert die vorgenommene Datenbereinigung zur Beseitigung bereits angesprochener Inkonsistenzen sowie in Bezug auf fehlende bzw. leere Werte. Kapitel 6.2 beleuchtet das Hinzufügen makroökonomischer Daten. Kapitel 6.3 erläutert die Erzeugung von Variablen, die Wechsel von Arbeitsmarktzuständen anzeigen, auf Basis von Variablen, die die jeweils aktuellen Arbeitsmarktzustände in den einzelnen Wellen angeben. Kapitel 6.4 beschreibt die Angleichung der Längen der extrahierten und erzeugten Wertereihen, die erfolgen muss, da die Länge der erzeugten Wertereihen der Zustandswechsel, also der späteren Zielattribute nicht mit der Länge der übrigen Wertereihen übereinstimmt. Kapitel 6.5 befasst sich dann mit der Transformation der Daten in eine für die spätere Analyse zweckdienliche Form. Basierend auf den transformierten Daten wurden außerdem weitere Merkmale generiert. Dies wird in Kapitel 6.6 erläutert. Da manche maschinelle Lernverfahren Einschränkungen hinsichtlich der Skalen der verwendeten Attribute machen, erklärt Kapitel 6.7 Verfahren, um numerische Attribute zu diskretisieren, d.h. in nominale Attribute zu überführen. Kapitel 6.8 nennt

Maßnahmen zur Beseitigung für die Analyse irrelevanter Daten. Kapitel 6.9 fasst schließlich die Ergebnisse zusammen und beschreibt die aus der Vorverarbeitung resultierenden Daten. Abschließend stellt Kapitel 6.10 das im Anschluss an die Vorverarbeitung für die eigentliche Anwendung von Data-Mining-Verfahren verwendete Experimentier-Framework in der Data-Mining-Software RapidMiner vor, welches eine automatische, sequentielle Bearbeitung aller Lernaufgaben mit diversen Verfahren erlaubt.

6.1 Datenbereinigung

Unter dem abstrakten Begriff *Datenbereinigung* werden Maßnahmen zur Beseitigung von Fehlern in den Daten und Korrektur der Darstellung von Daten zusammengefasst. Hierunter fallen zum Beispiel auch Schritte der Datenveränderung zur Vermeidung von Verzerrungen bei der Datenanalyse, die etwa durch falsche Kodierung der Daten entstehen.

6.1.1 Angleichung von Wertelabeln

Im SOEP-Datensatz sind alle Attributwerte, d.h. auch nominale Werte, durch reelle Zahlen kodiert. Zusätzlich existieren für nominale Werte sogenannte Wertelabel, die die eigentliche Bedeutung der Werte als Zeichenkette repräsentieren. Das im Rahmen dieser Arbeit implementierte Programm PanelX zur Extraktion von Daten aus dem SOEP-Datensatz speichert die extrahierten Daten jedoch im ARFF-Format, wobei die (nominalen) Werte direkt durch ihre Wertelabel, d.h. die zugeordneten Zeichenketten repräsentiert sind. Im Zusammenhang mit mehreren Variablen eines Items ist dies unter Umständen problematisch: Sind Zeichenketten, die in unterschiedlichen Variablen eines Items gleiche Sachverhalte, d.h. eigentlich *gleiche* Werte erfassen, unterschiedlich und somit nicht konsistent, so fasst auch die Analysesoftware RapidMiner diese beim Einlesen der ARFF-Datei als unterschiedliche Werte auf. Dies ist natürlich auch der Fall, wenn solche Zeichenketten nur geringfügig differieren und für Menschen leicht als äquivalent zu identifizieren sind, etwa durch nur leichte Abweichungen der Schreibweise, Groß-/Klein-Schreibung, etc. Leider sind solche Inkonsistenzen im SOEP nicht selten vorhanden. Ein Beispiel hierfür sind die verwendeten Indikatoren, ob Kinder unter 16 Jahren im Haushalt leben, in den Variablen `AH06` bis `UH53`. Während in den Variablen `AH06` und `CH01` (neben den obligatorisch definierten Werten für fehlende bzw. leere Werte) als mögliche Werte "*ja*" und "*nein*" definiert sind, sind in allen anderen Variablen dieser Wertereihe die Attributwerte als "*Ja*" und "*Nein*" definiert. Da bei der Extraktion keine explizite Korrektur solcher Inkonsistenzen der Wertelabel erfolgt, muss diese spätestens vor den eigentlichen Auswertungsschritten im Datenanalyseprozess in RapidMiner durchgeführt werden. Eine automatische Korrektur (eine Art Pattern Matching) ist aufgrund der häufig sehr unterschiedlichen Schreibweisen und Abkürzungen bei der Vergabe von Wertelabeln seitens des SOEP im Allgemeinen nicht möglich bzw. verbietet sich wegen des hohen Aufwandes bei immer noch vorhandener Erfolgsunsicherheit. Stattdessen muss eine manuelle Berichtigung der Wertelabel erfolgen. In RapidMiner existiert hierzu der Operator `AttributeValueMapper`. Dieser erlaubt es, Werte von Attributen zu ändern. Werte aus unterschiedlichen Variablen, die den gleichen Sachverhalt widerspiegeln und daher äquivalent sind, können so auf einen gleichen Wert geändert werden. Allerdings erlaubt eine Operatorinstanz nur das Ändern jeweils eines Wertes für ein Attribut. Aus diesem Grund wurde der Operator `AttributeValueChanger` implementiert, der nicht nur

ein Wertemapping, sondern multiple Wertemappings direkt in einer Instanz des Operators erlaubt.

6.1.2 Behandlung fehlender und leerer Werte

Nicht so geradlinig zu lösen bzw. zu beseitigen wie das Problem inkonsistenter Wertelabel ist das Problem der Existenz fehlender bzw. leerer Werte. Sowohl fehlende Werte, bedingt beispielsweise durch die Verweigerung einer Antwort, als auch leere Werte, d.h. die Information, dass es wegen des Nicht-Zutreffens der anderen Antwortmöglichkeiten keinen Wert geben kann, stellen Datenanalyseverfahren vor Probleme. Fehlende Werte werden von vielen Verfahren entweder einfach ignoriert oder aber als eigener für das jeweilige Attribut möglicher Domänenwert behandelt. Generell wird zumindest durch Analysesoftware (wie auch RapidMiner) die Kennzeichnung von Werten als fehlend unterstützt. Insofern obliegt es dann der Implementierung der späteren Datenanalyseverfahren selbst, inwiefern sie fehlende Werte handhaben[1]. Bei leeren Werten ist dies nicht gegeben. Eine eigene spezielle Kennzeichnung existiert für leere Werte auch in RapidMiner nicht. Dies ist bei nominalen Werten insofern kein Problem, als dass leere Werte - wie bereits in Kapitel 3.3.5 angedeutet - generell als eigener, möglicher Domänenwert aufgefasst werden können. Somit ist eine spezielle Kennzeichnung leerer Werte nicht nötig, sie können wie alle anderen nominalen Werte behandelt werden. Im numerischen Fall ist dies insofern nicht möglich, da jeder mögliche numerische Wert von den Datenanalyseverfahren als gültiger Wert behandelt und somit in die Analyse (und den darin stattfindenden Berechnungen) einbezogen wird.

Zusammengefasst ergeben sich bei der Behandlung fehlender und leerer Werte vier Fälle nach den Dimensionen der Unterscheidung in nominale und numerische Werte sowie der Unterscheidung in fehlende und leere Werte. Generell sind im SOEP-Datensatz fehlende bzw. leere Werte durch die entsprechenden numerischen Wertekodierungen aus 3.4 repräsentiert. Wie bereits erwähnt ersetzt PanelX bei der Extraktion und Speicherung der extrahierten Daten bei nominalen Attributen diese Wertkodierungen jedoch durch die entsprechenden Wertelabel. Für nominale Werte sind die zu berücksichtigenden fehlenden bzw. leeren Werte als *keine Angabe*, *trifft nicht zu* und *nicht valide* gekennzeichnet. Für numerische Werte sind für diese Bedeutungen die Werte -1, -2 und -3 im extrahierten Datensatz zu finden.

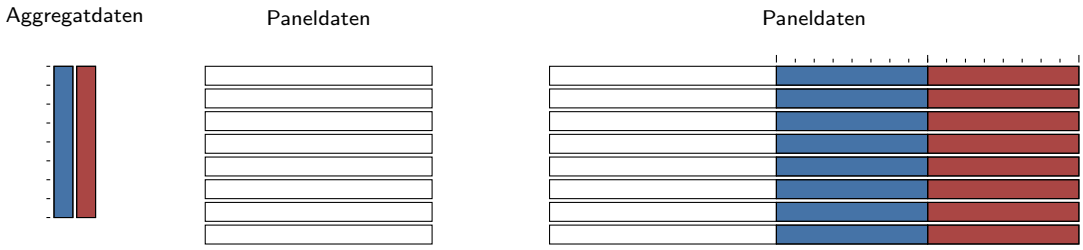
Für nominale Attribute ergeben sich folgende Implikationen hinsichtlich der Vorverarbeitung: fehlende Werte sollen als solche explizit gekennzeichnet werden, d.h. der Analysesoftware als solche bekannt gemacht werden. Dies geschieht ebenfalls durch Mapping der betreffenden Werte *keine Angabe* und *nicht valide* auf den von RapidMiner als fehlend definierten Wert mittels des bereits in RapidMiner vorhandenen Operators `AttributeValueMapper` oder des im Rahmen dieser Arbeit implementierten `AttributeValueChanger`. Feh-

[1] In Bezug auf das Phänomen und die Handhabung fehlender Werte existiert eine Fülle an Abhandlungen. Rubin (1976) schlägt die mittlerweile gebräuchlichste Form einer Taxonomie für verschiedene Mechanismen zur Entstehung und Verteilung fehlender Werte vor. Betrachtungen und Vergleiche verschiedener Verfahren zur Handhabung fehlender Werte erfolgen unter anderem durch Schafer und Graham (2002), Grzymala-Busse und Hu (2000) oder auch Quinlan (1989). Viele Ansätze dienen dabei der Ersetzung (Imputation) fehlender Werte. Hiermit befassen sich beispielsweise Lobo und Numao (1999) und Batista und Monard (2002). In Bezug auf sozio-ökonomische Daten sowie Längsschnittdaten finden sich Untersuchungen hinsichtlich fehlender Werte in Schnell (1986) und Spiess (2004). In Bezug auf diese Arbeit wurde aufgrund der bereits hohen Komplexität der Vorgehensweise die einfachste Form der Behandlung fehlender Werte, das simple ignorieren dieser bzw. die Verlagerung der diesbezüglichen Verantwortung auf die eigentlichen Datenanalyseverfahren, verwendet.

lende Werte werden dabei in RapidMiner durch ein Fragezeichen (*?*) repräsentiert. Leere Werte, gekennzeichnet durch den Wert *trifft nicht zu* können als eigener Domänenwert der jeweiligen nominalen Attribute bestehen bleiben und bedürfen somit keiner weiteren Behandlung.

Numerische Attribute sind hinsichtlich fehlender Werte ähnlich zu behandeln. Somit müssen numerische Werte, die gleich den im SOEP definierten Kodierungen -1 bzw. -3 sind, durch das Fragezeichen (*?*) ersetzt und damit als fehlend gekennzeichnet werden. Da der für das Panel-Plugin implementierte `AttributeValueChanger` nur mit nominalen Werten umgehen kann, muss hier in RapidMiner auf den ursprünglich vorhandenen Operator `AttributeValueMapper` zurückgegriffen werden.

Ungleich schwieriger ist die Berücksichtigung leerer Werte bei numerischen Attributen. Im Idealfall müsste auch diesbezüglich ein eigener möglicher Domänenwert zur Verfügung stehen. Da im Regelfall in RapidMiner jedoch jeder numerische Wert als gültiger Wert aufgefasst und in die Berechnung mit einbezogen wird, verbietet sich hier etwa die - im SOEP erfolgte - Definition eines negativen Wertes als leeren Wert. Obwohl etwa semantisch für eine Einkommensvariable kein negativer Wert als gültiger Wert in Frage kommen könnte, würden die in RapidMiner enthaltenen Datenanalyseverfahren diesen negativen Wert mit in die Berechnung (von Mittelwerten, etc.) einbeziehen. Dies würde die Ergebnisse eklatant verzerren. Dennoch gibt es drei Möglichkeiten, zumindest vordergründige Verzerrungen dieser Art zu vermeiden. Die erste bezieht sich auf Variablen, bei denen der leere Wert (d.h. der Wert *trifft nicht zu*) semantisch mit einem gültigen Domänenwert übereinstimmt. Dies ist etwa bei Variablen wie dem monatlichen Einkommen der Fall. Hier bedeutet *trifft nicht zu*, dass das monatliche Einkommen einer Person gleich null ist. Somit kann bei Variablen, für die eine solche Beziehung gilt, jedes Auftreten des Wertes *trifft nicht zu* durch den entsprechenden gültigen Wert (im Regelfall null) ersetzt werden. Bei vielen Variablen gilt eine solche Beziehung jedoch nicht, vor allem in Bezug auf Daten von Ereignissen. Ein aussagekräftiges Beispiel ist etwa die SOEP-Variable `IMMIYEAR`, die das Jahr der Immigration nach Deutschland erfasst, *falls* eine Person nach Deutschland immigrierte. Der Fall, dass eine Person schon immer in Deutschland lebte, wird auch hier durch den Wert -2 für *trifft nicht zu* gekennzeichnet. Hier kann dieser Wert jedoch nicht ohne Verzerrung der Analyseergebnisse durch einen gültigen Domänenwert ersetzt werden. Somit muss auf eine der beiden anderen Alternativen zurückgegriffen werden, die leider ebenfalls nachteilbehaftet sind: So könnte der Wert *trifft nicht zu* wie im Fall fehlender Werte ebenso als ein fehlender Wert markiert werden, allerdings resultierte auch hieraus eine Verzerrung. Im Beispiel der Immigration würde damit die Information, ob jemand nach Deutschland immigriert wurde oder nicht, vollständig ignoriert. Vielmehr würde durch die Kennzeichnung der Werte als fehlend eigentlich impliziert, dass diese Werte eigentlich (mit einer gültigen Ausprägung) existieren, sie jedoch nur nicht beobachtet wurden. Im hier besten Fall, dass Datenanalyseverfahren fehlende Werte einfach ignorieren, würden die Verfahren ihre Berechnung - zumindest in Bezug auf das betroffene Attribut - nur auf die tatsächlich vorhanden Werte (im Beispiel die der Immigranten) stützen. Die dritte alternative Möglichkeit zur Handhabung leerer Werte besteht darin, das numerische Attribut zu diskretisieren, d.h. in ein nominales Attribut zu überführen (zu alternativen Vorgehensweisen hierzu siehe Kapitel 6.7). Bei einem solchen nominalen Attribut dürfte dann die Ausprägung *trifft nicht zu* als eigenständiger Wert existieren. Allerdings ergibt sich zu Gunsten der Möglichkeit, leerer Werte als eigenständigen Attributwert einbeziehen zu können, ebenfalls der Nachteil einer ignorierten und vernachlässigten Information. So wird die Information über die Ordnung



(a) vorher (b) nachher

Abbildung 6.1: Verknüpfung aggregierter Daten mit Paneldaten

der einzelnen Attributwerte, d.h. die Ordnungsrelation, die bei einer numerischen Skala vorhanden ist, nicht mit auf die nominale Skala und damit ein nominales Attribut übertragen. Auch dieses hat potentielle Auswirkungen auf die Ergebnisse der später anzuwendenden Datenanalyseverfahren.

6.2 Hinzufügen makroökonomischer Daten

Für die in Abbildung 4.2 definierten Zustandswechsel werden neben den mikroökonomischen Attributen, die die Heterogenität der Personen durch Erfassung ihrer Eigenschaften abbilden, auch makroökonomische Größen als potentielle Einflussfaktoren der Arbeitslosigkeit einbezogen. Hierbei handelt es sich um die in Kapitel 4.3.1 identifizierten Größen des Wachstums des Bruttoinlandsprodukts sowie der Verbraucherpreise auf Jahresbasis mit zugehörigen, von der OECD veröffentlichten Daten, sowie der in Tabelle 4.2 aufgelisteten Daten. Diese wurden manuell in einer für RapidMiner lesbaren Datei gespeichert. In dieser Datei wurden die oben genannten Variablen als Spalten, die Jahre, also Zeitpunkte, als Zeilen erfasst. Im Einzelnen enthält die Datei nach Erzeugung basierend auf den OECD-Daten die Variablen `WACHSTUM` und `INFLATION` sowie `WACHSTUM_L1`, `WACHSTUM_L2` und `WACHSTUM_L3`, die das Wachstum um jeweils ein bis drei Jahre verzögert darstellen. Daneben enthält die Datei weiterhin gemäß der Definition aus Tabelle 4.2 die Variablen `ZYKLUS`, `LEGISLATUR` und `WIEDERVEREINIGUNG`.

Nachdem diese Daten in RapidMiner eingelesen wurden, müssen diese gesamtwirtschaftlichen Daten, die für jede Untersuchungseinheit des SOEP gelten, mit den ebenfalls bereits in RapidMiner vorliegenden SOEP-Daten verknüpft werden. Auch für diese - recht spezielle - Verknüpfung wurde ein neuer Operator `AppendAggregatedSeries` implementiert, der in obiger Form vorliegende Daten von Aggregaten mit Paneldaten verknüpft[2]. Dies geschieht derart, dass der Paneldatentabelle für jede Variable der Aggregatdaten und jeden in diesen Daten erfassten Zeitpunkt eine neue Variable hinzugefügt wird. Anschließend werden für alle Untersuchungseinheiten des Panels die identischen aggregierten Daten in den jeweiligen Variablen hinzugefügt. Die Transformation wird durch die schematische Darstellung der Ein- und Ausgabeformen einer exemplarischen Anwendung in Abbildung 6.1 illustriert. Dabei werden zwei Wertereihen (farbig dargestellt), die für alle acht Untersuchungseinheiten

[2]Das Vorliegen von Paneldaten ist zwar bei der hier durchgeführten Anwendung gegeben, ist allerdings nicht zwangsläufig verlangt: generell können mit diesem Operator Daten, die für alle Beispiele eines Datensatzes gelten, mit diesem Datensatz verknüpft werden.

des Beispiel-Panels gelten sollen, mit den eigentlichen Paneldaten (weiß dargestellt) derart verknüpft, dass jede Untersuchungseinheit die beiden Wertereihen als Merkmale besitzt.

6.3 Erzeugen von Übergangstriggern

Für die durchzuführende Analyseaufgaben müssen - wie in Kapitel 4.2 motiviert - Attribute erzeugt werden, die Arbeitsmarktzustandsübergänge von Personen erfassen. Dies geschieht auf Basis der Variablen `LFSxx`, die den jeweils aktuellen Arbeitsmarktzustand der Personen zu den Befragungszeitpunkten der einzelnen Wellen angeben. Vor dieser Merkmalsgenerierung wird dazu bei diesen Attributen eine Zuordnung von den betrachteten Arbeitsmarktzuständen zu den Originalwerten gemäß Tabelle 4.1 durchgeführt. Dieses Mapping geschieht mit Hilfe des Operators `AttributeValueChanger`. Aus den Variablen `LFSxx` mit den entsprechend zugeordneten Werten, die somit statische Zustände zu verschiedenen Zeitpunkten erfassen, müssen dann Zustandsübergänge generiert werden. Um diese Aufgabe zu lösen, wurde im Rahmen des Panel-Plugins der Operator `SeriesTrigger` implementiert. Mit diesem können in Paneldaten flexibel solche Zustandswechsel von Untersuchungseinheiten in nominalen Wertereihen erkannt werden. Dies geschieht bei der Vorverarbeitung nach folgender Regel: liegen etwa Zustandsbeobachtungen in den Attributen S_1 bis S_T vor, so werden Attribute $C_1^{ss'}$ bis $C_{T-1}^{ss'}$ defniert, die für $t = 1, \ldots, T-1$ einen Zustandswechsel vom Zustand s in den Zustand s' als

$$C_t^{ss'} := \begin{cases} \textit{positive}, & \text{wenn } S_t = s \ \wedge \ S_{t+1} = s' \\ \textit{negative}, & \text{wenn } S_t = s \ \wedge \ S_{t+1} \neq s' \ \wedge \ S_{t+1} \neq ? \\ ?, & \text{sonst} \end{cases}$$

erfassen. Sollen also etwa aus den 21 Attribute `LFS84` bis `LFS04` Zustandswechsel von *Working* nach *Unemployed* erkannt werden, so resultieren daraus 20 neue, binominale Attribute. Diese Attribute haben genau dann für Untersuchungseinheiten den Wert *positive*, falls ein Wechsel von *Working* nach *Unemployed* in den zugehörigen Variablen der Zustandsreihe bei diesen Untersuchungseinheiten stattgefunden hat. Sie haben den Wert *negative*, falls der erste Zustand zwar *Working*, der darauffolgende aber nicht *Unemployed* war. Hierunter entfallen somit Wechsel von *Working* in andere Zustände sowie ein Verbleiben im Zustand *Working*. In allen anderen Fällen, d.h. dass mindestens einer der beiden involvierten Zustände fehlend war, oder dass der erste Zustand ein anderer Zustand als *Working* war, wird der Zustandswechsel als fehlend gekennzeichnet. Anhand dieser Markierung können diese Beispiele für die spätere Analyse einzelner Zustandsübergänge ausgeschlossen werden. Mit Hilfe der geschilderten Methode wurde durch Anwendung des Operators `SeriesTrigger` für jeden Zustandsübergang aus Abbildung 4.2 eine Wertereihe von Übergangsindikatoren erzeugt. Den Daten wurden also die Variablen `LFSC_Working_Not-working`, `LFSC_Working_Unemployed`, usw. hinzugefügt.

6.4 Angleichung von Wertereihenlängen

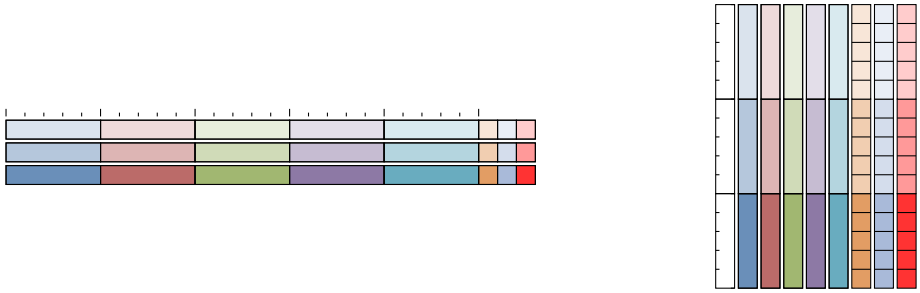
Vor weiteren Transformationen soll sichergestellt werden, dass alle Wertereihen, die Merkmale der Personen über die Zeit beobachten, die gleiche Länge haben. Prinzipiell umfassen alle Wertereihen, die aus dem SOEP extrahiert wurden, alle Wellen des SOEP. Auf Jahres-

basis ergibt sich damit eine Länge der einzelnen Wertereihen von 21 Variablen. Auch die hinzugefügten makroökonomischen Wertereihen haben diese Länge. Anders verhält es sich jedoch bei den nach Kapitel 6.3 erzeugten Variablen, die Übergänge zwischen den Zuständen erfassen: aus den 21 Zuständen können lediglich 20 mögliche Zustandsübergänge abgeleitet werden. Die Wertereihen, die die Zustandsübergänge erfassen, haben somit nur die Länge 20. Zur späteren Transformation sollen alle Wertereihen jedoch gleiche Länge haben. Daher müssen die übrigen Wertereihen potentiell erklärender Faktoren gekürzt werden. Bei diesen sind generell eher die Werte interessant bzw. relevant die *vor* den Zustandswechseln erfasst wurden. Daher wird die jeweils letzte Variable der erklärenden Wertereihen, die für keinen nachfolgenden Zustandsübergang als erklärende Variable herangezogen werden können, ignoriert und somit aus den Daten gelöscht. Prinzipiell wäre dies in RapidMiner mit dem Operator `FeatureNameFilter` möglich, jedoch müsste bei diesem der Name der zu löschenden Variable explizit angegeben werden. Etwas flexibler gestaltet sich die Lösung der Aufgabe durch Anwendung des im Rahmen des Panel-Plugins entwickelten Operators `SeriesFeatureFilter`. Dieser erlaubt, alternativ die ersten oder letzten n Variablen, oder auch alle Variablen einer Wertereihe, die durch einen regulären Ausdruck spezifiziert wird, aus den Daten zu löschen. Auch die Zahl n kann dabei als Parameter angegeben werden. Im Rahmen der hier durchzuführenden Vorverarbeitung können somit leicht die jeweils letzten Attribute einer Wertereihe der vorhandenen Wertereihen entfernt werden.

6.5 Wechsel der Analyseeinheit

Der zentrale Schritt der Vorverarbeitung im Rahmen dieser Arbeit ist die Transformation der Repräsentation der Paneldaten. So werden in diesem Schritt die in Ein-Tabellen-Form (siehe Abbildung 2.6(b)) vorliegenden Paneldaten in das Long-Format (siehe Abbildung 2.7) überführt. Dies geschieht durch den Operator `MultivariateAttributes2Examples`. Dieser transformiert die Paneldaten derart, dass multivariate Wertereihen, die durch mehrere Variablen (eine pro Zeitpunkt) erfasst werden, nach der Transformation im Datensatz nur noch in einer einzigen Variable (und damit Spalte) erfasst werden. Die Variablen, die Wertereihen bilden, müssen dem Operator dabei explizit mittels eines Parameters bekannt gemacht werden. Die Transformation hat zur Folge, dass Daten für eine Untersuchungseinheit nicht mehr nur in einer Zeile, also einem Beispiel, sondern in mehreren vorkommen. Damit werden in einer Zeile der Datentabelle nun Daten für eine Untersuchungseinheit zu genau einem Zeitpunkt erfasst. Dies bedeutet gleichzeitig, dass Daten, die für eine Untersuchungseinheit unveränderlich sind (wie etwa das Geschlecht, die Herkunft) in alle Zeilen, die Daten dieser Untersuchungseinheit - zu verschiedenen Zeitpunkten erfassen - geschrieben werden müssen und somit eigentlich eine gewisse Redundanz innerhalb der Daten eingeführt wird. Weiterhin wird bei der Transformation ein neues Attribut eingefügt, welches die Zeitpunkte, zu denen die Daten in den jeweiligen Zeilen gehören, angeben. Die beschriebene Transformation kann an einem Beispiel in Abbildung 6.2 schematisch nachvollzogen werden. Das beispielhaft dargestellte Panel besteht aus drei Untersuchungseinheiten, für die jeweils fünf Wertereihen und drei statische, d.h. unveränderliche Merkmale erhoben wurden. Befanden sich vor der Transformation die Wertereihen für die einzelnen Untersuchungseinheiten in Zeilen, so befinden sie sich nach der Transformation in Spalten.

Die soeben beschriebene Transformation hat bedeutende Auswirkungen auf die Analyse bzw. hinsichtlich der Anwendbarkeit von Analyseverfahren. Dies resultiert aus der impli-



(a) Ein-Tabellen-Form (b) Long-Format

Abbildung 6.2: Transformation von Paneldaten in Ein-Tabellen-Form ins Long-Format

ziten Änderung der Analyseeinheit bei der Transformation. Im Fall der hier bearbeiteten Daten haben sich durch die Transformation die betrachteten Analyseeinheiten von Personen auf Personen-Jahre geändert, d.h. es wird nicht mehr eine Person als Einheit betrachtet, sondern eine Person in einem bestimmten Jahr. Die immensen Vorteile dieser Transformation hinsichtlich der Analyse sind aus dem Vergleich der beiden Datenformen vor und nach der Transformation leicht ersichtlich. Vor der Transformation lagen sowohl die Zielvariablen (bei den hier betrachteten Daten die Zustandsübergangsindikatoren) als auch die ausgewählten, potentiell zur Erklärung dienenden Attribute für jede Analyseeinheit als Wertereihen vor. Somit verbot sich die Anwendung von gewöhnlichen Datenanalyseverfahren, die üblicherweise nur mit Querschnittsdaten umgehen können, da sie nicht explizit für zeitlich veränderliche Daten entworfen wurden. Nach der Transformation liegen jedoch genau solche Querschnittsdaten, d.h. ein Querschnitt von Personen-Jahren vor. Innerhalb der Merkmale einer Analyseeinheit, d.h. innerhalb eines Personen-Jahres, existiert damit keine zeitliche Variabilität eines Merkmals oder der Zielvariable mehr. Dies ermöglicht die Anwendung vieler üblicher, nicht explizit für Paneldaten bzw. sonstige zeitveränderliche Daten entwickelter Methoden wie etwa üblicher Klassifikationsverfahren (siehe hierzu auch Kapitel 9). Aus diesem Grund ist die oben beschriebene Transformation auch Grundlage bei ökonometrischen Analysen von Paneldaten üblicherweise angewandter Panelregressionsmodelle wie dem gepoolten Panelmodell, dem Fixed Effects Model oder dem Random Effects Model. Zeitliche Variabilität von Merkmalen einzelner Personen wird bei diesen Modellen allerdings dennoch teilweise durch gesonderte Annahmen der Modelle berücksichtigt. Interessierte Leser finden nähere Informationen dazu in umfangreichen Standardwerken zu diesem Thema wie denen von Greene (2003), Hsiao (2003) und Baltagi (2003).

Zusammenfassend liefert die mittels des Operators `MultivariateAttributes2Examples` durchgeführte Vorverarbeitung Daten, die im Long-Format vorliegen und einen Querschnitt von Beobachtungen von Personen pro Jahr bilden. Dabei wurden alle Wertereihen, die der Datensatz enthielt gemäß obiger Beschreibung transformiert sowie statische Attribute beibehalten. Als zusätzliche Variable wurde zudem ein Index, der die jeweiligen Wellen der einzelnen Beobachtungen indiziert, hinzugefügt. Da die transformierten Wertereihen Daten aus 20 Wellen und damit - vor der Transformation - 20 Variablen umfassten, hat diese Indexvariable einen Wertebereich von 1 bis 20. Zu erwähnen ist ebenfalls, dass sich die Anzahl der Beispiele des Datensatzes um den Faktor 20 erhöht hat. Gleichzeit ist jedoch die Zahl der Attribute sehr viel kleiner geworden. Ein Faktor 20 kann diesbezüglich nicht ganz erreicht werden, da statische Attribute erhalten blieben.

6.6 Merkmalsgenerierung

Manche Merkmale, deren Auswirkungen auf die Zielvariable in der Analyse überprüft werden soll, sind im SOEP und daher in den aus dem SOEP extrahierten Daten nicht explizit vorhanden, können jedoch aus vorhandenen Merkmalen konstruiert werden. Dies sind im Wesentlichen Merkmale, die zeitliche Effekte erfassen und häufig Resultat von Ereignissen sind. Hierunter fallen vor allem drei Bereiche, die im folgenden betrachtet werden.

6.6.1 Jahr und Alter

Nach der Transformation aus Kapitel 6.5 liegen die Beobachtungen so vor, dass sich einzelne Beispiele auf einzelne Wellen und somit Jahre beziehen. Zur Identifikation der jeweiligen Welle, für die ein Beispiel gilt, wurde bei der erwähnten Transformation eine Index-Variable hinzugefügt. Falls zeitliche Effekte zusätzlich zu den aus der Heterogenität der Untersuchungseinheiten resultierenden Effekten mit in die spätere Analyse einbezogen werden sollen, könnte die Index-Variable diesbezüglich bereits als potentiell erklärende Variable dienen und einbezogen werden. Alternativ wurde im Rahmen der hier beschriebenen Vorverarbeitung der Daten jedoch eine Variable `YEAR` explizit erzeugt, die das Jahr der Beobachtungen angibt. Da die Index-Variable die vom SOEP erfassten Wellen durch die Zahlen 1 bis 20 indiziert und die erste Befragungswelle des SOEP im Jahr 1984 durchgeführt wurde, kann diese Variable `YEAR` einfach durch Addition der Werte der Index-Variablen mit einem Offset 1983 berechnet werden. Hierzu kann der RapidMiner-Operator `FeatureGeneration` verwendet werden, der die Erzeugung neuer Attribute, die aus einer Berechnung anhand bestehender Attribute hervorgehen, erlaubt.

Auch das Alter der Personen (zum jeweiligen Betrachtungszeitpunkt) ist bislang nicht explizit in den Daten vorhanden. Ebenso wie das Jahr kann es jedoch leicht aus den vorhandenen Daten berechnet werden. Dies kann auf Basis der soeben erzeugten Variable `YEAR` in Kombination mit der aus dem SOEP extrahierten Variable `GEBJAHR`, die das Geburtsjahr der Personen angibt, geschehen. Im Rahmen der Vorverarbeitung wird daher ebenfalls durch Anwendung des Operators `FeatureGeneration` eine neue Variable `AGE` erzeugt, wobei hierzu einfach der Wert der Variablen `YEAR` von dem der Variable `GEBJAHR` subtrahiert wird.

6.6.2 Anzahl der Kinder

Wie aus Tabelle 5.2 hervorgeht, enthalten die aus dem SOEP extrahierten Daten auch die Variablen `KIDGEB01` bis `KIDGEB15`, die das Geburtsjahr der leiblichen Kinder einer Person angeben. Dieser zunächst nicht zielgerichtet erscheinende Einbezug dieser Variablen macht insofern Sinn, als dass nun aus diesen Variablen die Anzahl der leiblichen Kinder, die eine Person zu einem Zeitpunkt, also in einem Jahr hat, berechnet werden kann. Dabei wird abermals auf die erzeugte Variable `YEAR` zurückgegriffen. Konzeptionell ergibt sich die Anzahl der leiblichen Kinder einer Person in einem Jahr als Anzahl der Variablen aus den Variablen `KIDGEB01` bis `KIDGEB15`, für die zutrifft, dass das jeweilig erfasste Geburtsjahr des Kindes zahlenmäßig kleiner als das in der Variable `YEAR` erfasste ist. Prinzipiell könnte ein (zu implementierender) Operator diese Bedingung einfach prüfen. Allerdings wäre dieser hinsichtlich seines Einsatzbereiches oder einer Wiederverwendung arg beschränkt. Stattdessen wurde ein anderer Weg unter Einsatz generischerer Operatoren gewählt. Hierzu

wurden zunächst (erneut durch Einsatz des Operators `FeatureGeneration`) neue Variablen `KIDGEB01_REL` bis `KIDGEB15_REL` erzeugt, die die Differenzen aus dem aktuellen Jahr aus der Variable `YEAR` und den Geburtsjahren der Kinder aus den Variablen `KIDGEB01` bis `KIDGEB15` enthalten. Sind die Werte dieser Variablen negativ, so wurde das betreffende Kind erst nach dem aktuellen Jahr geboren. Sind die Werte positiv, so wurde das Kind vor dem aktuellen Jahr geboren. Daher wurde anschließend der Operator `UserBasedDiscretization` angewandt, mit dem die Attribute `KIDGEB_REL01` bis `KIDGEB_REL15` diskretisiert und damit in nominale Attribute überführt werden können. Der Operator erlaubt die Angabe von Intervallen der numerischen Attribute und nominaler Werte, die diesen Intervallen zugeordnet werden sollen. Somit erfolgte die Diskretisierung derart, dass negative numerische Werte auf die nominale Ausprägung *negativ* und positive auf die Ausprägung *positiv* abgebildet wurden. Zum Einsatz kam anschließend der im Rahmen dieser Arbeit entwickelte Operator `SeriesCounter`, mit dem eigentlich in nominalen multivariaten Wertereihen, d.h. Längsschnittdaten in Wide-Format, Vorkommnisse von Werten gezählt werden können. Somit wurden die Anzahlen der Vorkommnisse des Wertes *positiv* in den soeben diskretisierten Variablen bestimmt. Diese entspricht der Anzahl der Kinder, die eine Person in dem jeweiligen Jahr hat. Die Resultate wurden vom Operator `SeriesCounter` in der Variable *SUMKIDS*[3] gespeichert.

6.6.3 Immigration

Neben dem Geburtsjahr einer Person und den Geburtsjahren ihrer Kinder lässt sich auch das Jahr der Immigration bei Immigranten als Basis zur Merkmalsgenerierung nutzen. So kann durch Subtraktion des aktuellen Jahres und des Immigrationsjahres etwa berechnet werden, vor wie vielen Jahren (bezogen auf das jeweilige Jahr) eine Person nach Deutschland immigriert ist. Auch dies wurde durchgeführt, unter Berücksichtigung der Problematik der Existenz leerer Werte in Bezug auf Personen, die nicht immigriert sind, wurde die Merkmalsgenerierung jedoch noch um eine Diskretisierung ergänzt (siehe auch Kapitel 6.7).

6.7 Diskretisierung numerischer Attribute

Die Vorverarbeitung der Daten im Rahmen dieser Arbeit enthält als nächsten Schritt die Diskretisierung numerischer Attribute. Die Notwendigkeit einer Diskretisierung hängt dabei davon ab, ob die später anzuwendenden Analyseverfahren in der Lage sind, numerische Daten verarbeiten zu können. Die verwendeten Klassifikationsverfahren (siehe Kapitel 9) machen diesbezüglich keine Einschränkungen. Die Verfahren zur Subgruppenentdeckung (siehe Kapitel 10) schränken dagegen die Anwendbarkeit dahingehend ein, dass nur nominale Attribute betrachtet werden können. Somit ist zumindest für diese eine vorherige Diskretisierung numerischer Attribute Voraussetzung.

Zur Diskretisierung numerischer Attribute existieren in RapidMiner mit mehreren Operatoren unterschiedliche Möglichkeiten hinsichtlich der Wahl der Intervalle numerischer Werte, die auf nominale Werte abgebildet werden sollen. Schon angesprochen wurde der Operator `UserBasedDiscretization`, dessen initiale Version ebenfalls im Rahmen dieser

[3] Eine Variable SUMKIDS existiert auch im SOEP in der Datei BIOBIRTH bzw. BIOBRTHM. Allerdings werden von dieser die Anzahlen der leiblichen Kinder der Personen bis zu dem Jahr erfasst, welches als letztes im SOEP-Datensatz erfasst ist. Bei dem vorliegenden Datensatz wären somit Kinder bis zum Jahr 2004 erfasst.

Arbeit implementiert wurde, der allerdings nicht Teil des Panel-Plugins sondern mittlerweile direkt in RapidMiner verfügbarer Operator ist. Dieser Operator erlaubt die Definition der Intervalle, die auf nominale Werte abgebildet werden sollen, d.h. er erlaubt die Spezifikation der Grenzen dieser Intervalle sowie der nominalen Werte. Insofern obliegt es dem Benutzer, diese Grenzen aufgrund seines Domänenwissens auszuwählen. Da in dieser Arbeit prinzipiell eine recht offene Anwendung von Datenanalyseverfahren erfolgen soll, wurde dieser Ansatz nur für die im letzten Abschnitt angesprochene Variable, die die Zeitdauer seit Immigration in Bezug auf das aktuelle Jahr erfasst, gewählt. Das resultierende Attribut `IMMIGRATED` erhielt damit die Ausprägungen *im letzten Jahr*, *vor 1 bis 2 Jahren*, *vor 2 bis 5 Jahren*, *vor 5 bis 10 Jahren*, *vor 10 bis 20 Jahren*, *vor 20 bis 30 Jahren*, *vor mehr als 30 Jahren* und natürlich die Ausprägung *nicht immigriert* im Fall, dass die Person kein Immigrant ist.

Die meisten anderen numerischen Attribute sind, wie `IMMIGRATED`, ebenfalls Attribute, die nur ganzzahlige Werte enthalten. Die vorher erzeugten Attribute `YEAR`, `AGE` oder `SUMKIDS` sind Beispiele dafür. Anstatt eine benutzer-basierte Diskretisierung mittels des Operators `UserBasedDiscretization` durchzuführen, sollte bei der Diskretisierung dieser Variablen möglichst wenig in die vorliegende Verteilung der Werte eingegriffen werden und diese möglichst genau auf die resultierenden nominalen Attribute übertragen werden. Da ganzzahlige numerische Attribute eigentlich schon diskret sind, ist es naheliegend, sie einfach in nominale Attribute umzuwandeln, welche als Werte genau die ganzzahligen numerischen Werte haben. Dieses kann mittels des Operators `IntegerDiscretization` durchgeführt werden, der für das Panel-Plugin entwickelt wurde. Dieser diskretisiert numerische Werte, indem er sie auf nominale Werte abbildet, die die zugehörigen ganzzahligen numerischen Werte darstellen. Dies bedeutet, dass ein numerisches Attribut, welches nur ganzzahlige Werte enthält, bijektiv auf nominale Werte abgebildet wird. Enthält ein Attribut hingegen auch numerische Werte, die nicht ganzzahlig sind, gilt keine bijektive Beziehung. Stattdessen werden die nicht ganzzahligen Attributwerte entweder gerundet oder nur der ganzzahlige Anteil der Werte genommen und dann ebenfalls auf die nominalen Werte abgebildet. Alle aus dem SOEP extrahierten numerischen Daten wurden bis auf unten erläuterte Ausnahmen für nachfolgende Analysen, die nominale Werte verlangten, mittels des Operators `IntegerDiscretization` diskretisiert.

Alternativ zu den beiden oben beschriebenen Verfahren der Diskretisierung besteht auch die Möglichkeit einer datengestützten Diskretisierung, bei der die Intervallgrenzen für die zu bildenden nominalen Klassen automatisch auf Basis der Verteilung der numerischen Attributwerte gewählt werden. Die Anzahl der nominalen Klassen muss dabei vom Benutzer spezifiziert werden. RapidMiner bietet diesbezüglich unter anderem die Operatoren `BinDiscretization` sowie `FrequencyDiscretization`. Bei der ersten Form werden die Grenzen für ein Attribut derart gewählt, dass der Bereich zwischen Minimum und Maximum der in den gegebenen Beispielen für das Attribut auftretenden Werten äquidistant in die Anzahl der vorgegebenen Klassen unterteilt wird. Bei der zweiten Form der Diskretisierung wird ebenfalls die vorgegebene Anzahl an Intervallen gebildet, allerdings werden die Grenzen derart gewählt, dass die Anzahlen der Beispiele, die in jede Klasse fallen, identisch ist. Durch eine derartige Wahl der Grenzen wird damit eine Gleichverteilung der durch die Diskretisierung entstehenden nominalen Werte in den Daten erzeugt.

In einem ersten Schritt innerhalb dieser Arbeit wurden als Ausnahme zu obiger Vorgehensweise bei der Diskretisierung nur die Attribute `WACHSTUM`, `WACHSTUM_L1`, `WACHSTUM_L2`, `WACHSTUM_L3` sowie `INFLATION` mittels des Operators `BinDiscretization` in fünf nomina-

le Werteklassen diskretisiert und diese auf die Ausprägungen *sehr niedrig*, *niedrig*, *mittel*, *hoch* und *sehr hoch* abgebildet.

6.8 Filtern relevanter Daten

Als letzter Schritt muss im Rahmen der Vorverarbeitung das Filtern für die Analyse relevanter Daten erfolgen. Dies beinhaltet zum einen das Löschen nicht in die Analyse einzubeziehender Attribute. Hierunter ist noch nicht etwa eine datengestützte Selektion von Attributen zu verstehen, die sich als bedeutend hinsichtlich ihres Erklärungsgehaltes in Bezug auf die betrachteten Zustandsübergänge erweisen. Vielmehr sollen Daten ignoriert - d.h. beseitigt - werden, die von vornherein irrelevant für die eigentliche Analyse sind. Hierunter fallen etwa die zur Verknüpfung bei der Extraktion der Paneldaten benötigten und hinzugefügten Identifikationsschlüssel auf Personen- und Haushaltsebene. Somit wurden die Attribute `PERSNR` sowie `xHHNR` mittels der Operatoren `FeatureNameFilter` bzw. `SeriesFeatureFilter` aus den Daten gelöscht. Auch die bei der Transformation der Analyseeinheit hinzugefügte Index-Variable, die die Wellen indiziert, kann wieder entfernt werden, da die Zeitpunkte, denen die vorliegenden Beobachtungen zuzuordnen sind, durch die generierte Variable `YEAR` erfasst werden.

Zum anderen können und müssen aus den Daten auch Beispiele entfernt werden, die für die Analyse nicht relevant sind. Es können etwa alle Beispiele gelöscht werden, die nicht Daten von tatsächlichen Befragungspersonen enthalten, sondern etwa von Kindern oder von Personen, die in der jeweiligen Welle die Befragung verweigerten. Dies wurde durch Einsatz des RapidMiner-Operators `ExampleFilter` erreicht, die die Bedingung überprüft, ob das Attribut `NETTO` den Wert *Befragungsperson (_P)* hat. Nur Beispiele, für die dies zutrifft wurden in der Menge der zu analysierenden Beispiele belassen. Durch eine weitere Instanz des `ExampleFilter`-Operators wurden zudem möglicherweise auftretende Beispiele mit einem Gewicht von null (dies entspricht einem Wert des Attributs `PHRF` von null) gelöscht. Hiernach kann analog zur Personen- und Haushaltsnummer außerdem die Variable `NETTO` aus den Daten entfernt werden. Die Gewichte in der Variable `PHRF` sind jedoch für die spätere Analyse von Bedeutung und müssen deshalb in der Datentabelle verbleiben.

6.9 Resultierende Daten

Zur Entdeckung von Einflussfaktoren der definierten Zustandsübergänge wurden nach der Extraktion der in Tabelle 5.2 aufgelisteten Variablen Vorverarbeitungsschritte gemäß der obigen Beschreibungen durchgeführt. Zur Verdeutlichung der Komplexität des Vorverarbeitungsprozesses ist dieser in Anhang B in der Box-Plot-Darstellung von RapidMiner abgebildet. Der nach der Vorverarbeitung resultierende Datensatz enthält die in Tabelle 6.1 genannten Attribute und späteren Zielvariablen.

6.10 Framework für Experimente

Aufbauend auf den aus der Vorverarbeitung resultierenden Daten, geschieht die Analyse in einem dafür bereitgestellten Framework innerhalb eines RapidMiner-Prozesses. Dieser Prozess iteriert über die vorher erzeugten Zielattribute, d.h. die einzelnen Zustandswechsel. Für jeden Zustandswechsel wird der vorverarbeitete Datensatz eingelesen, das für die

Eingabe : Menge von Beispielen E, ausgewählte Attributmenge A
Ausgabe: Analyseergebnisse R

```
1 R = ∅
2 for Y ∈ {LFSC_Working_Not-working, LFSC_Working_Unemployed, ...} do
3     E ← LeseDaten
4     SetzeLabel(E,Y)
5     SetzeGewicht(E,PHRF)
6     LöscheFehlendeLabel(E,Y)
7     SelektiereAttribute(E,Y,A)
8     R ← R ∪ Analysiere(E)
9 end
```

Abbildung 6.3: Experiment-Framework für Analysen

Iteration aktuelle Zustandswechselattribut als Label ausgewählt und die anderen Zustandswechselattribute aus dem Datensatz entfernt. Des Weiteren wird das Attribut `PHRF` als Attribut markiert, das Beispielgewichte enthält. Zudem werden alle Beispiele aus den Daten (für diese Iteration) entfernt, in denen das aktuelle Label als fehlend markiert ist. Diese Beispiele sind für die durch das aktuell vorliegende Label definierte Lernaufgabe irrelevant. Ein letzter Schritt vor der Analyse dient zur (Vor-)Auswahl und Selektion relevanter Attribute. Prinzipiell ist diese schon durch die Extraktion explizit selektierter Variablen erfolgt. Nichtsdestotrotz können und sollten bei den Übergängen aus Zuständen, die weder erwerbstätig noch nebenerwerbstätig sind, die Variablen, die die Daten zur Erwerbstätigkeit der Personen erfassen, gelöscht werden, da sie für die Analyse dieser Übergänge stets unzutreffend sind. Schematisch ist der Ablauf zusammenfassend der Abbildung 6.3 zu entnehmen.

Tabelle 6.1: Resultierende Attribute und spätere Zielattribute

Variablenname	Beschreibung	Skala
SEX	Geschlecht	binominal
AGE	Alter	numerisch
YEAR	Jahr	numerisch
GERMBORN	in Deutschland geboren	binominal
CORIGIN	Herkunftsland	nominal
ORTKINDH	Ort der Kindheit	nominal
ORTKIND1	Immer noch Ort der Kindheit	nominal
LOC1989	Aufenthalt im Jahr 1989	nominal
BULA	Bundesland	nominal
FAMSTD	Familienstand	nominal
TYPHH	Haushaltstyp	nominal
HHGR	Haushaltsgröße	numerisch
HHCHILDREN	Kinder unter 16 Jahren im Haushalt	binominal
SUMKIDS	Anzahl der leiblichen Kinder	numerisch
PSBIL	Schulbildung	nominal
PBBIL01	Beruflicher Bildungsabschluss	nominal
PBBIL02	Hochschulabschluss	nominal
BILZEIT	Dauer der Ausbildung (in Jahren)	numerisch
NACE	Wirtschaftszweig/Branche	nominal
IS88	Berufsklassifikation nach ISCO88	nominal
OEFFD	im öffentlichen Dienst	nominal
ERLJOB	Tätigkeit im erlernten Beruf	nominal
AUSB	für Tätigkeit erforderliche Ausbildung	nominal
ERWZEIT	Dauer der Betriebszugehörigkeit	numerisch
AUTONO	berufliche Autonomie	nominal
BETR	Unternehmensgröße	nominal
ZYKLUS	Konjunkturzyklus	nominal
WACHSTUM	BIP-Wachstum (gegenüber Vorjahr)	numerisch
WACHSTUM_L1	BIP-Wachstum des letzten Jahres	numerisch
WACHSTUM_L2	BIP-Wachstum des vorletzten Jahres	numerisch
WACHSTUM_L3	BIP-Wachstum von vor drei Jahren	numerisch
INFLATION	Verbraucherpreissteigerung	numerisch
LEGISLATUR	Legislaturperiode	numerisch
WIEDERVEREINIGUNG	Indikator der Wiedervereinigung	binominal
LFSC_Working_Not-working	Wechsel erwerbstätig/nicht erwerbstätig	binominal
LFSC_Working_Unemployed	Wechsel erwerbstätig/arbeitslos	binominal
LFSC_Jobbing_Not-working	Wechsel nebenerwerbstätig/nicht erwerbstätig	binominal
LFSC_Jobbing_Unemployed	Wechsel nebenerwerbstätig/arbeitslos	binominal
LFSC_Unemployed_Working	Wechsel arbeitslos/erwerbstätig	binominal
LFSC_Unemployed_Jobbing	Wechsel arbeitslos/nebenerwerbstätig	binominal
LFSC_Training_Working	Wechsel in Ausbildung/erwerbstätig	binominal
LFSC_Training_Jobbing	Wechsel in Ausbildung/nebenerwerbstätig	binominal
LFSC_Parenthood_Working	Wechsel in Mutterschutz/erwerbstätig	binominal
LFSC_Parenthood_Jobbing	Wechsel in Mutterschutz/nebenerwerbstätig	binominal

7 Deskriptive Analyse

Nachdem die Daten aus dem SOEP-Datensatz extrahiert und vorverarbeitet wurden, können sie analysiert werden. Vor dem tatsächlichen Einsatz analytischer Verfahren empfiehlt es sich jedoch aus mehreren Gründen, die Daten zunächst deskriptiv zu untersuchen und zu beschreiben. Zum einen ist es vor der Analyse hilfreich, schon durch die deskriptive Analyse erkennbare Zusammenhänge sowie Eigenheiten der Daten gefunden zu haben. Aus diesen können unter Umständen Strategien hinsichtlich der eigentlichen Datenanalyse abgeleitet werden. Des Weiteren ist es in Bezug auf die später anzuwendenden maschinellen Lernverfahren etwa von Nöten, die empirischen Verteilungen der Zielvariablen zu kennen, um später die Performanz der Lernverfahren bewerten zu können.

In diesem Kapitel erfolgt daher eine kurze deskriptive Analyse der Daten. Die folgende Erörterung gliedert sich dabei in zwei Teile. Im ersten Teil erfolgt in Kapitel 7.1 eine kurze Beschreibung der Lernaufgaben. Diese erfolgt auf Basis des vorverarbeiteten Datensatzes bereits innerhalb des Experimentier-Frameworks, welches in Kapitel 6.10 vorgestellt wurde.

Der zweite Teil, d.h. Kapitel 7.2, befasst sich mit der Erfassung sowie Beschreibung zeitlicher Effekte in Bezug auf die Zielattribute der Lernaufgaben in den extrahierten Paneldaten. Die Erfassung dieser zeitlichen Effekte erfolgt auf den Daten in der Form, wie sie vor der Transformation der Paneldaten von der Ein-Tabellen-Form in das Long-Format (vgl. Kapitel 6.5) vorliegen. Sie ist damit in gewisser Weise ein vom eigentlichen Analyseprozess losgelöster und unabhängiger Exkurs. Dennoch ist die exkursartige Durchführung und Beschreibung dieser Erfassung sowie ihrer Resultate an dieser Stelle sinnvoll, da sie eine gesonderte Betrachtung einiger zeitlicher Effekte, deren Analyse im Rahmen der Paneldatenanalyse häufig besonders schwierig ist, ermöglicht. Dies ist wiederum hilfreich, da zeitliche Komponenten zwar innerhalb der Vorverarbeitung durch explizite Erzeugung des Attributs `YEAR` einbezogen wurden und somit die Entdeckung zeitlicher Effekte prinzipiell möglich ist, mit der Transformation ins Long-Format allerdings unter Umständen eine gewisse Konzentration auf Querschnittseffekte einhergeht. Eine vorherige Kenntnis eventuell auftretender zeitlicher Effekte ist daher keinesfalls nachteilig, bietet im Gegenteil sogar den Vorteil, die später zu erlangenden Analyseergebnisse zumindest ansatzweise hinsichtlich ihrer Korrektheit kontrollieren und in Bezug auf ihre Aussagekraft analysieren zu können.

7.1 Lernaufgaben

Nachdem die Paneldaten gemäß Kapitel 6 auf- und vorbereitet wurden, stehen sie im Experimentier-Framework, welches in Kapitel 6.10 vorgestellt wurde, zur Verfügung. Dies bedeutet, dass für jede Zielvariable, die jeweils einen der Übergänge aus Abbildung 4.2 beschreibt, die vorverarbeiteten Daten geladen, das entsprechende Label und dementsprechend vermeintlich erklärende Attribute ausgewählt werden. Anschließend werden wie Beispiele mit fehlendem Label, die nicht in die Analyse des jeweiligen Übergangs einzubeziehen sind, gelöscht. Der dann für eine Lernaufgabe entstehende Datensatz ist Eingabe für die

später anzuwendenden Datenanalyseverfahren und somit auch Grundlage einer ersten deskriptiven Beschreibung für die jeweilige Analyseaufgabe. Die Komplexität und der Umfang einer Analyseaufgabe werden durch verschiedene Komponenten bestimmt. Diesbezüglich ist zunächst der Umfang des Datensatzes, also die Anzahl der Beispiele, zu nennen. Des Weiteren spielen die Anzahl und Typen der Attribute eine Rolle. Diese sind für die zu betrachtenden zehn Analyseaufgaben in Tabelle 7.1 aufgelistet. Folgende Beobachtungen sollen dabei

Tabelle 7.1: Übersicht der Lernaufgaben

Label	Attribute		Beispiele	Klassenverteilung	
	nominal	numerisch		negativ	positiv
LFSC_Working_Not-working	22	12	154.892	97,42%	2,58%
LFSC_Working_Unemployed	22	12	154.892	96,76%	3,24%
LFSC_Jobbing_Not-working	22	12	8.141	88,39%	11,61%
LFSC_Jobbing_Unemployed	22	12	8.141	96,39%	3,61%
LFSC_Unemployed_Working	15	11	13.705	71,96%	28,04%
LFSC_Unemployed_Jobbing	15	11	13.705	96,17%	3,83%
LFSC_Training_Working	15	11	10.610	72,47%	27,53%
LFSC_Training_Jobbing	15	11	10.610	87,52%	12,48%
LFSC_Parenthood_Working	15	11	3.528	77,47%	22,53%
LFSC_Parenthood_Jobbing	15	11	3.528	92,96%	7,04%

kurz erläutert werden: Die Anzahl der Attribute unterscheidet sich für die ersten vier und die letzten sechs Analyseaufgaben. Dies rührt - wie schon zuvor angedeutet - daher, dass für die Übergänge aus einem Zustand der (Neben-)Erwerbstätigkeit solche Attribute einbezogen werden können, die die Erwerbstätigkeit vor dem potentiellen Übergang beschreiben. Im Umkehrschluss sind diese aus den Analyseaufgaben bzgl. der anderen Übergänge zu entfernen. Die Anzahl der Beispiele stimmt für jeweils die Analyseaufgaben überein, die aus einem (gemeinsamen) Zustand in andere Zustände übergehen. Dies ist insofern leicht einzusehen, als dass die Zielattribute jeweils den Übergang von einem Zustand in einen anderen Zustand beschreiben, unter der Bedingung, dass der erste Zustand vorgelegen hat. Somit haben diese Zielattribute für alle Beispiele, in denen der erste Zustand vorlag einen gültigen Wert. Die Anzahlen der Beispiele für Analyseaufgaben mit einem gemeinsamen vorherigen Zustand stimmen also überein.

Aus Perspektive der Datenanalyseverfahren sind neben den oben gemachten Beobachtungen vor allem die Klassenverteilungen relevant. Hierbei fällt auf, dass generell die negative Klasse (die der Nichtdurchführung des jeweiligen Übergangs entspricht) überwiegt. Unterschiede finden sich jedoch darin, wie stark die negative Klasse im Vergleich zur positiven Klasse überwiegt. Diesbezüglich ist festzustellen, dass die negative Klasse im niedrigsten Fall bereits eine relative Häufigkeit von über 70 Prozent hat, in den höchsten Fällen sogar eine relative Häufigkeit von etwa 97 Prozent besitzt. Demnach besitzt die positive Klasse relative Häufigkeiten von nur knapp 3 bis etwa knapp 30 Prozent. Vor allem in den Fällen, in denen der Anteil der positiven Klasse sehr gering ist, ergibt sich möglicherweise eine hohe Schwierigkeit des Erfassens der positiven Klasse durch die anzuwendenden Lernverfahren.

7.2 Beschreibung zeitlicher Effekte

Vor der Erläuterung der Anwendung von Datenanalyseverfahren auf den oben beschriebenen Daten sowie der Vorstellung der daraus resultierenden Ergebnisse in den nachfolgenden Kapiteln, beschreibt das vorliegende Unterkapitel zeitliche Effekte, die in Bezug auf die einzelnen Zielattribute, d.h. in Bezug auf die Übergänge zwischen Arbeitsmarktzuständen, existieren. Ein Mittel zur Beschreibung dieser Effekte ist die Darstellung von relativen Häufigkeiten der Arbeitsmarktzustände bzw. der Arbeitsmarktzustandsübergänge als Zeitreihe. Dies bedeutet, dass Zeitreihendaten erzeugt wurden, die für jeden Arbeitsmarktzustand in jedem Zeitpunkt den (relativen) Anteil der Personen angeben, die sich in diesem Zustand befinden. Analog wurden für die Arbeitsmarktzustandsübergänge Zeitreihen erzeugt, die die relative Häufigkeit der jeweiligen Zustandsübergänge in den einzelnen Perioden erfassen. Die letztgenannten relativen Häufigkeiten geben die empirischen, d.h. geschätzten, Wahrscheinlichkeiten an, in den einzelnen Perioden die jeweiligen Übergänge durchzuführen.

Die Berechnung der relativen Häufigkeiten wurde auf Basis der aus dem SOEP extrahierten Daten im Wide-Format durchgeführt. Um die in dieser Arbeit betrachteten Arbeitsmarktzustände und die in Abbildung 4.2 dargestellten Übergänge zwischen diesen zu beschreiben, wurden die Variablen `LFSxx`, die Arbeitsmarktzustände auf jährlicher Basis im SOEP erfassen, gemäß der in Kapitel 6.3 beschriebenen Schritte vorverarbeitet. Letztlich erfolgte also eine Vorverarbeitung der Daten gemäß der beschriebenen Schritte bis einschließlich Kapitel 6.3. Der resultierende Datensatz enthält zum einen die Arbeitsmarktzustände sowie die Zustandsübergänge als multivariate Wertereihen in Wide-Format. Diese Wertereihen wurden mittels des für diese Arbeit als Teil des Panel-Plugins implementierten RapidMiner-Operators `MultivariateSeriesAggregation` zu relativen Häufigkeiten aggregiert. Ausgangspunkt ist dabei, dass die multivariate Wertereihe, die zu aggregieren ist, als einzelne Attribute vorliegen, die jeweils einen Querschnitt von Werten für die einzelnen Personen zu einem Zeitpunkt darstellen. Der Operator berechnet daher für jedes Attribut, aus denen die zu aggregierende Wertereihe besteht, die relativen Häufigkeiten der vorkommenden Attributwerte. Anschließend fügt der Operator diese so berechneten relativen Häufigkeiten für die einzelnen Zeitpunkte zu Zeitreihen zusammen. Somit wird pro Attributwert eine Zeitreihe erzeugt, die die jeweilige relative Häufigkeit des Attributwerts in jedem Zeitpunkt angibt. Fehlende Werte sind bei der Aggregation sowohl bei der Berechnung der Häufigkeit des Attributwertes als auch bei der Berechnung der Größe der zu diesem Zeitpunkt betrachteten Stichproben, die zur Berechnung der relativen Häufigkeit erfasst werden muss, zu ignorieren.

7.2.1 Periodeneffekte

Durch Anwendung des Operators `MultivariateSeriesAggregation` wurden anhand des soeben beschriebenen Verfahrens Zeitreihendaten erzeugt, die die relativen Häufigkeiten der betrachteten Arbeitsmarktzustände pro Welle, d.h. pro Jahr, angeben. Diese sind in Abbildung 7.1 grafisch dargestellt und spiegeln Periodeneffekte in Bezug auf die (empirische) Verteilung von Arbeitsmarktzuständen in der SOEP-Stichprobe wider. Wie jedoch leicht zu erkennen ist, sind diese Effekte in der Verteilung begrenzt: der Anteil der erwerbstätigen schwankt nur in einem relativ engen Band um knapp 55 Prozent. Eine Ausnahme dessen ist in der Zeit um die Wiedervereinigung zu erkennen. Hier könnten jedoch auch einmalige

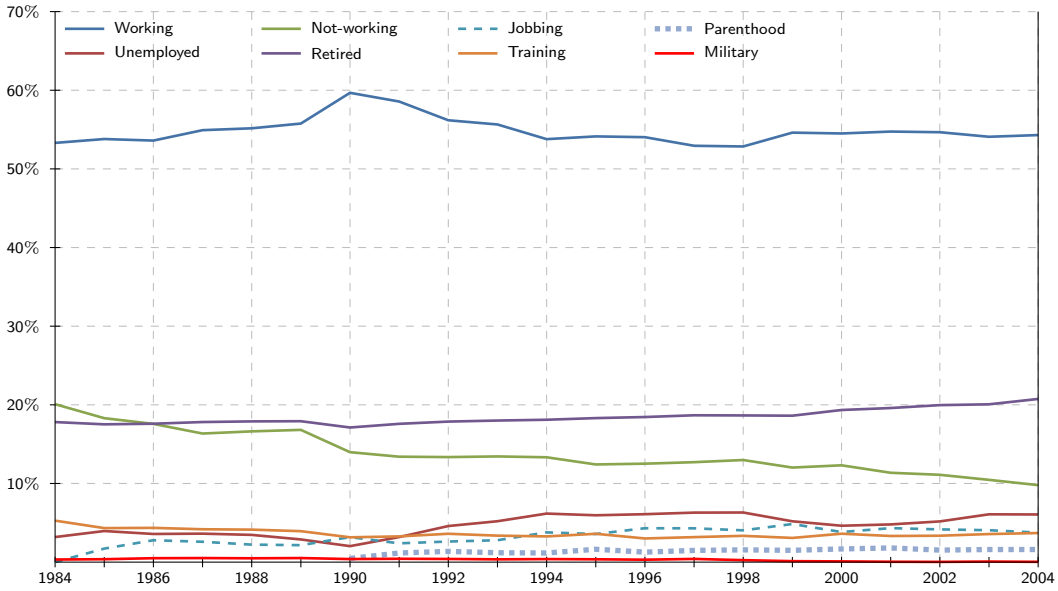


Abbildung 7.1: Arbeitsmarktzustände im Zeitverlauf (1984-2004). Jährliche Quoten der Arbeitsmarktzustände. Quelle: SOEP, eigene Berechnung

Effekte etwa durch Hinzunahme der Stichprobe von Ostdeutschen eine Rolle spielen. Der Anteil der sich im Ruhestand befindenden Personen steigt über den betrachteten Zeitraum nahezu kontinuierlich an. Auch hier ist Vorsicht in der Interpretation geboten, da auch dies unter Umständen auf erhebungstechnische Aspekte zurückzuführen ist. Hypothetisch denkbar wäre etwa, dass generell der Anteil der jüngeren Personen im SOEP sinkt, der Anteil der Rentner daher automatisch steigt. Auffälliger und vermutlich nicht alleinig durch erhebungstechnische Gesichtspunkte zu erklären ist dagegen der Abfall des Anteils der Nichterwerbstätigen von etwa 20 Prozent im Jahr 1984 bis auf etwa 10 Prozent im Jahr 2004. Die Kurve des Anteils von Arbeitslosen verläuft in etwa analog zur in Abbildung 4.1 abgebildeten Arbeitslosenquote - allerdings auf niedrigerem Niveau. Die relativen Häufigkeiten der anderen Zustände schwanken eher wenig, lediglich der Anteil der Nebenerwerbstätigen nimmt zeitweise - in den 1990er Jahren - etwas zu, danach jedoch auch wieder leicht ab.

Weitere interessante Gegebenheiten zeigen sich in der Darstellung aus Abbildung 7.2. Die abgebildeten Zeitreihen sind Ergebnis der - vollkommen analog durchgeführten - Aggregation der Wertereihen, die die einzelnen Arbeitsmarktzustandsübergänge und damit Flüsse zwischen den Zuständen erfassen. In Abbildung 7.2 sind für jeden betrachteten Zustandsübergang die relativen Häufigkeiten der positiven Übergänge dargestellt. Die dargestellten Zeitreihen erfassen also für jedes Jahr die Anzahl der Personen, die den jeweiligen Übergang durchgeführt haben, in Relation zur Anzahl der Personen, die sie sich im Vorjahr im Ausgangszustand des Überganges befunden haben. Somit beschreiben sie die bedingte Wahrscheinlichkeit eines Zustandsüberganges über die Zeit. Bei den meisten Graphen sind generell mittel- bis langfristige Veränderungen relativ deutlich auszumachen: Die relative Häufigkeit des Übergangs von erwerbstätig nach nichterwerbstätig ist im betrachteten Zeitraum im Wesentlichen rückläufig, überraschend ist allerdings der abrupte Anstieg von 1987 bis 1989. Der Graph der Zeitreihe zum Übergang von der Erwerbstätigkeit in die Arbeitslosigkeit bewegt sich ähnlich zur Arbeitslosenquote scheinbar konjunkturbe-

(a) LFSC_Working_Not-working

(b) LFSC_Working_Unemployed

(c) LFSC_Jobbing_Not-working

(d) LFSC_Jobbing_Unemployed

(e) LFSC_Unemployed_Working

(f) LFSC_Unemployed_Jobbing

(g) LFSC_Training_Working

(h) LFSC_Training_Jobbing

(i) LFSC_Parenthood_Working

(j) LFSC_Parenthood_Jobbing

Abbildung 7.2: Arbeitsmarktzustandsübergänge im Zeitverlauf (1984-2004). Jährliche Übergangsraten. Quelle: SOEP, eigene Berechnung

dingt mit Tiefständen in den Boom-Jahren 1990 und 2000 sowie starken Anstiegen in den darauffolgenden wirtschaftlichen Abschwüngen. Die relative Häufigkeit des Übergangs von nebenerwerbstätig nach nichterwerbstätig verläuft unter erhöhter Volatilität analog zur ersten beschriebenen und ist damit ebenfalls rückläufig. Die Zeitreihe von nebenerwerbstätig nach arbeitslos zeigt dagegen einen klar ansteigenden Verlauf, vor allem im Abschwung nach 2001. Wiederum rückläufig ist die (generell überraschend hohe) geschätzte Wahrscheinlichkeit eines Übergangs von der Arbeitslosigkeit in die Erwerbstätigkeit, ansteigend dagegen die relative Häufigkeit des Übergangs von der Arbeitslosigkeit in die Nebenerwerbstätigkeit. Diese Diskrepanz, d.h. einen Rückgang bei Wechsel in die Erwerbstätigkeit bei gleichzeitigem Anstieg der Wechsel in die Nebenerwerbstätigkeit ist auch bei Wechseln aus dem Zustand der Ausbildung zu beobachten. Deutliche Entwicklungen in den Zeitreihen zu den Übergängen `LFSC_Parenthood_Working` und `LFSC_Parenthood_Jobbing` sind gesichert nicht zu beschreiben, da sie einzelne Ausreißer bzw. extrem schwankende Verläufe aufweisen. Dies rührt unter Umständen daher, dass die Werte einer *selektiven Verzerrung* (beispielsweise resultierend aus einer Erweiterung der SOEP-Stichprobe) unterliegen. Die geringe Anzahl von Beispielen, die zur Analyse speziell dieser Übergänge herangezogen werden können (vgl. auch Tabelle 7.1), verstärkt dabei die Inkonsistenzen.

7.2.2 Alterseffekte

Ergänzend zur deskriptiven Analyse der Periodeneffekte ist auch eine Untersuchung der Alterseffekte bei den Arbeitsmarktzuständen und den Übergängen zwischen ihnen anhand einer ähnlichen Vorgehensweise möglich und wurde im Rahmen dieser Arbeit durchgeführt. Mit dem Operator `MultivariateSeriesAggregation` konnte die dazu notwendige Aggregation prinzipiell analog zur oben beschriebenen Aggregation durchgeführt werden. Zuvor mussten die Daten jedoch dergestalt transformiert werden, dass in der betrachteten Wertereihe ein Attribut nicht mehr den Beobachtungen in einer Periode, d.h. einem Jahr, sondern den Beobachtungen zu einem bestimmten Alter, d.h. einem Lebensjahr, entspricht. Für die Durchführung einer solchen Transformation wurde im Rahmen dieser Arbeit der RapidMiner-Operator `MultivariateSeriesAdjustment` implementiert, der in der Lage ist, die Beobachtungen für die Untersuchungseinheiten am Zeitpunkt eines Ereignisses, der durch ein Attribut für jede Untersuchungseinheit angegeben ist, auszurichten. Dies sei verdeutlicht anhand des schematisch in Abbildung 2.9(b) dargestellten Panels mit dem markierten Ereignis. Durch Anwendung des Operators werden die Daten in die in Abbildung 7.3 gezeigte Form überführt.

Die aus der Aggregation hervorgehenden relativen Häufigkeiten der Arbeitsmarktzustände in Abhängigkeit vom Lebensalter sind in Abbildung 7.4 dargestellt. Die Werte für 16- bzw. 17-Jährige sind aufgrund von anomalen Verzerrungen mit Vorsicht zu erörtern. Ansonsten entsprechen die meisten Verläufe klar den Erwartungen. Für die Analyse im Rahmen dieser Arbeit interessant ist vor allem der erhöhte Anteil von Arbeitslosen bei über 55-Jährigen sowie der relativ hohe Anteil von Nebenerwerbstätigen bei unter 20-Jährigen. Auch der mit einem Lebensalter von 60 Jahren etwas ansteigende Anteil der Nebenerwerbstätigen ist erwähnenswert. Die in Abbildung 7.4 gezeigten relativen Häufigkeiten von Arbeitsmarktzustandsübergängen in Bezug auf das Lebensalter bestätigen teilweise das aus der vorherigen Abbildung entnommene Bild. Die für die Altersgruppe der 55- bis 60-Jährigen erhöhte Wahrscheinlichkeit eines Übergangs in die Arbeitslosigkeit so bestätigt. Auch der Graph zum Übergang von die Arbeitslosigkeit in die Erwerbstätigkeit zeigt klar eine Abhängigkeit

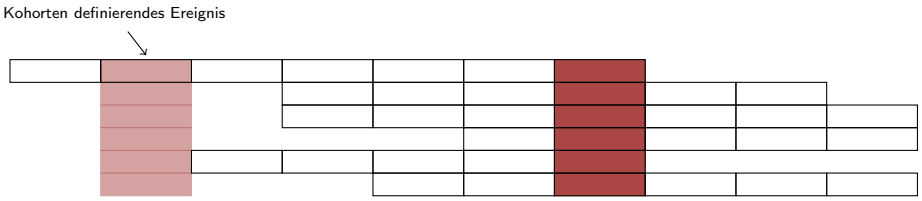


Abbildung 7.3: Kohorteneffekt in ereignis-adjustierten Paneldaten

dieses Übergangs vom Alter: die (geschätzte) Wahrscheinlichkeit eines Übergangs nimmt mit zunehmendem Alter rapide ab. Die Abbildungen bzgl. der Übergänge in die Nichterwerbstätigkeit zeigen mit dem Alter zunehmende Wechsel. Die Wechsel aus dem Zustand erwerbstätig nehmen allerdings erst ab einem Alter von etwa 55 stark zu. Bei Wechseln aus der Nebenerwerbstätigkeit verläuft der Anstieg dagegen ab einem Alter von 20 unter relativ hoher Volatilität beinahe linear. Die Interpretation der anderen Graphen gestaltet sich aufgrund der hohen Volatilität und vielen Ausreißer, die teilweise auf Dateninkonsistenzen hindeuten, schwierig. Eindeutige sowie auffällige Muster sind eher nicht zu erkennen.

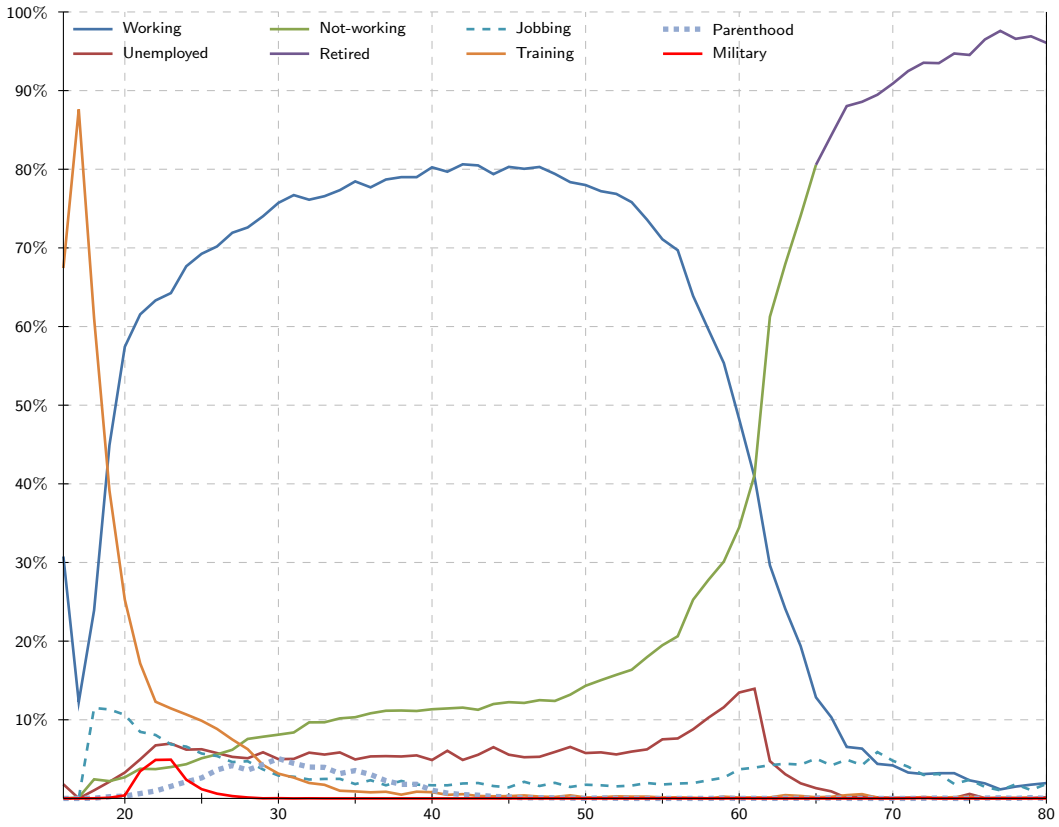
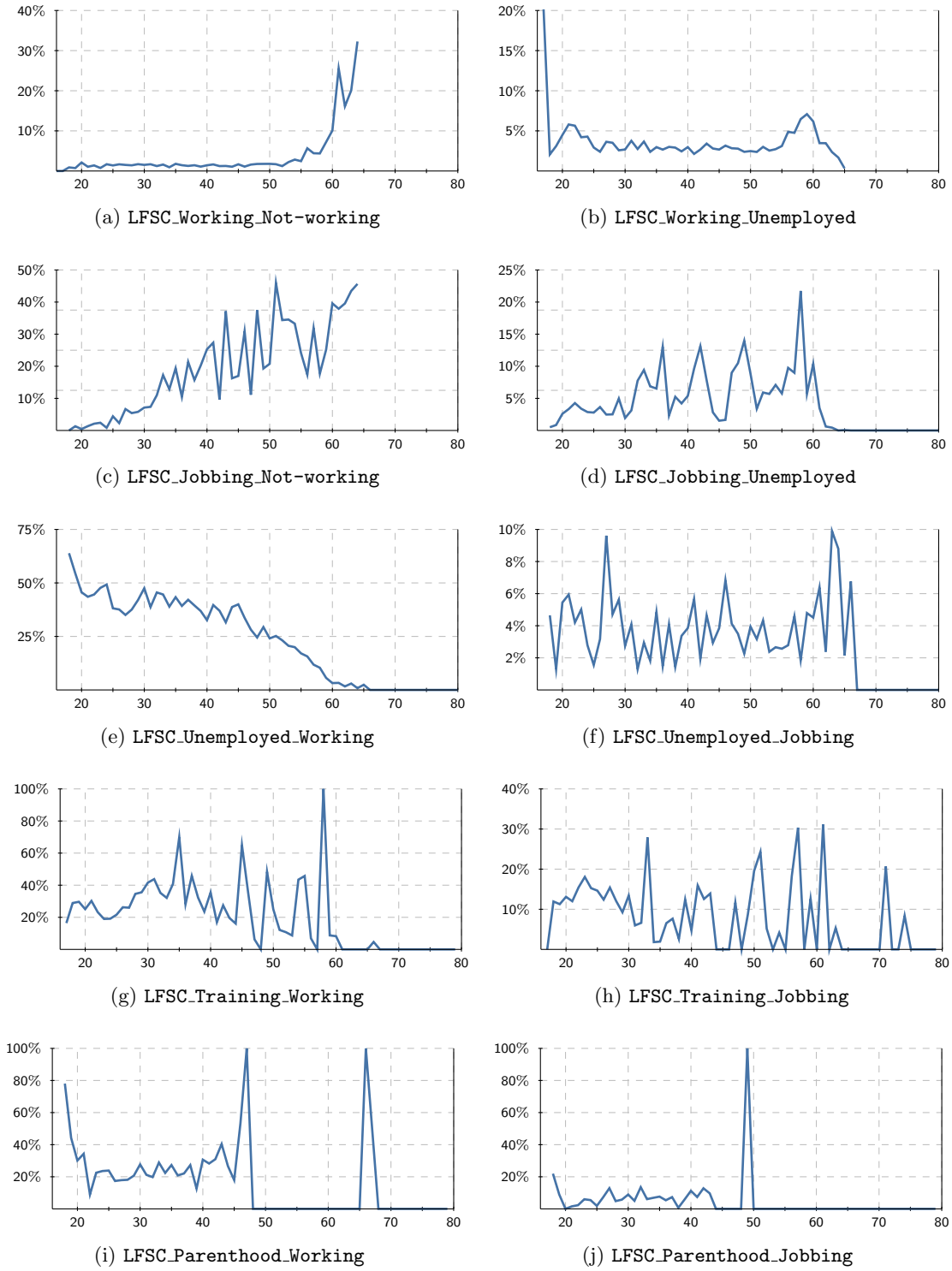


Abbildung 7.4: Arbeitsmarktzustände nach Lebensalter. Quoten der Arbeitsmarktzustände bezogen auf das Lebensalter von Personen in Jahren. Quelle: SOEP, eigene Berechnung



(a) `LFSC_Working_Not-working` (b) `LFSC_Working_Unemployed`

(c) `LFSC_Jobbing_Not-working` (d) `LFSC_Jobbing_Unemployed`

(e) `LFSC_Unemployed_Working` (f) `LFSC_Unemployed_Jobbing`

(g) `LFSC_Training_Working` (h) `LFSC_Training_Jobbing`

(i) `LFSC_Parenthood_Working` (j) `LFSC_Parenthood_Jobbing`

Abbildung 7.5: Arbeitsmarktzustandsübergänge nach Lebensalter. Übergangsraten von Arbeitsmarktzuständen bezogen auf das Lebensalter von Personen in Jahren. Quelle: SOEP, eigene Berechnung

8 Merkmalskorrelation, -gewichtung und -selektion

Will man wie in der gegebenen Analyseaufgabe den Einfluss von Attributen auf eine Zielvariable bestimmen, so kann man einen ersten Eindruck der Wichtigkeit der Attribute hinsichtlich ihres Einflusses durch Berechnung einfacher Kennziffern wie etwa Korrelationen zwischen den Attributen bzw. zwischen den Attributen und dem Zielattribut gewinnen. Auf Basis dieser Kennziffern kann zudem eine Vorauswahl vermeintlich einzubeziehender Attribute vor der eigentlichen Anwendung eines maschinellen Lernverfahrens erfolgen. Ebenfalls diesem Zweck können Verfahren zur Berechnung von Attributgewichten dienen.

Dieses Kapitel erläutert die Berechnung von Korrelationen und die Anwendung von Verfahren zur Gewichtung von Attributen auf dem in Kapitel 6.9 beschriebenen Datensatz und stellt die resultierenden Ergebnisse vor. Kapitel 8.1 beschreibt kurz die Berechnung von Korrelationen zwischen numerischen und Kontingenzen zwischen nominalen Attributen. Attributgewichte können bestimmt werden mittels der Berechnung des Informationsgewinns einzelner Attribute. Eine Erläuterung dessen erfolgt in Kapitel 8.2. Als drittes und letztes Verfahren zur Gewichtung von Attributen wird in Kapitel 8.3 schließlich das Verfahren Relief vorgestellt. Alle genannten Verfahren wurden innerhalb der in Kapitel 6.10 beschriebenen Experimentierumgebung durchgeführt und Korrelationen, Kontingenzen sowie Attributgewichte anhand der genannten Verfahren berechnet. Die erzielten Ergebnisse werden in Kapitel 8.4 zusammengefasst.

8.1 Merkmalskorrelationen und -kontingenzen

Mit dem Begriff *Korrelation* wird abstrakt ein Zusammenhang zwischen Merkmalen bezeichnet. Aus der reinen Existenz einer Korrelation können jedoch keine kausalen Zusammenhänge geschlossen werden. D.h. es kann nicht gefolgert werden, dass etwa das eine Merkmal das andere beeinflusst oder umgekehrt. Genauso gut könnten beide Merkmale in einem kausalen Zusammenhang zu einem dritten, unbeobachteten Attribut stehen oder aber es existiert überhaupt kein kausaler Zusammenhang. Dennoch bietet die eher deskriptive Beobachtung bzw. Berechnung von Korrelationen vor der eigentlichen Anwendung von analytischen Verfahren zwei gute Ansatzpunkte für die Auswahl von Attributen für den Einbezug in die spätere Anwendung dieser Verfahren. Zum einen deuten vorhandene Korrelationen zwischen Attributen und der Zielvariable auf potentielle Einflüsse dieser Attribute auf die Zielvariable hin. Zum zweiten können mögliche Redundanzen zwischen Attributen, die hoch korrelieren, entdeckt werden. Somit kann die Berechnung der Korrelation zwischen Merkmalen und Zielvariable einen ersten Eindruck existierender Zusammenhänge geben.

Zur Beobachtung der Korrelation zwischen Merkmalen bedarf es eines Maßes, welches die Korrelation quantifiziert. Verschiedene solche Maße wurden entwickelt und unterliegen teilweise einer weiten Verbreitung. Das bekannteste (empirische) Maß für die Korrelation

zwischen Merkmalen ist der *Pearsonsche Korrelationskoeffizient*, der durch folgende Definition, die beispielsweise auch Hartung et al. (2005) entnommen werden kann, gegeben ist:

Definition 8.1 (Korrelationskoeffizient) *Seien X und Y zwei Zufallsvariablen. Dann ist die* Korrelation *gegeben durch*

$$\rho(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

Liegen für X und Y Beobachtungen $x_1,\ldots,x_n$ sowie $y_1,\ldots,y_n$ vor, so ist ein Schätzer für die Korrelation gegeben durch die Stichprobenkorrelation*, die auch* Pearsonscher Korrelationskoeffizient *genannt wird. Sie wird berechnet durch*

$$\hat{\rho}(X,Y) := \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

Der Pearsonsche Korrelationskoeffizient ist also gleich dem Quotienten aus der empirischen Kovarianz und dem Produkt der empirischen Standardabweichungen zweier beobachteter Merkmale. Im Regelfall sind diese nicht bekannt und müssen geschätzt werden, um so den empirischen Korrelationskoeffizienten als Schätzung zu berechnen. Der oben definierte Korrelationskoeffizient berechnet ein Maß für einen linearen Zusammenhang zweier Merkmale. Er kann Werte zwischen -1 für einen perfekt negativen bis 1 für einen perfekt positiven linearen Zusammenhang annehmen. Ein Wert von 0 liegt vor, wenn kein Zusammenhang zwischen den Merkmalen existiert. Die Einschränkung hinsichtlich der Linearität des Zusammenhangs impliziert zwei gravierende Nachteile des Korrelationskoeffizienten. Zum einen müssen die betrachteten Merkmale mindestens intervallskaliert sein, um Mittelwerte sowie Differenzen berechnen zu können. Zum zweiten können auch bei intervallskalierten Merkmalen andere Formen der Zusammenhänge, also etwa quadratische oder exponentielle, nur unzureichend dargestellt werden. Unzureichend dargestellt heißt in diesem Fall, dass zwar beispielsweise ein perfekter quadratischer Zusammenhang zwischen zwei Merkmalen besteht, der berechnete Korrelationskoeffizient jedoch weit unter dem möglichen Maximalwert liegen kann.

Im Rahmen dieser Arbeit ist die Berechnung obiger Korrelationskoeffizienten insofern problematisch, als dass viele der verwendeten Attribute (und auch die Zielattribute) nominal sind. Für polynominale Attribute kann der Pearsonsche Korrelationskoeffizient im Allgemeinen nicht angewendet werden. Vielfach werden nominale Werte jedoch durch Werte auf einer reellen Skala abgebildet (vor allem etwa in Softwarepaketen zur statistischen Analyse) und Korrelationskoeffizienten können auf diesen reellen Werten berechnet werden. Allerdings ist dann bei der Interpretation dieser Korrelationskoeffizienten große Vorsicht geboten, da je nach Zuordnung der nominalen Werte zu reellen Werten der Wert des zu berechnenden Korrelationskoeffizienten schwanken kann.

Im Gegensatz zu obiger Beschreibung kann der Korrelationskoeffizient für binäre Attribute, d.h. Attribute, die nur den Wert 0 oder 1 annehmen können, sehr wohl berechnet werden. Diese entsprechen jedoch binominalen Attributen mit entsprechend gewählter Skala. Um das oben beschriebene Problem zu umgehen, bietet sich daher in Bezug auf nominale Werte die Möglichkeit, die nominalen Attribute zuvor in Attribute mit binärer Skala zu überführen.

Alternativ können zum Korrelationskoeffizienten ähnliche Kennzahlen verwendet werden, die speziell für nominale Daten entworfen wurden, beispielsweise der *korrigierte Kontingenzkoeffizient*.

Definition 8.2 (Korrigierter Kontingenzkoeffizient) *Seien für zwei nominale Zufallsvariablen X und Y, die k_x respektive k_y verschiedene Werte annehmen können, Beobachtungen $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ gegeben. Seien ferner $h_{i\cdot}$ (mit $i = 1, \ldots, k_x$) sowie $h_{\cdot j}$ (mit $j = 1, \ldots, k_y$) die Randhäufigkeiten der gemeinsamen, empirischen Verteilung der beiden Variablen und h_{ij} die gemeinsame Häufigkeit des i-ten möglichen Attributwerts für X sowie des j-ten Wert für Y. Dann ist der χ^2-Koeffizient gegeben durch*

$$\chi^2 = \sum_{i=1}^{k_x} \sum_{j=1}^{k_y} \frac{\left(h_{ij} - \frac{h_{i\cdot} h_{\cdot j}}{n}\right)^2}{\frac{h_{i\cdot} h_{\cdot j}}{n}}.$$

Basierend darauf ist der korrigierte Kontingenzkoeffizient *gegeben durch*

$$c_{korrigiert} = \sqrt{\frac{k}{k-1}} \sqrt{\frac{\chi^2}{n + \chi^2}},$$

wobei $k := \min\{k_x, k_y\}$ ist.

Der χ^2-Koeffizient misst demnach Abweichungen der gemeinsamen, empirischen Verteilung von der durch die Randverteilungen erwarteten. Der Kontingenzkoeffizient bezieht zudem die Anzahl der Beobachtungen mit ein. Die Korrektur durch den ersten Wurzelfaktor erzeugt den resultierenden Wertebereich von 0 bis 1, wobei der Wert 1 den größtmöglichen Zusammenhang bedeutet. Eine ausführliche Herleitung des korrigierten Kontingenzkoeffizienten und anderer Assoziationsmaße ist ebenfalls in Hartung et al. (2005) zu finden.

Mit den in diesem Abschnitt definierten Kennzahlen können Zusammenhänge der vermeintlich erklärenden Variablen mit den Zielvariablen erfasst werden. Allerdings bestehen Nachteile dahingehend, dass Korrelations- und Kontingenzkoeffizienten untereinander nicht miteinander verglichen werden können, und die Kennzahlen für numerische und nominale Attribute daher nicht miteinander in Beziehung gesetzt werden können. Dennoch bieten beide Konzepte eine erste Ansatzmöglichkeit zur Identifikation für spätere Analysen wichtiger Attribute.

8.2 Gewichtung nach Informationsgewinn

Attributgewichte können - je nach Verfahren - erste Anzeichen der Wichtigkeit ihrer Einbeziehung in analytische Verfahren bieten und so - falls gewünscht - zur Vorauswahl von Attributen, die in einem Analyseverfahren berücksichtigt werden sollen, dienen. Das im Folgenden vorgestellte Verfahren berechnet als Attributgewichte den Informationsgewinn der einzelnen Attribute auf einer Menge von Beispielen. Der Informationsgewinn findet vielfach auch in maschinellen Lernverfahren, etwa bei Baum- oder Regellernern Verwendung und wird in diesem Zusammenhang beispielsweise von Quinlan (1993) beschrieben. Der Informationsgewinn basiert auf dem Maß der Entropie, einem Maß aus der Informationstheorie, welches den Informationsgehalt von Daten in Bits misst.

Die Entropie ist wie folgt definiert:

Definition 8.3 (Entropie) *Sei Y eine Zufallsvariable. Dann ist die* Entropie *gegeben durch*[1]

$$\mathrm{H}(Y) := -\sum_{y \in Y} \Pr(y) \log \Pr(y).$$

Für eine gegebene Menge von Beispielen kann die Berechnung der Entropie erfolgen mittels Schätzung der Wahrscheinlichkeiten durch die relativen Häufigkeiten in der jeweiligen Menge. Basierend auf der Entropie gibt der Informationsgewinn an, um wie viel sich die Entropie einer Beispielmenge erhöht, wenn man sie nach den Werten eines Attributes aufspaltet. Formal lässt sich dies folgendermaßen definieren:

Definition 8.4 (Informationsgewinn) *Seien zwei Zufallsvariablen X und Y gegeben. Dann heißt*

$$\begin{aligned} \mathrm{IG}_Y(X) &:= \mathrm{H}(Y) - \sum_{x \in X} \Pr(x)\, \mathrm{H}(Y|x) \\ &= -\sum_{y \in Y} \Pr(y) \log \Pr(y) + \sum_{x \in X} \Pr(x) \sum_{y \in Y} \Pr(y|x) \log \Pr(y|x) \end{aligned}$$

der Informationsgewinn *von X in Bezug auf Y.*

Um Attributgewichte als Indikator der Wichtigkeit der einzelnen Attribute als Erklärungskomponenten der Zielvariablen zu berechnen, ist in der Definition des Informationsgewinns als Zufallsvariable Y natürlich das Zielattribut zu verwenden. Das Gewicht eines Attributs ergibt sich dann einfach als Informationsgewinn dieses Attributes in Bezug auf das Label. Hat ein Attribut einen höheren Informationsgewinn als ein anderes Attribut, so ist es vermutlich bedeutender bzw. genauer in der Erklärung der Zielvariable. Es erhält dann logischerweise ein höheres Gewicht.

8.3 Relief

Ein anderes, alternatives Verfahren zur Gewichtung von Attributen ist das Verfahren *Relief*. Initial eingeführt von Kira und Rendell (1992), wurde es beispielsweise von Kononenko (1994) und Sun und Li (2006) weiterentwickelt. Das Ziel von Relief ähnelt dem der Gewichtung nach Informationsgewinn. Relief gewichtet Attribute ebenfalls danach, ob sie in der Lage sind, anhand ihrer auftretenden Attributwerte die Beispielmenge so zu partitionieren, dass gleichzeitig eine möglichst gute und reine Aufspaltung hinsichtlich des Zielattributes resultiert. Im Folgenden wird das Verfahren kurz erläutert, wobei nur der Zweiklassenfall betrachtet wird, da nur dieser für diese Arbeit relevant ist. Die grundsätzliche Vorgehensweise ist in Abbildung 8.1 dargestellt. Aus der Menge gegebener Beispiele werden iterativ zufällig einzelne Beispiele $(\mathbf{x}, y)$ ausgewählt. Für jedes dieser Beispiele werden zwei nächste Nachbarn, d.h. Beispiele, die dem gezogenen Beispielen am ähnlichsten sind, gefunden. Eines davon, der sogenannte *nearest hit* muss den gleichen Zielwert wie das gezogene Beispiel

[1] Zur Erhöhung der Lesbarkeit wurde dabei (sowie im Folgenden) eine verkürzte Schreibweise gewählt. Darin ist mit $\Pr(y)$ die Wahrscheinlichkeit $\Pr(Y = y)$ gemeint. Auf welches Attribut bzw. welche Zufallsvariable sich die Wahrscheinlichkeit jeweils bezieht, sollte durch die Wahl der Buchstaben dabei leicht ersichtlich sein.

Eingabe : Menge von Beispielen E, Distanzfunktionen $d_1, \ldots, d_m$, Abbruchkriterium
Ausgabe: Attributgewichte $w_1, \ldots, w_m$

```
1  for j ← 1 to m do
2      w_j := 0
3  end
4  repeat
5      e ← ZieheBeispiel(E)
6      e^NM ← NearestHit(e)
7      e^NH ← NearestMiss(e)
8      for j ← 1 to m do
9          w_j := w_j − d_j(x_j, x_j^NH) + d_j(x_j, x_j^NM)
10     end
11 until Abbruchkriterium erfüllt
```

Abbildung 8.1: Grundlegender Algorithmus von Relief

aufweisen, das andere Beispiel, der *nearest miss*, muss die andere Klasse als Zielwert aufweisen. In jedem Iterationsschritt wird dann das Attributgewicht mittels der abgebildeten Formel aktualisiert. Die Funktion d_j (mit $j = 1, \ldots, m$) bezeichnet dabei ein für das jeweilige Attribut X_j geeignetes Distanzmaß, also allgemein eine Funktion $d : X \times X \to \mathbb{R}_+$, für die Definitheit, Symmetrie und die Dreiecksungleichung gilt. Beispiele hierfür sind der für nominale Attribute vornehmlich verwendete *0-1-loss*, gegeben durch

$$d_{\text{0-1-loss}}(x, x') := \begin{cases} 0, & \text{wenn } x = x' \\ 1, & \text{wenn } x \neq x' \end{cases} \tag{8.1}$$

oder die im Regelfall für numerische Daten benutzte *euklidische Distanz*

$$d_{\text{euklidisch}}(x, x') := |x - x'|. \tag{8.2}$$

Da im Fall von numerischen Attributen die Wertebereiche stark unterschiedlich sein können, wird - um eine Vergleichbarkeit der Attributgewichte zu erreichen - in Relief jedoch eher die Distanzfunktion verwendet, die durch

$$d(x, x') := \frac{|x - x'|}{|x_{\max} - x_{\min}|} \tag{8.3}$$

gegeben ist. Die Werte $x_{\min}$ respektive $x_{\max}$ bezeichnen die minimalen bzw. maximalen in den Beispielen auftretenden Werte. Nach obiger Definition der Gewichtsaktualisierung wirkt sich demnach eine Differenz, d.h. fehlende Übereinstimmung, zwischen dem gezogenen Beispiel und nearest hit vermindernd, eine eine Differenz zwischen gezogenem Beispiel und nearest miss vergrößernd auf das jeweilige Attributgewicht aus. Als Erweiterung der beschriebenen Strategie können beispielsweise auch mehr als zwei nächste Nachbarn berechnet werden und im Aktualisierungsschritt der Durchschnitt der Differenzen von den einzelnen gezogenen Beispielen zu ihren jeweiligen nächsten Nachbarn verwendet werden,

wie in Scherf und Brauer (1997) notiert.

8.4 Ergebnisse

Die in diesem Kapitel vorgestellten Verfahren zur Messung eines Zusammenhangs zwischen den in den Daten vorhandenen Attributen und dem jeweiligen Zielattribut wurden auf den gemäß den Ausführungen in Kapitel 6 vorverarbeiteten Daten innerhalb des in Kapitel 6.10 vorgestellten Experimentier-Frameworks angewandt. Für jedes der zehn Zielattribute wurden daher erstens Korrelationskoeffizienten zwischen den numerischen Attributen und den Zielattributen und zweitens Kontingenzkoeffizienten zwischen den nominalen Attributen und den Zielattributen berechnet. Drittens wurden Informationsgewinne von Attributen in Bezug auf die Label berechnet und viertens das Verfahren Relief für die einzelnen Zielattribute angewandt.

Die absolute Höhe der aus den Verfahren resultierenden einzelnen Attributgewichte ist nur mäßig interessant. Von größerer Bedeutung sind stattdessen zum einen die relative Anordnung von Attributen durch die Gewichte eines Verfahrens, d.h. die durch die Gewichtung implizierte Reihenfolge in der Wichtigkeit der Attribute. Zum anderen ist auch ein Vergleich der aus verschiedenen Verfahren resultierenden Attributgewichte wünschenswert. Grundsätzlich ist ein solcher Vergleich der von den verschiedenen Verfahren für ein Attribut errechneten Gewichte jedoch problematisch. Dies resultiert zum einen daraus, dass die Verfahren unterschiedliche Wertebereiche für die Gewichte implizieren. Zum anderen folgt dies aus der semantischen Bedeutung der Attributgewichte. Allerdings können ebenfalls aus den bereits erwähnten Rangfolgen der einzelnen Attribute, die sich aus der Höhe der Attributgewichte ergeben, Unterschiede bzw. Übereinstimmungen der angewandten Verfahren in Bezug auf die durch die Verfahren propagierte Wichtigkeit von Attributen gefolgert werden. Exemplarisch wird dies an zwei der zehn betrachteten Analyseaufgaben, d.h. in Bezug auf zwei Zielattribute, ausführlich dargestellt. Aufgrund der besonderen Relevanz der Übergänge in die Arbeitslosigkeit bzw. aus der Arbeitslosigkeit heraus, wurden für diese ausführliche Darstellung die Zielattribute `LFSC_Working_Unemployed` und `LFSC_Unemployed_Working` gewählt.

Um die relative Anordnung der Attributgewichte zu analysieren bzw. bei unterschiedlichen Verfahren zu vergleichen, ist es sinnvoll, die Attributgewichte auf einen einheitlichen Wertebereich zu normalisieren. Eine recht einfache, in RapidMiner zur Verfügung stehende Methode hierzu ist die Transformation

$$w_j^{\text{norm}} = \frac{|w_j| - \min\{|w_1|, \ldots, |w_m|\}}{\max\{|w_1|, \ldots, |w_m|\} - \min\{|w_1|, \ldots, |w_m|\}},$$

wobei w_j die ursprünglichen Attributwerte sind. Hierbei ist $j = 1, \ldots, m$ und m die Anzahl der Attribute. Diese Transformation bewirkt eine Normalisierung der Gewichte auf den Bereich von 0 bis 1, wobei das niedrigste vorkommende Attributgewicht normalisiert dem Wert 0, das höchste dem Wert 1 entspricht. Tabelle 8.1 stellt die anhand dieser Transformation normalisierten Attributgewichte, die aus den angewandten Verfahren für die Lernaufgabe mit dem Label `LFSC_Working_Unemployed` resultieren, grafisch dar. Bezüglich der einzelnen Verfahren ergeben sich folgende Beobachtungen: unter den numerischen Variablen haben die Variablen `YEAR` und `LEGISLATUR` sowie - etwas geringer - `AGE` die höchsten Korrelationen zum Zielattribut. Die mit Abstand niedrigsten Korrelationen zum Label weisen die

Tabelle 8.1: Normalisierte Attributgewichte für `LFSC_Working_Unemployed`. Attributgewichte sind innerhalb eines Verfahrens auf den Wertebereich $[0, 1]$ normalisiert.

Variablenname	Attributgewichte			
	Korrelation	Kontingenz	Infogain	Relief
SEX				
AGE				
YEAR				
GERMBORN				
CORIGIN				
ORTKINDH				
ORTKIND1				
LOC1989				
BULA				
FAMSTD				
TYPHH				
HHGR				
HHCHILDREN				
SUMKIDS				
PSBIL				
PBBIL01				
PBBIL02				
BILZEIT				
NACE				
IS88				
OEFFD				
ERLJOB				
AUSB				
ERWZEIT				
AUTONO				
BETR				
ZYKLUS				
WACHSTUM				
WACHSTUM_L1				
WACHSTUM_L2				
WACHSTUM_L3				
INFLATION				
LEGISLATUR				
WIEDERVEREINIGUNG				

Attribute `ERWZEIT` und `SUMKIDS` auf. Bei den nominalen Attributen besitzt die Variable `NACE` die höchste Kontingenz zum Zielattribut, die Variablen `SEX` und `HHCHILDREN` besitzen die niedrigste. Die Berechnung des Informationsgewinns kann sowohl für nominale als auch numerische Attribute durchgeführt werden. Den höchsten Informationsgewinn in Bezug auf das Zielattribut `LFSC_Working_Unemployed` hat die Variable `IS88`. Weiterhin im Vergleich zu den anderen Variablen recht hohe Werte erreichen die Variablen `LOC1989` und `BULA` sowie `ERWZEIT` und `AUTONO`. Die geringsten Informationsgewinne weisen die Variablen `SEX`, `ORTKINDH` sowie `ORTKIND1`, `HHGR`, `HHCHILDREN`, `SUMKIDS` und die makroökonomische Variable `WACHSTUM_L1` auf. Das Verfahren Relief gewichtet wiederum die Variable `NACE` am höchsten, sowie die ebenfalls auf die Erwerbstätigkeit bezogenen Variablen `OEFFD`, `ERLJOB` und `IS88` recht hoch. Die niedrigsten Gewichte erhalten die Attribute `BILZEIT` und `WACHSTUM_L3`.

Generell fällt auf, dass die Verfahren keine vollkommen einheitliche Reihenfolge der Attributgewichtung liefern, sodass sich kein gänzlich eindeutiges Bild hinsichtlich der Bedeu-

tung von Attributen ergibt. Dennoch bieten die Gewichte erste Anhaltspunkte dafür, welche Einflussgrößen und welche thematischen Aspekte in Bezug auf die Zielattribute erklärenden Einfluss haben könnten. So scheint in Bezug auf den hier betrachteten Übergang von der Erwerbstätigkeit in die Arbeitslosigkeit die Branche bzw. der ausgeübte Beruf, erfasst in den Variablen `NACE` bzw. `IS88`, in allen Verfahren eine wichtige Rolle zu spielen. Bedenkt man, dass viele Berufe häufig bereits eine Tätigkeit in einer bestimmten Branche implizieren und diese Attribute daher vermutlich stark korrelieren, so ist sicherlich ein Einfluss der Komponenten Branche und Beruf abzuleiten. Neben den relativ hohen Gewichten der beiden genannten Attribute deuten jedoch auch andere Attributgewichte eine potentielle Bedeutung von Attributen für die Analyseaufgabe an, etwa wenn mittlere Gewichte für ein Attribut von allen Verfahren gleichermaßen berechnet werden. Als Beispiel hierfür sind etwa die regionalen Indikatoren `LOC1989` sowie `BULA` zu nennen. Auch das Attribut `AUTONO` wird beispielsweise zumindest durch die Gewichtung nach Informationsgewinn und durch Relief als bedeutsam erkannt und ist auch in Bezug auf seine Kontingenz mit dem Zielattribut nicht zu vernachlässigen.

Auch bei der zweiten, hier exemplarisch vorgestellten Analyseaufgabe mit dem Zielattribut `LFSC_Unemployed_Working` beschreiben die Ergebnisse zum einen eine vielfach uneinheitliche Gewichtung der Attribute. Zum anderen lassen sich jedoch auch bei dieser Analyseaufgabe Anzeichen für eine Bedeutsamkeit einzelner Attribute ableiten. Analog zu den vorherigen Betrachtungen sind in Tabelle 8.2 die normalisierten Attributgewich-

Tabelle 8.2: Normalisierte Attributgewichte für `LFSC_Unemployed_Working`. Attributgewichte sind innerhalb eines Verfahrens auf den Wertebereich $[0, 1]$ normalisiert.

Variablenname	Attributgewichte			
	Korrelation	Kontingenz	Infogain	Relief
SEX				
AGE				
YEAR				
GERMBORN				
CORIGIN				
ORTKINDH				
ORTKIND1				
LOC1989				
BULA				
FAMSTD				
TYPHH				
HHGR				
HHCHILDREN				
SUMKIDS				
PSBIL				
PBBIL01				
PBBIL02				
BILZEIT				
ZYKLUS				
WACHSTUM				
WACHSTUM_L1				
WACHSTUM_L2				
WACHSTUM_L3				
INFLATION				
LEGISLATUR				
WIEDERVEREINIGUNG				

te für diese Analyseaufgabe grafisch dargestellt. Wie zuvor ist leicht zu erkennen, dass die Rangfolgen der Attributgewichte hinsichtlich ihrer Höhe, die durch die verschiedenen Verfahren impliziert werden, sehr unterschiedlich sind. Während etwa die Kontingenz des Labels mit dem Attribut `TYPHH` am höchsten ist, besitzt das Attribut `AGE` den mit Abstand höchsten Informationsgewinn. Relief gewichtet die Variable `BULA` am höchsten, erachtet die Variable `AGE` dagegen für mit am wenigsten wichtig. Somit zeigen sich zwar teilweise Widersprüchlichkeiten in der Bestimmung der Bedeutung von Attributen. Allerdings lassen auch hier manche Attributgewichte auf hohe Einflüsse der jeweiligen Attribute schließen. Beispielhaft sind hier etwa die durch die einzelnen Verfahren am höchsten gewichteten Attribute `AGE`, `BULA` und `TYPHH` zu nennen. Obschon die Verfahren sich nicht gänzlich einig in der Bedeutung der Attribute sind, scheinen auch die Attribute `SEX`, `ORTKINDH`, `ORTKIND1` sowie die Bildungsindikatoren `PSBIL`, `PBBIL01` und `PBBIL02` einen nicht zu vernachlässigenden potentiellen Einfluss auf den Übergang von der Arbeitslosigkeit in die Erwerbstätigkeit zu besitzen.

Generell setzt sich auch bei den hier nicht dokumentierten Ergebnissen für die Attributgewichte der anderen Analyseaufgaben auf der einen Seite die partielle Uneindeutigkeit der oben dargestellten Ergebnisse hinsichtlich der Gewichtung der Attribute durch die einzelnen Verfahren fort. Auf der anderen Seite lassen sich jedoch stets Einflüsse solcher Attribute vermuten bzw. erkennen, die von mindestens einem Verfahren sehr hoch gewichtet wurden, oder aber die von allen Verfahren mittlere und damit robuste Gewichte erhielten.

Basierend auf den vorgestellten Ergebnissen und den erfolgten Beobachtungen kann somit zwar nicht eindeutig auf die Wichtigkeit einzelner Attribute geschlossen werden. Allerdings deuten die Gewichte der Attribute häufig einen bestehenden Einfluss einzelner Attribute auf die jeweiligen Zielattribute an. Generell bedeutet dies allerdings auch für solche Attribute, die kein außergewöhnlich hohes Gewicht als Indiz für ihre potentielle Bedeutsamkeit aufweisen, nicht, dass sie vernachlässigbar sind und daher etwa ohne Einschränkung der Analyseaufgabe deselektiert werden können. Vielmehr ist zu beachten, dass die verwendeten Verfahren zur Gewichtung der Attribute stets nur den Einfluss einzelner Attribute auf das Zielattribut erfassen. Interaktionen zwischen Attributen in Bezug auf das Zielattribut können somit nicht erkannt werden. Ein niedriges Gewicht bedeutet daher nicht automatisch, dass es in den geltenden Zusammenhängen keine Rolle spielt. So könnten etwa auch mehrere Attribute, die für sich genommen niedrige Attributgewichte erhielten, in Kombination eine gute Erklärung des Zielattributes ermöglichen. Eine Aufdeckung eines solchen Sachverhalts ist allerdings mit den in diesem Kapitel vorgestellten Verfahren nicht möglich. Hierzu bedarf es der Anwendung von Verfahren, die Interaktionen zwischen Attributen und damit das Erklärungspotential von Attributgruppen berücksichtigen. Eine Anwendung solcher Verfahren wird im nächsten Kapitel beschrieben. Aus dem genannten Grund wurde für diese Anwendung keine Deselektion von Attributen durchgeführt. Dies entspricht im Übrigen auch der Intention einer möglichst offenen Anwendung der tatsächlichen Datenanalyseverfahren zur Bildung von Modellen.

9 Klassifikationslernen

Die wohl verbreitetste Lernaufgabe innerhalb des maschinellen Lernens ist das *Funktionslernen* aus einer Menge gegebener Beispiele. Diese Lernaufgabe besteht darin, einen globalen Zusammenhang zwischen Beschreibungen von Instanzen und Zielwerten für Instanzen zu lernen. Obwohl den Verfahren üblicherweise nur Daten einer Stichprobe vorgelegt werden, soll das erlernte Modell global, also für alle Daten des Instanzenraumes gültig sein. Je nach Art der Zielmenge, der die erwähnten Zielwerte entstammen, definiert man verschiedene Subtypen des Funktionslernens, u.a. das Klassifikationslernen, auf das sich diese Arbeit hauptsächlich konzentriert. Im folgenden Abschnitt 9.1 erfolgt eine formale Definition des Funktionslernens aus Beispielen sowie der einzelnen Subtypen. In den nachfolgenden Abschnitten 9.2 und 9.3 werden anschließend einzelne Verfahren des Klassifikationslernens erläutert. Kapitel 9.4 beschreibt die bei der Anwendung der Verfahren erzielten Ergebnisse.

9.1 Funktionslernen aus Beispielen

Die Eingabe der Lernaufgabe *Funktionslernen aus Beispielen* besteht aus einer Menge von typischerweise als Merkmalsvektoren vorliegenden Instanzenbeschreibungen und einer Menge von beobachteten Zielwerten für die Instanzen. Beim Funktionslernen macht man die Annahme, dass die beobachteten Zielwerte den Instanzen gemäß einer unbekannten Funktion (im Idealfall in Abhängigkeit von den jeweiligen Instanzenbeschreibungen) zugeordnet wurden. Die Aufgabe Funktionslernen besteht darin, eine Funktion zu lernen, die diese unbekannte Funktion möglichst gut approximiert. Formal lässt sich dies folgendermaßen definieren:

Definition 9.1 (Funktionslernen) *Seien X ein Instanzenraum, Y eine Menge möglicher Zielwerte und $\mathcal{L}_H := \mathcal{F}(X, Y) = \{h \mid h : X \to Y\}$ die Menge möglicher Funktionen von X nach Y, der sogenannte Hypothesenraum. Sei $E \subseteq X \times Y$ eine Menge möglicher Beispiele, wobei für eine unbekannte Funktion $f : X \to Y$ gilt, dass $f(x) = y$ für alle $(x, y) \in E$. Sei außerdem* error $: Y \times Y \to \mathbb{R}$ *eine Fehlerfunktion. Ferner sei D eine Wahrscheinlichkeitsverteilung auf X. Die Lernaufgabe* Funktionslernen *ist dann definiert als:*

Gegeben X, Y und E sowie error*. Bestimme die Funktion $h \in \mathcal{L}_H$ mit $h : X \to Y$, für die gilt:*

$$h = \operatorname{argmin}_{h \in \mathcal{L}_H} \operatorname{error}_D(h, f).$$

Hierbei ist

$$\operatorname{error}_D(h, f) := E_D\big[\operatorname{error}\left(f(x), h(x)\right)\big],$$

der erwartete durchschnittliche Fehler für gemäß der Wahrscheinlichkeitsverteilung D aus X gezogene Instanzen. Dieser wird als wahrer Fehler *bezeichnet.*

In Abhängigkeit von der Zielmenge Y werden unterschiedliche Spezialfälle der Lernaufgabe Funktionslernen benannt. Ist Y nominal, nennt man die Lernaufgabe *Klassifikationslernen*. Der Spezialfall des Klassifikationslernens, in dem Y binominal ist, also nur zwei diskrete Werte enthält, wird auch als *Begriffslernen* oder *Konzeptlernen* bezeichnet. Ist Y numerisch, also etwa $Y = \mathbb{R}$, so bezeichnet man die Lernaufgabe als *Regression*.

Im Regelfall kann der wahre Fehler in obiger Definition nicht berechnet werden, da üblicherweise weder f noch D bekannt sind. In diesem Fall müsste der wahre Fehler geschätzt werden. Meist wird zur Auswahl einer Hypothese stattdessen ein Gütemaß auf einer Menge vorliegender Beispiele berechnet und dieses zur Bestimmung der Hypothese h minimiert. Handelt es sich dabei um die gesamte in obiger Definition angesprochene Menge vorliegender Beispiele E, so wird das berechnete Gütemaß als Trainingsfehler einer Hypothese bezeichnet, und es gilt

$$\text{error}_E(h) = \sum_{(x,y)\in E} \text{error}\,(y, h(x)).$$

Die Fehlerfunktion error wird je nach Lernaufgabe gewählt. Beim Klassifikationslernen wird meistens der schon in Gleichung 8.1 definierte 0-1-loss verwendet. Häufig ist es allerdings nicht sinnvoll, den Trainingsfehler als Minimierungskriterium heranzuziehen, da die aus der Minimierung resultierende Hypothese h zwar möglicherweise den Trainingsfehler minimiert, sich damit jedoch den vorliegenden Trainingsdaten zu genau anpasst, auf neuen ungesehenen Daten jedoch schlechte Vorhersageergebnisse liefert. Dieses Phänomen wird als *Überanpassung* (*engl.* Overfitting) bezeichnet. Diese ist schematisch in Abbildung 9.1 dargestellt. Angenommen, die Trainingsmenge enthielte alle Beispiele, die durch die roten

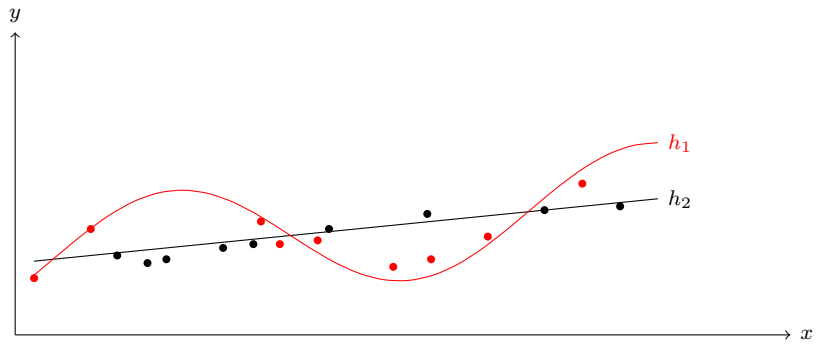


Abbildung 9.1: Phänomen der Überanpassung

Punkte dargestellt sind. Dann ist h_1 eine Hypothese mit nahezu minimalem Trainingsfehler. Sofern die schwarzen Punkte jedoch auch zur Grundgesamtheit gehören und nur nicht zur Trainingsmenge ausgewählt sind, bietet h_2 jedoch sicherlich eine generell bessere Lösung. Somit erfolgt häufig eine Validierung des Fehlers auf einer Testmenge anstatt der Trainingsmenge direkt.

Von der erfolgten Beschreibung ausgehend soll in dieser Arbeit ein global gültiger, funktionaler Zusammenhang zwischen den extrahierten sowie erzeugten Attributen und den einzelnen erzeugten Labeln, den Übergangswechseln zwischen Arbeitsmarktzuständen, gelernt werden. Da es sich bei den Zielattributen um (bi)nominale Attribute handelt, liegt als Lernaufgabe das Begriffs- bzw. Klassfikationslernen vor. Hierbei ist es jedoch nicht so sehr

von Bedeutung, dass ein erlernter Klassifikator in allen Situation gute Vorhersagen macht. Die Minimierung des Vorhersagefehlers auf einer Testmenge steht damit nicht direkt im Fokus. Vielmehr soll auf den Trainingsdaten ein deskriptives Modell gelernt werden, welches die Trainingsdaten möglichst gut beschreibt. Dies bedeutet erstens, dass eine möglichst gute Anpassung des erlernten Modells an die Trainingsdaten erfolgen soll, welche durch den Vorhersagefehler auf den Trainingsdaten gemessen werden kann. Eine Validierung des erlernten Modells auf Testdaten bzw. eine Kreuzvalidierung kann ergänzend sicherstellen, dass eine gewisse Universalität des Modells gegeben ist. Zweitens soll das erlernte Modell möglichst leicht verständlich und umfassend interpretierbar sein, um aus dem Modell Wissen hinsichtlich der realen Wirkungszusammenhänge ableiten zu können. Als diesbezüglich geeignete Verfahren sind *Naïve Bayes* und *Entscheidungsbaumverfahren* anzusehen, die in dieser Arbeit angewandt wurden, und die im Folgenden beschrieben werden.

9.2 Naïve Bayes

Das von Duda und Hart (1973) vorgestellte Verfahren Naïve Bayes versucht, die Wahrscheinlichkeit zu bestimmen, dass eine Untersuchungseinheit, die einen Merkmalsvektor $\mathbf{x} = (x_1, \ldots, x_m)$ für Attribute $X_1 \times \cdots \times X_m$ besitzt, für ein Zielattribut Y den Wert y annimmt. Naïve Bayes beruht dabei auf der "naiven" Annahme, dass alle Attribute stochastisch unabhängig sind. Diese Annahme bedeutet, dass für alle Attributepaare X_i und X_j mit $i, j = 1, \ldots, m$ gilt, dass $\Pr(x_i, x_j) = \Pr(x_i)\Pr(x_j)$, woraus folgt, dass

$$\Pr(x_1, \ldots, x_j) = \prod_{i=1}^{m} \Pr(x_i). \tag{9.1}$$

Basierend auf dieser Annahme gilt nach dem Satz von Bayes für die zu berechnende, bedingte Wahrscheinlichkeit, dass Y den Wert y annimmt unter der Bedingung, dass ein Merkmalsvektor $\mathbf{x} = (x_1, \ldots, x_m)$ beobachtet wurde, dass

$$\begin{aligned} \Pr(y|\mathbf{x}) &= \frac{\Pr(y) \cdot \Pr(\mathbf{x}|y)}{\Pr(\mathbf{x})} \\ &= \frac{\Pr(y) \cdot \Pr(x_1, \ldots, x_m|y)}{\Pr(x_1, \ldots, x_m)} \\ &\overset{(9.1)}{=} \frac{\Pr(y) \cdot \prod_{i=1}^{m} \Pr(x_i|y)}{\Pr(x_1, \ldots, x_m)}. \end{aligned} \tag{9.2}$$

Hierbei ist $P(y)$ die A-Priori-Wahrscheinlichkeit, dass Untersuchungseinheiten das Label y annehmen, $P(x_1, \ldots, x_m)$ die A-Priori-Wahrscheinlichkeit, dass für eine Untersuchungseinheit die Merkmale $x_1, \ldots, x_m$ beobachtet werden, sowie $P(x_1, \ldots, x_m|y)$ die bedingte Wahrscheinlichkeit, dass eine Untersuchungseinheit, für die bekannt ist, dass sie als Label den Wert y hat, die Merkmale $x_1, \ldots, x_m$ aufweist. Im Lernschritt schätzt Naïve Bayes genau diese Wahrscheinlichkeiten durch Berechnung der jeweiligen relativen Häufigkeiten auf den vorgelegten Daten. Zur Klassifikation wird dann die Funktion

$$\begin{aligned} h : X_1 \times \cdots \times X_m &\to Y \\ (x_1, \ldots, x_m) &\mapsto \operatorname{argmax}_{y \in Y} \Pr(y|x_1, \ldots, x_n) \end{aligned}$$

unter Berücksichtigung von Gleichung (9.2) und den geschätzten Wahrscheinlichkeiten verwendet.

Im Fall, dass die Unabhängigkeit der Attribute vorliegt, ist Naïve Bayes theoretisch das beste Lernverfahren. Im Regelfall gilt die von Naïve Bayes instrumentalisierte Annahme der Unabhängigkeit von Attributen jedoch nicht. Nichtsdestotrotz bietet Naïve Bayes häufig relativ gute Lernergebnisse (vgl. Domingos und Pazzani (1996)). Im Rahmen dieser Arbeit ist darüber hinaus nicht die Vorhersage-Performanz des Verfahrens entscheidend. Vielmehr kann Naïve Bayes als Erweiterung der Verfahren aus Kapitel 8 dazu dienen, Zusammenhänge zwischen dem Label und einzelnen Attributen aufzudecken, indem auf Basis der oben berechneten Wahrscheinlichkeiten die Wahrscheinlichkeiten $P(y|x_i)$ für die möglichen Werte der Attribute X_i (mit $i = 1, \ldots, m$) berechnet werden.

9.3 Entscheidungsbäume

Verfahren zur Erzeugung von *Entscheidungsbäumen* (*engl.* decision trees) als Modell haben eine lange Tradition innerhalb des maschinellen Lernens. So wurden viele Varianten von Entscheidungsbaumlernern entwickelt, etwa CART von Breiman et al. (1984), ID3 von Quinlan (1986) oder dessen Nachfolger C4.5 ebenfalls von Quinlan (1993). Die genannten Systeme basieren jedoch alle auf dem gleichen algorithmischen Konzept, der rekursiven Erzeugung einer hierarchischen Partitionierung der Beispielmenge E anhand von Attributwerten.

Die grundsätzliche Vorgehensweise ist dabei die folgende: Anhand einer noch zu spezifizierenden Heuristik wird ein Attribut X_i (mit $i \in \{1, \ldots, m\}$) ausgewählt und die vorliegende Beispielmenge E anhand der einzelnen Werte v_j (mit $j \in \{1, \ldots, |X_i|\}$) des Attributes X_i partitioniert, d.h. in Teilmengen $E_j := \{(x_1, \ldots, x_m, y) \in E \mid x_i = v_j\}$ unterteilt, sodass für jeden Attributwert des ausgewählten Attributes eine Teilmenge von Beispielen, die genau diesen Attributwert gemeinsam haben, existiert. Im Fall, dass für eine der resultierenden Teilmengen der Partition alle Beispiele das gleiche Label haben, braucht diese Teilmenge nicht weiter unterteilt werden. Die Teilmenge bildet ein Blatt im Baum mit dem Label, welches die in der Teilmenge enthaltenen Beispiele gemeinsam haben. Im entgegengesetzten Fall bildet die Teilmenge einen inneren Knoten im Baum, der einen Split anhand der Werte eines Attributes repräsentiert. Es erfolgt dann rekursiv eine weitere Partition dieser Teilmenge nach dem soeben beschriebenen Verfahren. Der vollständige Algorithmus, wie er beispielsweise von Quinlan (1993) erläutert wird, ist in Abbildung 9.2 dargestellt. Neben der Rekursivität ist die heuristische Bestimmung des Attributes, welches die hoffentlich beste Aufspaltung der Beispiele in möglichst reine Teilmengen in Bezug auf die Labelwerte bietet, zentraler Bestandteil des Algorithmus. Diese ist im Algorithmus in der Funktion BestimmeBestesAttribut gekapselt. Zur Durchführung dieser Bestimmung wurden einige Heuristiken vorgeschlagen. Eine dieser Heuristiken verwendet das in Definition 8.4 eingeführte Maß des Informationsgewinns und wählt das Attribut mit dem höchsten Informationsgewinn. Alternativ kann die *Gain-Ratio-Heuristik* verwendet werden, die durch

$$\mathrm{GR}_Y(X) = \frac{\mathrm{IG}_Y(X)}{H(X)}.$$

gegeben ist. Sie setzt den Informationsgewinn eines Attributes ins Verhältnis zur Entropie (siehe Definition 8.3) desselben Attributes. Einen breiten Überblick über diese und weitere heuristische Maße zur Selektion eines guten Attributes in Bezug auf die Partitionierung der

Eingabe : Menge von Beispielen E, Testheuristik h
Ausgabe: Entscheidungsbaum T

```
1  function InduziereEntscheidungsbaum(E, h)
2      T ← ErzeugeKnoten(E)
3      if EnthältUnterschiedlicheLabel(E) then
4          i ← BestimmeBestesAttribut(E, h)
5          foreach v_1, ..., v_k ∈ X_i do
6              E_j ← {(x_1, ..., x_m, y) ∈ E | x_i = v_j}
7              T_j ← InduziereEntscheidungsbaum(E_j, h)
8              FügeTeilbaumHinzu(T, T_j)
9          end
10     else
11         MarkiereAlsBlatt(T, E)
12     end
13 return T
```

Abbildung 9.2: Rekursiver Algorithmus zum Aufbau eines Entscheidungsbaumes

Beispiele geben Murthy (1998) und Rokach und Maimon (2005). Dort sind des Weiteren auch eine Übersicht über Abbruchkriterien des rekursiven Prozesses zu finden, die alternativ zur Forderung, dass ein Blatt nur Beispiele einer Klasse enthalten darf, genutzt werden können. Eines dieser Abbruchkriterien ist etwa das Erreichen einer vorgegebenen Maximaltiefe des Baumes. Außerdem resümieren sowohl Murthy (1998) als auch Rokach und Maimon (2005) verschiedene Verfahren, um den Baum zu beschneiden (*engl.* Pruning), um eine Überanpassung an die Trainingsdaten und damit eine höhere Generalität und prädiktive Performanz zu erreichen, und sie erläutern die Behandlung fehlender Werte beim Aufbau des Entscheidungsbaumes.

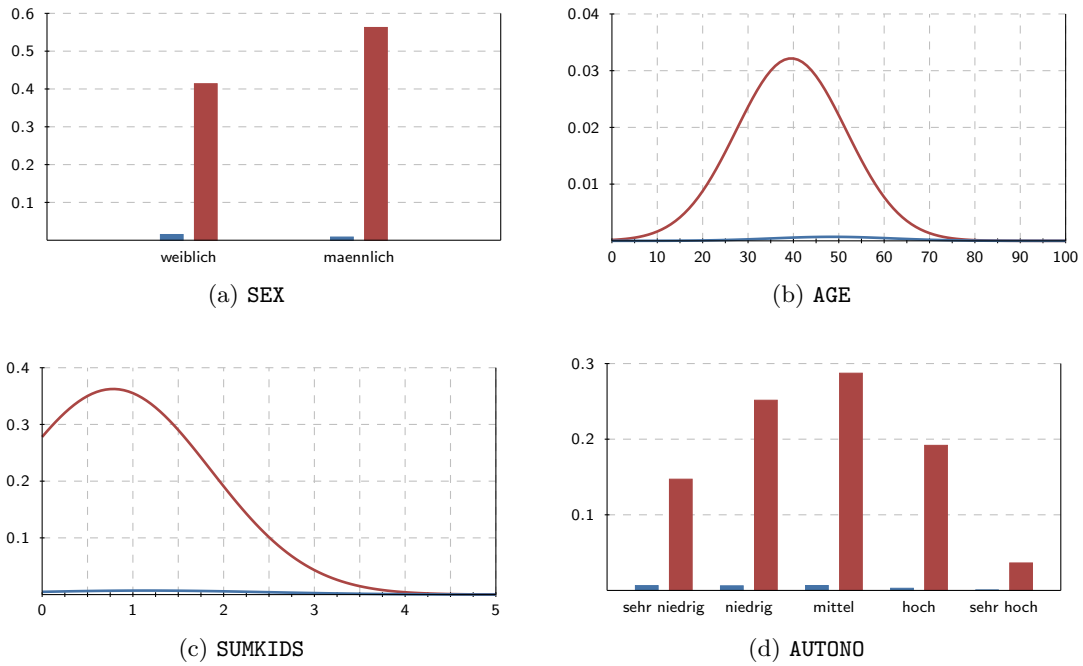
Die Klassifikation eines Beispiels erfolgt durch Traversierung des Baumes, wobei bei der Wurzel begonnen wird. Ist der aktuelle Knoten ein innerer Knoten und stellt damit eine Aufteilung anhand von Werten eines Attributes dar, wird in den Teilbaum übergegangen, der bzgl. des Attributes den gleichen Wert repräsentiert wie das zu klassifizierende Beispiel. Wird ein Blattknoten erreicht, so wird das Beispiel mit dem Label, mit dem das Blatt markiert ist, klassifiziert.

9.4 Ergebnisse

Im Rahmen dieser Arbeit wurden sowohl das Verfahren Naïve Bayes als auch ein Entscheidungsbaumlernverfahren für die vorgestellten Lernaufgaben auf den Daten angewandt. Von den Verfahren gelernte Modelle werden im Folgenden exemplarisch vorgestellt, bevor im Anschluss daran eine kurze Evaluation der resultierenden Ergebnisse in Bezug auf die Performanz der Verfahren erfolgt.

Nach Kapitel 9.2 extrahiert Naïve Bayes aus den Daten als Modell vordergründig die bedingten Wahrscheinlichkeitsverteilungen von Attributwerten unter der Bedingung des Vorliegens der einzelnen Zielwerte, d.h. das Verfahren berechnet für alle Klassen (d.h. alle $y \in$

Y) die Wahrscheinlichkeit $\Pr(X = x|Y = y)$. Für numerische Attribute X wird dabei häufig eine Normalverteilung der Attributwerte unterstellt. Auch der in dieser Arbeit verwendete Operator `NaiveBayes` in RapidMiner verfährt so. In Anlehnung an die Darstellungsweise des vom Operator ausgegebenen Modells innerhalb der Software RapidMiner erfolgt auch die Darstellung der erlernten Modelle in dieser Arbeit. Dabei werden die erwähnten bedingten Wahrscheinlichkeiten noch mit der A-Priori-Wahrscheinlichkeit $\Pr(Y = y)$ gewichtet. Abbildung 9.3 zeigt exemplarisch für vier Attribute die Ergebnisse der Anwendung



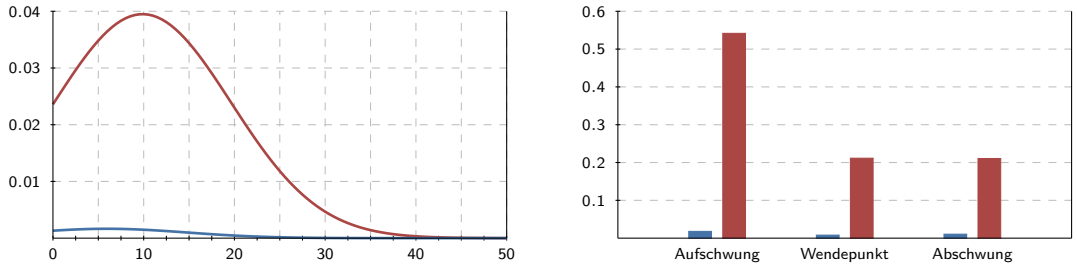
(a) `SEX` (b) `AGE`

(c) `SUMKIDS` (d) `AUTONO`

Abbildung 9.3: Ausgewählte Ergebnisse von Naïve Bayes für `LFSC_Working_Not-working`. Bedingte Wahrscheinlichkeiten $\Pr(X = x|Y = y)$ gewichtet mit A-priori-Klassenwahrscheinlichkeiten $\Pr(Y = y)$. Bei nominalen Attributen Abbildung der fünf am häufigsten vorkommenden Attributwerte. rot: negative Klasse, blau: positive Klasse

von Naïve Bayes bezüglich der Analyseaufgabe, die den Übergang von erwerbstätig nach nichterwerbstätig, erfasst durch die Zielvariable `LFSC_Working_Not-working`, untersucht. Dabei werden die auffälligsten und interessantesten Sachverhalte, die in den Modellen beschrieben sind, aufgezeigt. Interessant bedeutet im Fall von Naïve Bayes, dass die bedingten Verteilungen der Attributwerte für die positive Klasse und die negative Klasse möglichst weit differieren. Dann ergeben sich Unterschiede hinsichtlich der Eigenschaften von Untersuchungseinheiten[1], die den durch die Zielvariable erfassten Wechsel aufweisen (positive Klasse) und denen, die dies nicht tun (negative Klasse). Dies lässt unter Umständen auf einen (möglicherweise erklärenden) Zusammenhang zwischen den angenommen Attributausprägungen und den jeweils angenommenen Werten des Labels schließen. Aus Abbildung 9.3 sind diesbezüglich folgende Sachverhalte zu beobachten: Untersuchungseinheiten, die vom Zustand erwerbstätig in den Zustand nichterwerbstätig wechseln, sind größtenteils

[1]Man beachte, dass die Untersuchungseinheit aufgrund der Transformation aus Kapitel 6.5 nicht Person sondern Personen-Jahr, d.h. Person in einem Jahr, ist.

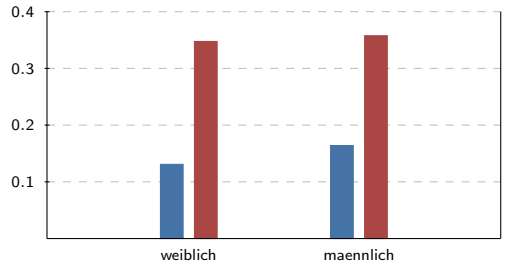


(a) ERWZEIT (b) ZYKLUS

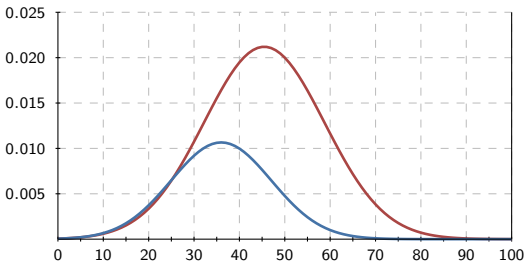
Abbildung 9.4: Ausgewählte Ergebnisse von Naïve Bayes für LFSC_Working_Unemployed. Bedingte Wahrscheinlichkeiten $\Pr(X = x|Y = y)$ gewichtet mit A-priori-Klassenwahrscheinlichkeiten $\Pr(Y = y)$. rot: negative Klasse, blau: positive Klasse

weiblich. Bei Untersuchungseinheiten, die diesen Übergang nicht machen, ist dies umgekehrt und die Eigenschaft *maennlich* am häufigsten vorkommend. Somit ist zu schlussfolgern, dass unter Frauen die Tendenz, in die Nichterwerbstätigkeit überzugehen größer ist als bei Männern. Auch bezüglich des Alters ergibt sich ein recht deutlicher Zusammenhang. Der Mittelwert des Alters von Untersuchungseinheiten, die nicht aus der Erwerbstätigkeit in die Nichterwerbstätigkeit übergehen liegt bei etwa 39,5 Jahren, der von Untersuchungseinheiten, die diesen Übergang aufweisen hingegen bei etwa 48,5 Jahren. Eine weitere Auffälligkeit besteht hinsichtlich der Anzahl leiblicher Kinder, erfasst durch die Variable SUMKIDS: Bei Untersuchungseinheiten, die den angegebenen Übergang nicht durchführen, liegt der Mittelwert der Variable SUMKIDS bei etwa 0,783, bei Untersuchungseinheiten, die den Übergang durchführen jedoch bei 1,146. Dies kann derart gedeutet werden, dass die Wahrscheinlichkeit, in den Zustand nichterwerbstätig überzugehen, mit zunehmender Anzahl leiblicher Kinder steigt. Weitere Differenzen in den bedingten Verteilungen ergeben sich bei dem Attribut AUTONO. Während bei den Untersuchungseinheiten ohne den spezifizierten Übergang die Wahrscheinlichkeiten für das Vorliegen einer beruflichen Tätigkeit mit mittlerer, niedriger und sehr niedriger Autonomie unterschiedlich sind, sind die Wahrscheinlichkeiten in der Gruppe der Untersuchungseinheiten mit Übergang in etwa gleich groß. Folgern hieraus ließe sich, dass eine niedrige berufliche Autonomie die Entscheidung eines Überganges in die Nichterwerbstätigkeit positiv beeinflusst. Die restlichen, nicht dargestellten Attribute weisen im Wesentlichen keine besonderen Auffälligkeiten in den bedingten Verteilungen auf, sie sind daher nicht dargestellt.

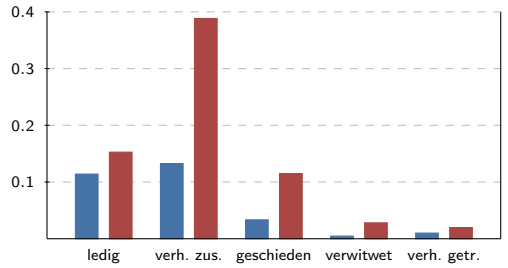
Hinsichtlich des Überganges von der Erwerbstätigkeit in die Arbeitslosigkeit (erfasst durch LFSC_Working_Unemployed) sind ebenfalls nur wenige Auffälligkeiten in den Attributverteilungen von Untersuchungseinheiten mit bzw. ohne Übergang vorhanden. Die größte Auffälligkeit zeigt sich bei der Variable ERWZEIT, die die Dauer der Betriebszugehörigkeit von Untersuchungspersonen in Jahren erfasst. Diesbezüglich ist zu verzeichnen, dass der Mittelwert dieser Dauer bei Untersuchungseinheiten, die arbeitslos werden, nur etwa 5,952 Jahre beträgt gegenüber einem Mittelwert bei Untersuchungseinheiten ohne diesen Übergang von etwa 9,879 Jahren. Somit wird die intuitive Annahme belegt, dass im Regelfall zunächst die Personen entlassen werden, die als letztes zum Betrieb gestoßen sind. Die beiden zum Attribut ERWZEIT gehörigen bedingten Verteilungen sind in Abbildung 9.4(a) dargestellt. Durch die bedingten Verteilungen der Werte der Variable ZYKLUS wird zudem



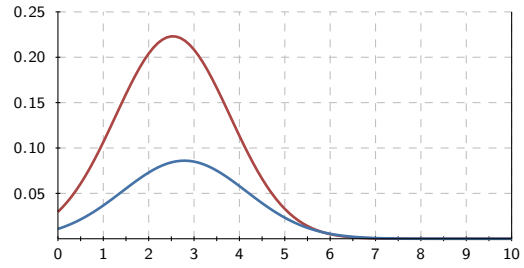
(a) SEX



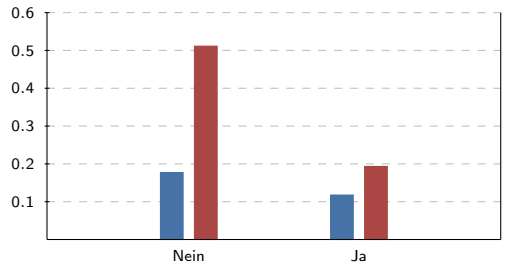
(b) AGE



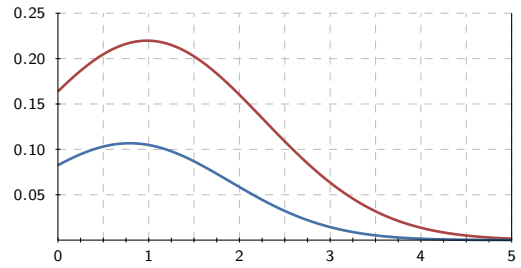
(c) FAMSTD



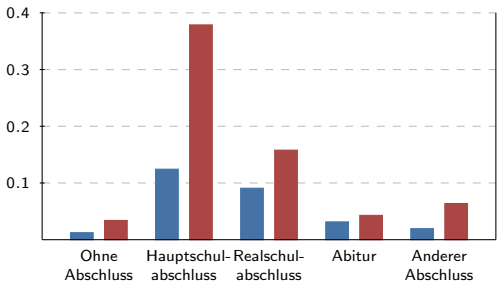
(d) HHGR



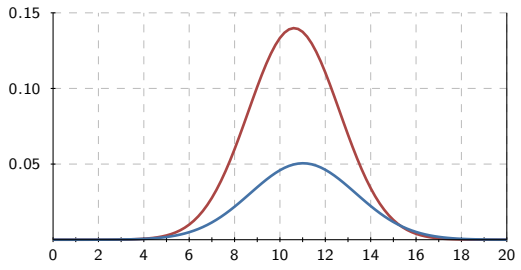
(e) HHCHILDREN



(f) SUMKIDS



(g) PSBIL



(h) BILZEIT

Abbildung 9.5: Ausgewählte Ergebnisse von Naïve Bayes für LFSC_Unemployed_Working. Bedingte Wahrscheinlichkeiten $\Pr(X = x|Y = y)$ gewichtet mit A-priori-Klassenwahrscheinlichkeiten $\Pr(Y = y)$. Bei nominalen Attributen Abbildung der fünf am häufigsten vorkommenden Attributwerte. rot: negative Klasse, blau: positive Klasse

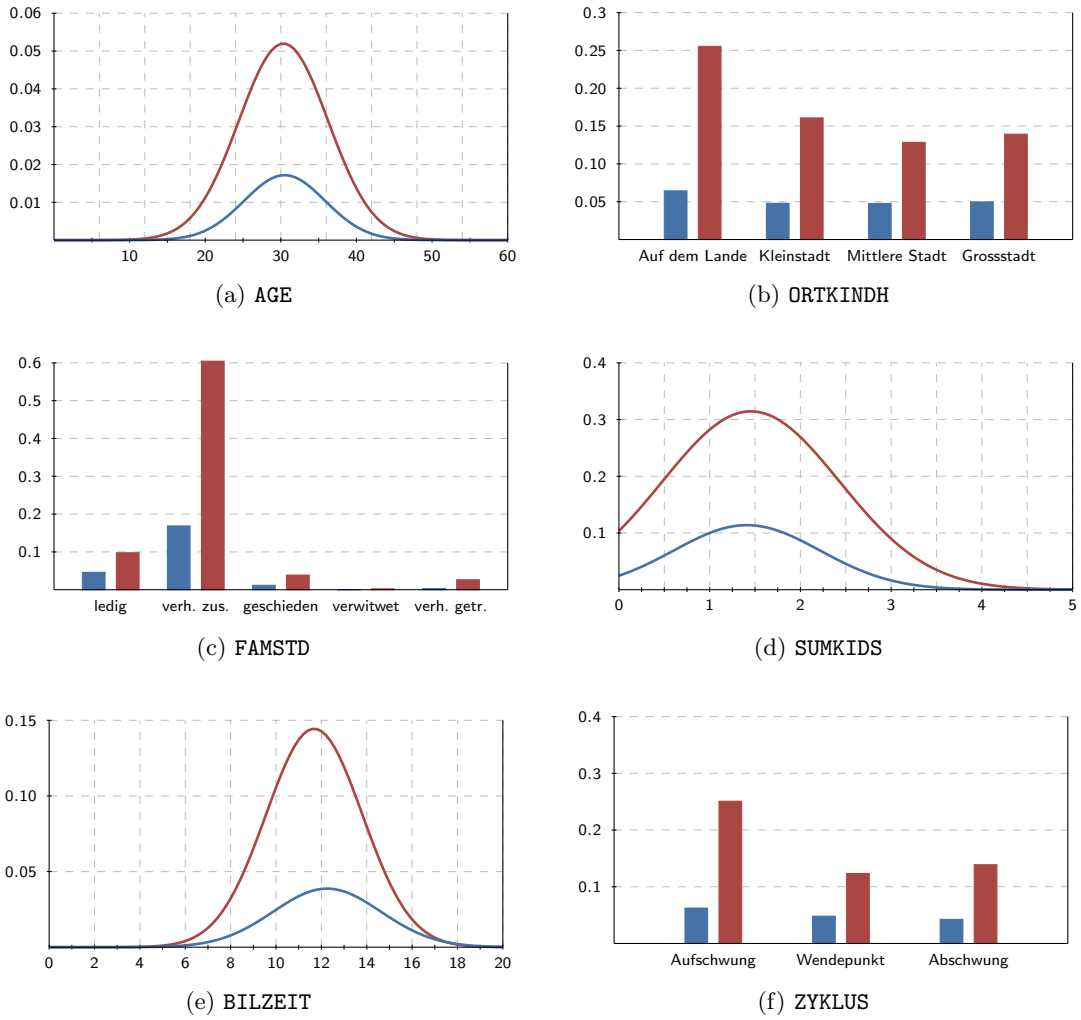
die gängige, bereits aus Abbildung 7.2(b) abgeleitete Vermutung des Einflusses der Konjunkturphase gestützt. Dies zeigt sich in den Verhältnissen aus den Wahrscheinlichkeiten des Vorliegens eines Abschwungs bzw. eines Aufschwungs bei Untersuchungseinheiten mit Übergang in die Arbeitslosigkeit im Vergleich zu denen ohne diesen Übergang.

Im Gegensatz zu den vorher gezeigten Ergebnissen profitieren die Ergebnisdarstellungen für den Übergang von arbeitslos nach erwerbstätig durch die hohe A-priori-Wahrscheinlichkeit der positiven Klasse, wie in Abbildung 9.5 leicht zu erkennen ist. Dennoch sind auch die bedingten Verteilungen für sich genommen bei manchen Attributen in dieser Lernaufgabe aufschlussreicher als bei anderen Übergängen. Ein Beispiel hierfür ist das Attribut `SEX`. Während die Eigenschaften *maennlich* und *weiblich* in den Untersuchungseinheiten ohne Übergang von arbeitslos nach erwerbstätig mit 50,7% zu 49,3% in etwa gleich verteilt sind, beträgt das Verhältnis bei Untersuchungseinheiten mit diesem Übergang etwa 55,6% zu 44,4%. Weitere, recht deutliche Schlüsse erlauben die Ergebnisse in Bezug auf das Attribut `AGE`. Liegt der Mittelwert von Untersuchungseinheiten ohne Übergang bei über 45 Jahren, so liegt er bei Untersuchungseinheiten mit Übergang nur bei knapp 36 Jahren. Dies unterstützt die gängige These, dass es für Personen höheren Alters schwieriger ist, bei Vorliegen von Arbeitslosigkeit wieder einen Arbeitsplatz zu finden. Weitere Auffälligkeiten ergeben sich etwa dahingehend, dass Untersuchungseinheiten mit Übergang von der Arbeitslosigkeit in die Erwerbstätigkeit einen vergleichsweise hohen Anteil von Untersuchungseinheiten mit der Eigenschaft *ledig*, eine höhere mittlere Haushaltsgröße, einen relativ hohen Anteil von Untersuchungseinheiten mit Kindern im Haushalt sowie einen etwas geringeren Mittelwert der Anzahl der leiblichen Kindern haben. Ferner lässt sich diese Gruppe von Untersuchungseinheiten als vergleichsweise gebildeter charakterisieren, was aus den bedingten Werteverteilungen der Attribute `PSBIL` und `BILZEIT` hervorgeht.

Als letzte exemplarisch hier vorgestellte Ergebnisse von Naïve Bayes sind in Abbildung 9.6 einige Ergebnisse der Lernaufgabe mit dem Label `LFSC_Parenthood_Working` aufgezeigt. Aus diesen ist etwa zu erkennen, dass das Alter bei der Entscheidung, wieder in die Erwerbstätigkeit überzugehen, kaum eine Rolle zu spielen scheint. Erstaunlicherweise gilt dies auch für die Anzahl der leiblichen Kinder, deren Mittelwert für Untersuchungseinheiten mit bzw. ohne Übergang nahezu identisch ist. Einen positiven Einfluss auf den Übergang scheint allerdings beispielsweise die Eigenschaft *ledig* als Familienstand von Untersuchungseinheiten zu haben. Dies ist evtl. dadurch begründet, dass ledige Personen, die zu einem Großteil alleinerziehend sein dürften, selbst für den Lebensunterhalt sorgen müssen. Ein positiver Effekt auf den Übergang scheint weiterhin von einem höheren Bildungsniveau auszugehen. Dies wird durch die Ergebnisse bzgl. der Variable `BILZEIT` nahegelegt. Unter Umständen interessanter - jedoch in ihrer Deutung schwieriger sowie weniger eindeutig - sind die Ergebnisse bzgl. der Variablen `ORTKINDH` bzw. `ZYKLUS`. So sind unter den Untersuchungseinheiten mit Übergang beispielsweise diejenigen, die ihre Kindheit in einer mittleren Stadt verbracht haben, anteilsmäßig überproportional vertreten. Eine Begründung dessen ohne weitere Informationen wäre eher spekulativ und unterbleibt daher. Ähnliches gilt für die Ergebnisse bzgl. des Attributes `ZYKLUS`, die innerhalb der Untersuchungseinheiten mit Übergang nahezu eine Gleichverteilung der Konjunkturphasen beschreiben - im Gegensatz zur Verteilung innerhalb der Untersuchungseinheiten ohne Übergang (vgl. Abbildung 9.6(f)).

Als zweites Lernverfahren wurde der in RapidMiner verfügbare Entscheidungsbaumlerner `DecisionTree` zur Lösung der Lernaufgaben angewendet. Im Folgenden werden die Ergebnisse dessen - ebenfalls exemplarisch - beschrieben.

Der Entscheidungsbaum, der Ergebnis der Anwendung des Lerners für die Lernaufgabe



(a) `AGE` (b) `ORTKINDH`

(c) `FAMSTD` (d) `SUMKIDS`

(e) `BILZEIT` (f) `ZYKLUS`

Abbildung 9.6: Ausgewählte Ergebnisse von Naïve Bayes für `LFSC_Parenthood_Working`. Bedingte Wahrscheinlichkeiten $\Pr(X = x|Y = y)$ gewichtet mit A-priori-Klassenwahrscheinlichkeiten $\Pr(Y = y)$. Bei nominalen Attributen Abbildung der fünf am häufigsten vorkommenden Attributwerte. rot: negative Klasse, blau: positive Klasse

mit dem Label `LFSC_Working_Not-working` war, ist in Abbildung 9.7 dargestellt. Dieser fällt durch seine Simplizität auf: er besteht aus lediglich einem inneren Knoten, der die

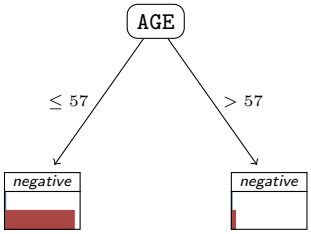


Abbildung 9.7: Entscheidungsbaum für `LFSC_Working_Not-working`

Beispielmenge hinsichtlich des Alters (durch das Attribute `AGE`) in den Fälle jünger als 58 bzw. älter als 57 Jahre aufspaltet. Trotzdem ist die Vorhersage, die durch die beiden resultierenden Blattknoten für das Label getroffen wird, in beiden Fällen *negative*. Somit kann dieses durch den Baum repräsentierte Modell keine bessere Vorhersageperformanz erreichen, als durch die einfache Vorhersage der Default-Klasse *negative* erreicht würde. Doch nicht nur die prädiktive Performanz ist eher ernüchternd, auch die deskriptive Aussagekraft ist gering. So ist dem Entscheidungsbaum lediglich zu entnehmen, dass das Attribut `AGE` einen hohen Informationsgewinn, der als Kriterium für die Selektion von Attributen zum Erzeugen der Beispielpartitionen für den Baumlerner verwendet wurde, hat. Lediglich die Altersgrenze, anhand der die Beispielmenge aufgespalten wurde, erscheint intuitiv bedeutsam für die Lernaufgabe.

Auch für die Lernaufgabe mit dem Zielattribut `LFSC_Working_Unemployed` liefert der Entscheidungsbaumlerner keine verwertbaren Ergebnisse in Form eines aussagekräftigen deskriptiven Modells. Wie in Abbildung 9.8 zu erkennen ist, besteht dieser Baum lediglich

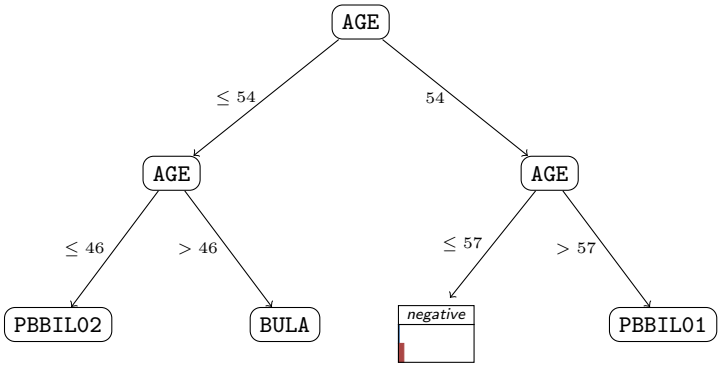
Abbildung 9.8: Entscheidungsbaum für `LFSC_Working_Unemployed`

aus einem Wurzelknoten, der damit gleichzeitig einziges Blatt ist und die häufigste in der gesamten Beispielmenge vorkommende Klasse *negative*, also keinen Übergang, vorhersagt. Damit ist weder eine über die Vorhersage der häufigsten Klasse hinausgehende prädiktive Performanz gegeben, noch ein deskriptives Modell, welches Zusammenhänge in den Daten jenseits der Nennung der häufigsten Klasse beschreibt.

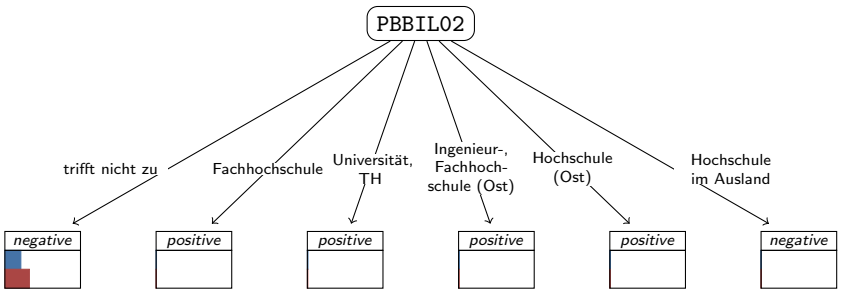
Tatsächlich war das Lernen eines aussagekräftigen Entscheidungsbaumes, der ein befriedigendes deskriptives Modell bei gleichzeitig akzeptabler Vorhersageperformanz darstellt, für viele der betrachteten Lernaufgaben nicht möglich. Hiervon betroffen waren im Wesentlichen die Lernaufgaben, bei denen der Anteil der negativen Klasse als der am häufigsten vorkommenden Klasse bereits sehr hoch war. Bei diesen Lernaufgaben war der Entscheidungsbaumlerner auch bei einer Parameterwahl, die ein Pruning möglichst verhindern sollte, nicht in der Lage, ein aussagekräftiges Modell zu erzeugen. Im Einzelnen war dies bei den Lernaufgaben bzgl. der Zielattribute `LFSC_Jobbing_Not-working`, `LFSC_Unemployed_Jobbing`, `LFSC_Training_Jobbing` und `LFSC_Parenthood_Jobbing` der Fall.

Im Gegensatz zu den vorherig dargestellten Modellen stellt der in Abbildung 9.9 partiell

gezeigte Entscheidungsbaum für die Lernaufgabe zum Wechsel von der Arbeitslosigkeit in die Erwerbstätigkeit ein komplexeres Modell dar, in dem die Beispielmenge anhand des Entscheidungsbaumes auf mehreren Ebenen aufgespalten wird und dann eine dedizierte



(a) Oberer Teil des Baumes



(b) Teilbaum PBBIL02

Abbildung 9.9: Entscheidungsbaum für LFSC_Unemployed_Working. Darstellung des oberen Teils des Baumes sowie (beispielhaft) eines nachfolgenden Teilbaumes.

Vorhersage hinsichtlich des Überganges getroffen wird. Die vordergründige Aufspaltung nach dem Alter der Untersuchungseinheiten durch die Knoten auf den ersten beiden Ebenen bestätigt den hohen Einfluss des Alters auf den Wechsel, den schon die Ergebnisse von Naïve Bayes vermuten ließen. Des Weiteren postuliert der Baum bei den Untersuchungseinheiten mit einem Höchstalter von 46 Jahren einen positiven Einfluss der Hochschulbildung (mit Ausnahme einer solchen im Ausland) auf den Übergang in die Erwerbstätigkeit (vgl. den in Abbildung 9.9(b) beispielhaft dargestellten Teilbaum). Allerdings ist zu beachten, dass mit dem Split anhand des Attributes PBBIL02 nur kleine Gruppen von Untersuchungseinheiten abgespalten und relativ eindeutig klassifiziert werden. Die große Gruppe der Untersuchungseinheiten, die keinen Hochschulabschluss besitzen, wird nicht weiter betrachtet und als *negative* klassifiziert. Weitere Aufspaltungen erfolgen beispielsweise anhand der Attribute BULA und PBBIL01. Grundsätzlich ist jedoch keines der Attribute in der Lage, anhand seiner Werte die jeweiligen Teilmengen der Beispiele in solche Mengen zu zerlegen, die bis auf einen zu vernachlässigend kleinen Anteil nur Untersuchungseinheiten einer Klasse enthalten, gleichzeitig aber auch noch eine relevante Größe und damit Abdeckung der betrachteten Daten gewährleisten. So stellen Blattknoten mit einer größeren Abdeckung meist zwei relativ große Teilmengen an positiven und negativen Beispielen dar, wobei der Anteil der negativen Beispiele in solchen Knoten stets überwiegt. Zerlegt der Entscheidungsbaumlerner in feiner gegliederte Beispielmengen, so sind diese zwar häufig eindeutiger einer

Klasse zuzuordnen, allerdings enthalten diese Mengen dann meist nur noch eine eher unbedeutende Zahl von Beispielen. Somit bietet das erlernte Modell entweder wenig präzise Charakterisierungen oder aber einigermaßen präzise, dafür aber wenig allgemeine Charakterisierungen. Diese präzisen Charakterisierungen muten zudem häufig beinahe willkürlich an, was insofern nicht verwunderlich ist, als dass sie nur einen sehr kleinen Teil der Population, d.h. nahezu Einzelfälle, beschreiben. Letztlich ist daraus zu folgern, dass der Entscheidungsbaumlerner nicht in der Lage ist, ein nicht zu triviales aber auch nicht zu detailliertes Modell zu erzeugen und damit einen sinnvollen Kompromiss der Komponenten Allgemeingültigkeit, Komplexität und Verständlichkeit des Modells, die schlussendlich die Akzeptanz des Modells bestimmen, zu finden. Dies gilt im Übrigen nicht nur für das hier exemplarisch vorgestellte Modell für die Lernaufgabe mit dem Label `LFSC_Unemployed_Working`, sondern im Wesentlichen auch für die weiteren, bislang in diesem Abschnitt nicht genannten Lernaufgaben mit den Zielattributen `LFSC_Jobbing_Unemployed`, `LFSC_Training_Working` und `LFSC_Parenthood_Working`.

Wurden die erlernten Klassifikationsmodelle bislang im Wesentlichen hinsichtlich ihrer deskriptiven Aussagen und ihrer Aussagekraft betrachtet, so folgt nun eine Untersuchung der prädiktiven Güte der Modelle. Eines der gebräuchlichsten Performanzmaße im Bereich des maschinellen Lernens ist die *Accuracy*, die den Anteil der durch das Modell richtig klassifizierten Beispiele angibt. Gängiger Konsens zur Bewertung der prädiktiven Performanz von Modellen ist die Messung der durchschnittlichen Accuracy bei einer - üblicherweise zehnfachen - Kreuzvalidierung. Nichtsdestotrotz zeigt Tabelle 9.1 zunächst die Performanz des Modells auf den Trainingsdaten, also den Daten, die auch zum Lernen des Modells verwendet wurden. Bereits hier ist zu erkennen, dass sowohl Naïve Bayes als auch der Entscheidungsbaumlerner selbst auf den Trainingsdaten häufig nicht in der Lage sind, genauere und damit bessere Vorhersagen zu machen als es die einfache Vorhersage der häufigsten Klasse machen würde. Dies ist vor allem der Fall bei den Lernaufgaben, die einen sehr hohen Anteil der häufigsten Klasse haben. Die Modelle schaffen es bei diesen Aufgaben nicht, die positive Klasse zu isolieren und im Modell zu beschreiben. Lediglich bei den Lernaufgaben mit den Zielattributen `LFSC_Jobbing_Unemployed`, `LFSC_Unemployed_Working`, `LFSC_Training_Working` sowie `LFSC_Parenthood_Working` können leichte Verbesserungen gegenüber der Vorhersage der Default-Klasse verzeichnet werden. Dieses Bild zeigt sich im Wesentlichen auch bei der Betrachtung der Performanzmaße, die bei einer zehnfachen

Tabelle 9.1: Trainingsperformanz von Naïve Bayes und Entscheidungsbaumlerner. Accuracy auf Trainingsdaten. Relative Häufigkeit der Default-Klasse als Referenz.

Label	Default	Naïve Bayes	Entscheidungsbaum
`LFSC_Working_Not-working`	97,415%	97,415%	97,415%
`LFSC_Working_Unemployed`	96,757%	96,757%	96,757%
`LFSC_Jobbing_Not-working`	88,386%	88,486%	88,386%
`LFSC_Jobbing_Unemployed`	96,391%	96,406%	96,765%
`LFSC_Unemployed_Working`	71,956%	72,046%	73,209%
`LFSC_Unemployed_Jobbing`	96,165%	96,165%	96,165%
`LFSC_Training_Working`	72,470%	72,515%	73,543%
`LFSC_Training_Jobbing`	87,523%	87,539%	87,523%
`LFSC_Parenthood_Working`	77,467%	77,613%	79,487%
`LFSC_Parenthood_Jobbing`	92,956%	92,956%	92,956%

Kreuzvalidierung unter Einsatz der beschriebenen Lernverfahren berechnet wurden. Diese sind in Tabelle 9.2 dargestellt. Die Performanzen entsprechen dabei im Regelfall in etwa der relativen Häufigkeit der Default-Klasse. Auch signifikante bzw. überhaupt nennenswerte Unterschiede zwischen den Performanzen der beiden Verfahren Naïve Bayes und Entscheidungsbaumlerner existieren nicht.

Tabelle 9.2: Klassifikationsperformanz von Naïve Bayes und Entscheidungsbaumlerner. Durchschnittliche Accuracy und Standardabweichung der Accuracy bei einer 10-fachen Kreuzvalidierung auf Trainingsdaten. Relative Häufigkeit der Default-Klasse als Referenz.

Label	Default-Lerner		Naïve Bayes		Entscheidungsbaum	
	Avg. Acc.	(Std.dev.)	Avg. Acc.	(Std.dev.)	Avg. Acc.	(Std.dev.)
LFSC_Working_Not-working	97,415%	(0,124%)	97,413%	(0,122%)	97,415%	(0,124%)
LFSC_Working_Unemployed	96,757%	(0,171%)	96,757%	(0,171%)	96,757%	(0,171%)
LFSC_Jobbing_Not-working	88,381%	(1,212%)	88,379%	(1,215%)	88,316%	(1,311%)
LFSC_Jobbing_Unemployed	96,393%	(0,381%)	96,376%	(0,376%)	96,151%	(0,328%)
LFSC_Unemployed_Working	71,974%	(1,617%)	71,962%	(1,770%)	71,168%	(1,457%)
LFSC_Unemployed_Jobbing	96,157%	(0,672%)	96,138%	(0,678%)	96,157%	(0,672%)
LFSC_Training_Working	72,483%	(1,251%)	72,424%	(1,248%)	72,190%	(1,513%)
LFSC_Training_Jobbing	87,533%	(0,735%)	87,532%	(0,733%)	87,172%	(0,987%)
LFSC_Parenthood_Working	77,462%	(2,134%)	77,355%	(2,094%)	77,323%	(2,294%)
LFSC_Parenthood_Jobbing	93,029%	(3,138%)	93,008%	(3,122%)	92,986%	(3,221%)

10 Subgruppenentdeckung

Die im vorangegangenen Abschnitt vorgestellte Lernaufgabe des Funktionslernens aus Beispielen, löst - sofern geeignete Verfahren verwendet werden - das Problem der Bildung eines Globalmodells. Dieses Modell bezieht sich auf die Gesamtheit der vorgelegten Daten und kann verwendet werden, um für ungesehene Beispiele - und zwar für jedes Element des Instanzenraumes - Vorhersagen zu machen. Je nach Art des Modells (etwa bei den betrachteten Entscheidungsbäumen) kann den Daten zudem eine globale Beschreibung der Zusammenhänge entnommen werden. Diese Beschreibung gilt dann ebenfalls global, d.h. für die gesamte Stichprobe bzw. Population. Alternativ ist man jedoch häufig daran interessiert, Modelle zu extrahieren, die beschreibende Aussagen oder Vorhersagen nur für einen Teil der Stichprobe bzw. Population, z.B. für Instanzen, die eine bestimmte Eigenschaft besitzen, machen. Solche Teilmengen der Stichprobe bzw. Population werden auch als *Subgruppen* bezeichnet. Modelle, die nur für eine Subgruppe, d.h. nur *lokal* und nicht global gelten, werden folgerichtig *lokale Modelle* genannt.

Dieses Kapitel befasst sich mit der Entdeckung von interessanten, lokalen Modellen. Kapitel 10.1 führt dazu zunächst lokale Modelle als Abweichung von global gültigen Modellen ein. Das Kapitel motiviert zudem die Deutung des Grades der Abweichung lokaler Modelle von der durch globale Modelle erzeugten Erwartung als Interessantheit, anhand derer der Nutzen lokaler Modelle gemessen werden kann. Kapitel 10.2 definiert die daraus abgeleitete Lernaufgabe der Subgruppenentdeckung und erläutert die Repräsentation lokaler Modelle für Subgruppen durch Regeln. Kapitel 10.3 stellt einen intuitiven Algorithmus zur Entdeckung von Subgruppen und zugehöriger lokaler Modelle vor. Ein weiterer Algorithmus zum Auffinden lokaler Modelle, der auf dem zuvor vorgestellten, intuitiven Algorithmus aufbaut, dessen Nachteile jedoch weitestgehend vermeidet, wird schließlich in Kapitel 10.4 erläutert, bevor in Kapitel 10.5 die Ergebnisse der Anwendung dieses Algorithmus präsentiert werden.

10.1 Lokale vs. globale Modelle

Modelle beschreiben Muster in Daten, die wiederum auf geltende Zusammenhänge in der realen Welt in Bezug auf die Untersuchungseinheiten, die durch die Daten repräsentiert werden, hindeuten. Modelle können bezüglich ihrer Gültigkeit in globale Modelle auf der einen Seite und lokale Modelle auf der anderen Seite unterschieden werden. Globale Modelle beziehen sich auf die gesamten vorgelegten Daten (d.h. auf die gesamte Stichprobe bzw. - bei induktivem Schluss - auf die Gesamtpopulation). Im Gegensatz dazu beziehen sich lokale Modelle nur auf eine (echte) Teilmenge der Daten. Sie sind also nur für diese Teilmenge gültig und beschreiben bzw. machen Vorhersagen eben nur für diesen Teil der Daten (bzw. der Stichprobe oder Population).

Im Allgemeinen koexistieren globale und lokale Modelle für gegebene Daten, d.h. es lassen sich sowohl globale Modelle als auch lokale Modelle aus denselben Daten extrahieren. Diese

Koexistenz globaler und lokaler Modelle fasst Hand (2002) in der Formel

$$\text{data} = \text{background model} + \text{pattern} + \text{random component}$$

zusammen. Unter der Anwendung der hier verwendeten Begrifflichkeiten können Daten nach dieser Betrachtungsweise von Hand (2002) als Interaktion aus Globalmodell, lokalen Modellen und einer zufälligen Komponente (Rauschen) aufgefasst werden. Um Daten umfassend charakterisieren zu können, ist die Extraktion sowohl des Globalmodells als auch der vorhandenen lokalen Modelle wünschenswert bzw. nötig. Aus Sicht des Datenanalysten sind dabei im Regelfall allerdings lediglich solche lokalen Modelle interessant, die eine Abweichung von dem gefundenen Globalmodell beschreiben. Vielmehr werden lokale Modelle meist direkt als Abweichung von einem Globalmodell aufgefasst. In der Tat ist ein lokales Modell, welches dieselben Regelmäßigkeiten für einen Teil der Daten wie ein globales Modell für die Gesamtheit der Daten beschreibt, nicht sehr hilfreich. Dies macht folgendes hypothetisches Beispiel intuitiv klar: betrüge die Anzahl der Ehescheidungen pro Jahr im Verhältnis zur Anzahl der Einwohner beispielsweise etwa 0,2 Prozent, wäre es relativ uninteressant zu erfahren, dass dieser Anteil in einem bestimmten Bundesland ebenfalls 0,2 Prozent betrüge. Dies wäre jedoch nicht der Fall, wiche der Anteil in diesem Bundesland signifikant von dem auf Bundesebene ab. Ein lokales Modell ist also dann potentiell interessant, wenn es eine Abweichung von der aufgrund der Kenntnis eines Globalmodells erwarteten Situation beschreibt. Allerdings ist diese Abweichung zumeist nicht alleinige Determinante der Interessantheit eines lokalen Modells. Zwar mag etwa in obigem Beispiel eine signifikante Abweichung der betrachteten Größe in einem Bundesland von der im gesamten Bundesgebiet für die Datenanalysten (und potentielle Empfänger bzw. Nutzer der Forschungsresultate wie beispielsweise die Politik) interessant sein. Unter Umständen mag es jedoch kaum interessant sein, wiese beispielsweise - und rein hypothetisch - ein Dorf mit vielleicht 300 Einwohneren eine sogar höhere Abweichung zum bundesweiten Durchschnitt auf als ein Bundesland. Demnach spielt auch die Größe der Menge der durch das lokale Modell beschriebenen Untersuchungseinheiten, d.h. die Abdeckung der insgesamt betrachteten Untersuchungseinheiten durch das lokale Modell, eine Rolle. Insgesamt definiert sich die Interessantheit von lokalen Modellen also einerseits aus der Abweichung von einem Globalmodell, andererseits durch die Abdeckung der betrachteten Untersuchungseinheiten durch das lokale Modell. Zwischen diesen beiden Determinanten der Interessantheit besteht im Regelfall ein Trade-Off: kleine Subgruppen mögen trotz hoher Abweichung ebenso wie nur gering abweichende, dafür sehr große Subgruppen wenig interessant sein. In den meisten Fällen obliegt es dem Datenanalysten, durch entsprechende Wahl von Verfahren bzw. Parametern diese Determinanten gegeneinander abzuwägen. Dies wird in Abschnitt 10.2 tiefergehend thematisiert werden.

Während sich viele Data-Mining-Verfahren (wie auch etwa die im vorherigen Kapitel betrachteten Entscheidungsbäume) auf die Bildung eines Globalmodells konzentrieren, gewinnt zunehmend auch die Erkennung von lokalen Modellen und damit die Modellierung von Abweichungen bzw. Ausnahmen vom Globalmodell an Bedeutung in der Literatur (siehe beispielsweise Hand et al. (2002) und Morik et al. (2005)). Welche Bedeutung solche lokalen Modelle hinsichtlich ihrer Interpretation haben können, sei an einem weiteren, hypothetischen Beispiel zum Thema Arbeitslosigkeit erläutert. Ein einfaches Globalmodell könnte etwa die Arbeitslosenquote hinsichtlich ihrer Konjunkturabhängigkeit analysieren und eine gewisse Schwankung attestieren, die auf die Expansion und Kontraktion der wirt-

schaftlichen Aktivität zurückzuführen ist. Angenommen es existierten zwei lokale Modelle in Abhängigkeit des Bildungsgrades. Das lokale Modell könnte für niedrig qualifizierte Personen etwa eine überproportionale Schwankung im Konjunkturverlauf, das für hoch qualifizierte eine unterproportionale Schwankung beschreiben. Während das Globalmodell nur die Implikation eines politischen Eingriffs hinsichtlich der Abmilderung der Konjunkturschwankungen bzw. der Expansion des Wirtschaftswachstums zuließe, erlaubten die lokalen Modelle zudem eine Ableitung des Ziels der Bildungsverbesserung.

Grundsätzlich folgt aus den obigen Beobachtungen zusammenfassend, dass die Interessantheit von lokalen Modellen einerseits abhängig ist von der Abweichung von den auf dem Globalmodell basierenden Erwartungen, andererseits ebenfalls abhängt von der Abdeckung der Stichprobe bzw. Population durch das lokale Modell. Somit sollen lokale Modelle gefunden werden, die vom Globalmodell möglichst deutlich abweichen, gleichzeitig jedoch eine hinreichende Abdeckung gewährleisten. Dies wird im Kontext des maschinellen Lernens häufig gelöst durch Instrumentalisierung der Lernaufgabe der Subgruppenentdeckung und Anwendung entsprechender Verfahren. Bei der Subgruppenentdeckung sollen die Subgruppen gefunden werden, die sich in Bezug auf die global gültigen Zusammenhänge abweichend und damit unerwartet bzw. ungewöhnlich verhalten. Obwohl der Begriff Subgruppenentdeckung zunächst Unabhängigkeit der Entdeckung solcher auffälligen Subgruppen von der Bildung lokaler Modelle, die das (auffällige) Verhalten dieser Subgruppen beschreiben, impliziert, sind diese Komponenten eng verwoben. Der folgende Abschnitt beschreibt daher sowohl die Lernaufgabe der Subgruppenentdeckung als auch die Beschreibung von Subgruppen durch Regeln, die lokalen Modellen entsprechen.

10.2 Repräsentation und Entdeckung lokaler Modelle

Das Auffinden lokaler Modelle erfolgt im Bereich des maschinellen Lernens zumeist anhand der Lernaufgabe der Subgruppenentdeckung. Eine *Subgruppe* kann dabei formal als Teilmenge eines Instanzenraumes $X := X_1 \times \cdots \times X_m$ aufgefasst werden. Da die Entdeckung von Subgruppen in dieser Arbeit jedoch eher das Finden lokaler Modelle bedeuten soll, ist es hilfreich, Subgruppenentdeckung in einem überwachten Kontext zu betrachten. Somit sind hier Subgruppen als Teilmenge der Menge $X \times Y$ zu definieren. Ein Modell (oder eine Hypothese) ist hier wie im Szenario des Klassifikationslernens eine Funktion von X nach Y. Lokale Modelle sind im Allgemeinen allerdings nur für eine Teilmenge von X definiert. Die Lernaufgabe der Subgruppenentdeckung kann dann wie folgt definiert werden:

Definition 10.1 (Subgruppenentdeckung) *Seien X ein Instanzenraum, Y eine Menge möglicher Zielwerte und $\mathcal{L}_H := \mathcal{F}(X,Y) = \{h \mid h : X \to Y\}$ die Menge der Funktionen von X nach Y, der sogenannte Hypothesenraum. Sei $E \subseteq X \times Y$ eine Menge gegebener Beispiele und $q : \mathcal{L}_H \to \mathbb{R}$ eine* Qualitätsfunktion*. Die Lernaufgabe* Subgruppenentdeckung *ist dann alternativ definiert durch:*

1. *Gegeben X, Y und E sowie q nebst einer Mindestqualität $q_{\min} \in \mathbb{R}$. Bestimme alle Funktionen $h \in \mathcal{L}_H$ für die gilt, dass $q(h) \geq q_{\min}$.*

2. *Gegeben X, Y, E und q sowie eine natürliche Zahl $k \in \mathbb{N}$. Bestimme die Hypothesenmenge $H \subseteq \mathcal{L}_H$ mit $|H| = k$ derart, dass es keine $h' \in \mathcal{L}_H \setminus H$ und $h \in H$ gibt, für die gilt, dass $q(h') \geq q(h)$.*

Die Lernaufgabe Subgruppenentdeckung besteht also alternativ darin, alle (lokalen) Modelle zu finden, deren Qualität eine Schwelle $q_{\min}$ überschreitet, oder die k hinsichtlich ihrer Qualität besten Modelle zu finden. Die Qualität der Modelle wird dabei durch eine Qualitätsfunktion q, die die Interessantheit der Modelle bewertet, bestimmt. Die Subgruppen werden insofern implizit gefunden, als dass sie über den Definitionsbereich der Hypothesen gegeben sind.

In Kapitel 9 wurden bereits zwei Arten von Modellen, bayessche Klassifizierer als Produkt des Verfahrens Naïve Bayes sowie Entscheidungsbäume, eingeführt. Im Kontext der Subgruppenentdeckung und damit dem Auffinden lokaler Modelle werden als Modelle meist *Regeln* (auch *Entscheidungsregeln* genannt) verwendet bzw. gelernt. Eine einfache Form solcher Regeln sind *Hornregeln*. Eine Hornregel besteht aus einem Körper (auch Prämisse genannt) und einem Kopf (der sogenannten Konklusion). Der Körper ist eine Konjunktion von Atomen, d.h. bei propositionaler Logik eine Konjunktion von Paaren aus Attribut und zugehörigem Attributwert. Der Kopf einer Hornregel ist eine Vorhersage eines Wertes für das Label. Wird der Körper einer Hornregel als A, der Kopf der Regel als B bezeichnet, so schreibt sich die Regel als $A \rightarrow B$[1]. Regeln implizieren einen funktionalen Zusammenhang zwischen einer Teilmenge von X und Y. Diese Teilmenge von X ist genau die Menge von Elementen aus X bzgl. der die Atome im Körper der Regel zutreffen. Für diese Elemente wird der Kopf der Regel als Wert für Y vorhergesagt, während für die anderen Elemente aus X keine Aussage getroffen wird. Hornregeln stellen insofern eine natürliche Repräsentationsform für lokale Modelle dar, als dass sie explizit eine Vorhersage bzgl. des Zielwertes nur für einen Teil des Instanzenraumes machen. Dieser wird anhand des Körpers der Regel sowohl leicht definiert als auch identifiziert.

Bei Verwendung von Regeln als Repräsentation lokaler Modelle für Subgruppen bedeutet die Entdeckung von Subgruppen das Auffinden solcher Regeln, die eine Abweichung vom normalen, erwarteten Verhalten der Stichprobe bzw. Population als Ganzem, beschreiben. Die Abweichung als Indikator der Interessantheit der Regel und damit des lokalen Modells wird dabei durch die Qualitätsfunktion q gemessen, die die Modelle durch eine reelle Zahl bewertet und damit bezüglich ihrer Interessantheit anordnen lässt. Die meisten Qualitätsfunktionen setzen dazu für eine gegebene Regel die Wahrscheinlichkeit, das durch die Regel vorhergesagte Label innerhalb der durch die Regel definierten Subgruppe zu beobachten, in Bezug zur A-Priori-Wahrscheinlichkeit, dieses Label in der gesamten Stichprobe (bzw. Population) zu beobachten. Sei etwa eine Regel $A \rightarrow B$ gegeben. Dann ist die A-Priori-Wahrscheinlichkeit, das durch den Kopf B der Regel vorhergesagte Label zu beobachten, durch $\Pr(B)$ gegeben. Die Wahrscheinlichkeit, dieses Label in der Subgruppe, die durch den Körper A der Regel definiert ist, zu beobachten, ist gleich der bedingten Wahrscheinlichkeit $\Pr(B|A)$. Diese Wahrscheinlichkeit gibt an, wie exakt die Vorhersage durch die Regel für die betrachtete Subgruppe ist. Sie wird daher auch als *Precision* bezeichnet.

Definition 10.2 (Precision) *Sei $A \rightarrow B$ eine Regel. Dann ist die* Precision *der Regel definiert durch*

$$\text{Precision}(A \rightarrow B) := \Pr(B|A).$$

[1] Hierbei bezeichnen A und B nicht nur die logischen Konstrukte, d.h. die Konjunktionen von Atomen, sondern auch die jeweiligen zugehörigen Mengen von Beispielen, d.h. die Menge der Beispiele, für die die Prämisse zutrifft, bzw. die Menge, die das durch die Regel vorhergesagte Label haben.

Die genannten Wahrscheinlichkeiten können durch Berechnung der korrespondierenden relativen Häufigkeiten auf der Menge der gegebenen Beispiele leicht berechnet werden. Die einfachste Form, die Precision einer Regel in Bezug zur A-Priori-Wahrscheinlichkeit $P(B)$ als eigentliche Erwartung zu setzen, ist die Bildung der Differenz:

Definition 10.3 (Bias) *Sei $A \rightarrow B$ eine Regel. Der* Bias *oder auch* Gain *dieser Regel ist dann gegeben durch*

$$\begin{aligned} \text{Bias}(A \rightarrow B) &:= \text{Precision}(A \rightarrow B) - \Pr(B) \\ &= \Pr(B|A) - \Pr(B). \end{aligned}$$

Wäre etwa als hypothetisches Beispiel die Wahrscheinlichkeit, arbeitslos zu werden, gleich 2 Prozent und betrüge sie bei Personen, die keinen Schulabschluss haben, 5 Prozent, so betrüge der Bias der Regel, die für Personen ohne Schulabschluss den Übergang in die Arbeitslosigkeit vorhersagt, 3 Prozent. Die meisten Verfahren zur Subgruppenentdeckung inkorporieren den Bias innerhalb der verwendeten Qualitätsfunktion. Aus den Betrachtungen im vorherigen Abschnitt ist zu folgern, dass eine alleinige Verwendung des Bias zur Messung der Qualität von Regeln jedoch problematisch ist. Dies liegt daran, dass kleine Subgruppen, die nur aus wenigen Beispielen bestehen, leicht eine sehr hohe Precision haben können. So existieren möglicherweise sogar Subgruppen, die nur aus einem einzigen Beispiel bestehen. Regeln für eine solche haben dann automatisch eine Precision von 100 Prozent und erreichen damit den maximal möglichen Wert, sofern das tatsächliche Label des Beispiels durch die Regel vorhergesagt wird. Bei einer sehr großen Menge vorgelegter Beispiele sind sowohl Subgruppen, die nur aus einem oder aber wenigen Beispiel bestehen, als auch die zugehörigen Modelle jedoch relativ uninteressant, da die Abdeckung des Modells zu gering ist. Aufgrunddessen ist ein Einbezug der Größe der durch die Regeln abgedeckten Subgruppen in die Qualitätsfunktion sinnvoll. Dies kann anhand der Wahrscheinlichkeit für das Beobachten eines Elementes der Subgruppe leicht geschehen. In Bezug auf eine Regel $A \rightarrow B$ entspricht diese der Wahrscheinlichkeit, dass ein zufällig gezogenes Beispiel die durch die Atome im Körper A der Regel gegebenen Werte bzgl. der jeweiligen Attribute aufweist. Diese Wahrscheinlichkeit ist durch $\Pr(A)$ gegeben und wird häufig als *Coverage* einer Subgruppe bzw. Regel bezeichnet.

Basierend auf Coverage und Bias einer Regel wurden verschiedene Qualitätsfunktionen vorgeschlagen, die diese beiden Komponenten multiplikativ gegeneinander abwägen. Die einfachste Form ist die folgende:

Definition 10.4 (Weighted Relative Accuracy) *Sei $A \rightarrow B$ eine Regel. Dann wird durch*

$$\begin{aligned} \text{WRAcc}(A \rightarrow B) &:= \text{Coverage}(A \rightarrow B) \cdot \text{Bias}(A \rightarrow B) \\ &= \Pr(A) \cdot (\Pr(B|A) - \Pr(B)) \end{aligned}$$

die Qualitätsfunktion Weighted Relative Accuracy *definiert.*

Andere Qualitätsfunktionen favorisieren eine höhere Gewichtung entweder der Coverage oder des Bias der Regeln etwa durch Verwenden der Quadratwurzel der Coverage, was zu einer faktischen Erhöhung des Gewichts der Coverage führt, oder die Quadrierung der Coverage mit gegenteiligem Effekt[2]. Fürnkranz (2005) analysiert diesen Trade-Off zwischen

[2]Man beachte hierbei, dass der Wertebereich der Coverage von Subgruppen gleich dem Intervall $[0, 1]$ ist.

Coverage und Bias und beschreibt hierzu verschiedene Qualitätsfunktionen hinsichtlich ihrer Eigenschaften in Bezug auf etwa die Bevorzugung kleinerer oder größerer Subgruppen bei der Subgruppenentdeckung.

Anhand solcher Qualitätsfunktionen können Regeln bzgl. ihrer Interessantheit bewertet und angeordnet werde. Bislang wurde jedoch noch nicht betrachtet, wie solche Regeln erzeugt werden, um sie dann anhand ihrer Qualität evaluieren und vergleichen zu können. Dies wird im nächsten Abschnitt nachgeholt.

10.3 Top-Down-Subgruppenentdeckung

Die grundsätzliche Vorgehensweise bei der Subgruppenentdeckung, für die lokale Modelle durch Regeln repräsentiert werden, basiert auf der Erzeugung solcher Regeln und der anschließenden Evaluation dieser anhand der gewählten Qualitätsfunktion. Die Erzeugung der Regeln bedeutet dabei, anhand der in den Daten vorhandenen Attribute sowie den möglichen Werten für diese Attribute einzelne Atome (also Attribut-Werte-Paare) zu bilden und diese danach zu Regeln zusammenzusetzen. Üblicherweise geschieht dies durch schrittweise Verfeinerung der Regeln. Dies kann beispielsweise durch eine Breitensuche geschehen. Hierbei werden - ausgehend von der leeren Regel, d.h. einer Regel mit leerem Körper - in einem ersten Schritt alle Regeln der Länge eins (d.h. sie enthalten genau ein Atom) erzeugt und diese evaluiert. Danach werden alle Regeln der Länge zwei erzeugt, usw. (siehe Abbildung 10.1). Durch Anwendung einer einmalig geordneten Liste der Attribute kann dabei die

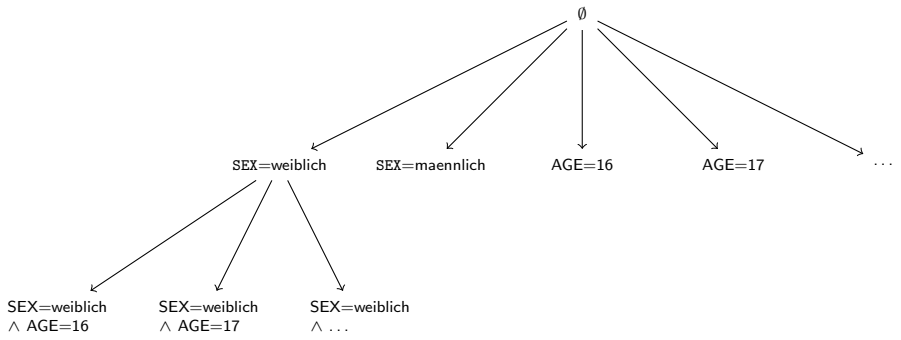


Abbildung 10.1: Breitensuche zur Erzeugung von Regeln

mehrfache Erzeugung von Regeln vermieden werden, indem für die Teilbäume der Belegung des ersten Attributes (im Beispiel `SEX`) unter der Wurzel des Baumes noch alle anderen Attribute als mögliche Verfeinerungen der Regel zugelassen werden, für jene Teilbäume, die zuerst das zweite Attribut (im Beispiel `AGE`) belegen, das erste Attribut (`SEX`) jedoch nicht mehr für mögliche Verfeinerungen der Regeln zugelassen wird, usw. Bezüglich des Labels sollten bei dieser Vorgehensweise die gebildeten Regelkörper mit allen möglichen Zielwerten kombiniert und die resultierenden Regeln evaluiert werden. Aus dieser Vorgehensweise resultiert der Top-Down-Ansatz, der im relationalen Kontext beispielsweise durch den von Wrobel (1997) propagierten Algorithmus Midos implementiert wird. Der Algorithmus bildet alle möglichen Regeln und evaluiert diese. Er ist damit vollständig und findet garantiert die besten k Regeln bzw. alle Regeln mit einer Mindestqualität $q_{\min}$. Der Nachteil zu Gunsten der Vollständigkeit des Algorithmus ist allerdings die immense Laufzeit, die durch die Er-

zeugung (und Evaluation) aller Regeln entsteht. Es existieren jedoch einige Möglichkeiten zum Abschneiden des Suchraumes in obig dargestelltem Prozess.

Die ersten beiden Varianten beschneiden den Suchraum nach Benutzervorgaben - allerdings im Allgemeinen mit Verlust der Vollständigkeit. Die erste Variante basiert auf der Verwendung einer vom Benutzer vorgegebenen Mindestcoverage der jeweiligen Subgruppen. Hat eine Regel diese Schranke unterschritten, so braucht sie nicht weiter verfeinert werden, da die Coverage von Regeln monoton bzgl. der Verfeinerung der Regeln ist. Als zweite Variante könnte ebenso eine maximale Tiefe der Suche vorgegeben und damit die Länge der erzeugten Regeln beschränkt werden. Dies ist unter Umständen auch vorteilhaft in Bezug auf die Verständlichkeit und Interpretierbarkeit der Regeln.

Eine weitere Möglichkeit des Beschneidens des Suchraumes resultiert dagegen nicht in dem Verlust der Vollständigkeit. Sie kann wie folgt resümiert werden: Prinzipiell existiert eine Monotonie wie bei der Coverage für Qualitätsfunktionen im Allgemeinen nicht. Daher kann auf Basis einer Auswertung der Qualität von Regeln anhand der Qualitätsfunktion direkt kein Beschneiden des Suchraumes erfolgen. Dies resultiert für die oben beschriebene Qualitätsfunktion Weighted Relative Accuracy zum Beispiel daraus, dass bei der Verfeinerung einer Regel der Bias sowohl zu- als auch abnehmen kann und damit die Qualitätsfunktion nicht monoton bezüglich der Verfeinerung von Regeln ist. Wrobel (1997) schlägt jedoch die Ausnutzung folgender Möglichkeit vor: Für viele Qualitätsfunktionen kann eine sogenannte *optimistische Schätzfunktion* gefunden werden, die eine obere Schranke der Qualität einer Regel und aller Verfeinerungen dieser Regel angibt. Ist der Wert dieser optimistischen Schätzfunktion für eine Regel kleiner als die vorausgesetzte Mindestqualität (bei der ersten Alternative der Subgruppenentdeckung) bzw. ist der Wert (bei der zweiten Alternative der Subgruppenentdeckung) kleiner als die Qualität der besten k Regeln, die bisher gefunden wurden, so braucht diese Regel sowie alle Verfeinerungen der Regel nicht mehr betrachtet zu werden. Für die Qualitätsfunktion Weighted Relative Accuracy ist beispielsweise

$$\begin{aligned} \mathrm{WRAcc}_{\max}(A \rightarrow B) &:= \Pr(A) \cdot (\max(\Pr(B|A)) - \Pr(B)) \\ &\quad\ \Pr(A) \cdot (1 - \Pr(B)) \end{aligned}$$

eine optimistische Schätzfunktion.

Zusammenfassend ist in Abbildung 10.2 ein Algorithmus zur Entdeckung der k besten Subgruppen (bzw. Hypothesen) exemplarisch abgebildet, der Regeln top-down per Verfeinerung erzeugt und diese anhand der Qualitätsfunktion evaluiert. In diesem Algorithmus ist beispielhaft sowohl die Möglichkeit der Benutzerangabe einer maximalen Tiefe der Suche als auch das Konzept des Beschneidens des Suchraums auf Basis optimistischer Schätzfunktionen implementiert. Als weitere Alternativen zu einer vollständigen Suche bieten sich (zur Verringerung der Laufzeit) etwa heuristische Suchen oder Sampling-Algorithmen, die die Qualitätsfunktionen nicht auf den gesamten zur Verfügung stehenden Daten evaluieren, an. So beschreiben Scheffer und Wrobel (2000, 2002) sampling-basierte Verfahren, die die k besten Regeln auf zufällig ausgewählten Beispielen zwar nicht garantieren, jedoch mit einer gewissen Konfidenz, d.h. einer gewissen Wahrscheinlichkeit finden.

Eingabe : Menge von Beispielen E, Hypothesenanzahl k und maximale Tiefe der Verfeinerung d_{max}
Ausgabe: Menge der k besten Hypothesen H und zugehörige Werte $q(h)$

```
1  H ← ∅
2  h_0 ← leere Regel
3  Q ← {h_0}
4  for i ← 1 to d_max do
5      Q' ← Verfeinerung(Q)
6      forall h ∈ Q' do
7          q(h) ← Qualität(h)
8          h_min ← argmin_{h'∈H} q(h')
9          if q(h) > q(h_min) then ersetze schlechteste Hypothese in H
10             H ← (H \ {h}) ∪ {h_min}
11         end
12         if q(h_min) > OptimischeSchätzung(h) then beschneide Suchraum
13             Q' ← Q' \ {h}
14         end
15         Q ← Q'
16     end
17 end
```

Abbildung 10.2: Top-Down-Algorithmus zur Subgruppenentdeckung

10.4 Knowledge-Based Sampling

Trotz der diversen, im letzten Abschnitt beschriebenen Einschränkungen sowie Optimierungen der Suche nach qualitativ bedeutsamen Subgruppen ist ein immanenter Nachteil einer alleinigen Anwendung des Top-Down-Verfahrens, dass die gefundenen Subgruppen häufig eine große Überlappung aufweisen. Zur Verdeutlichung dieses Sachverhaltes sei folgendes Beispiel gegeben: Angenommen es seien innerhalb einer Untersuchung der Bildungsabschluss, die Häufigkeit des Schulbesuchs sowie Übergänge in die Arbeitslosigkeit erfasst. Nun könnte wie in obigem Beispiel die Wahrscheinlichkeit eines Übergangs in die Arbeitslosigkeit bei Personen ohne Schulabschluss im Vergleich zur Gesamtwahrscheinlichkeit erhöht sein und somit eine interessante Subgruppe gefunden worden sein. Sei weiterhin hypothetisch angenommen, dass Personen ohne Schulabschluss meist eine geringe Häufigkeit des Schulbesuchs haben. Dann würde eine weitere Subgruppe gefunden werden, die mit der erstgenannten im Wesentlichen übereinstimmt, bei der unter Umständen nur die Beschreibung (etwa anhand die Prämisse einer Regel) durch das lokale Modell differiert. Zwar sind generell beide lokalen Modelle auch für sich interessant. Eventuell könnten jedoch viele solcher überlappenden Subgruppen (und die dementsprechenden lokalen Modelle) das Auffinden weiterer interessanter Subgruppen, die jedoch eine etwas geringere Qualität als die vielen überlappenden haben, verhindern (etwa durch zu restriktive Wahl des Parameters k oder $q_{\min}$ aufgrund der Fülle gefundener Regeln).

Eine elegante Lösung zur Vermeidung der soeben beschriebenen Problematik stellt Scholz (2005a,b) vor. Die zentrale Eigenschaft dieses Ansatzes besteht darin, dass in die Ent-

deckung von Subgruppen Vor- bzw. Hintergrundwissen in Form eines bereits bekannten Modells einbezogen wird, sodass bereits bekanntes Wissen nicht erneut als Modell aus den Daten extrahiert wird. Wäre in obigem Beispiel bereits eine höhere Wahrscheinlichkeit des Überganges in die Arbeitslosigkeit für Personen ohne Schulabschluss als Hintergrundwissen bekannt, so würde durch den Ansatz beispielsweise die Bildung eines lokalen Modells, welches die gleiche Subgruppe beschreibt, verhindert und somit die Regel, die die gleiche Vorhersage für Personen mit seltenem Schulbesuch trifft, ignoriert.

Die Methode, die dessen fähig ist, basiert auf Sampling und wird als *Knowledge-Based Sampling* bezeichnet, da sie vorheriges Wissen inkorporiert. Dabei wird vorher bekanntes Wissen aus den vorliegenden Daten herausgesamplet, sodass es nicht erneut in Form eines Modells extrahiert werden kann. Angenommen die vorliegenden Beispiele wurden anhand einer festen Wahrscheinlichkeitsverteilung D unabhängig und identisch verteilt gezogen. Dann bedeutet die Einbeziehung des Vorwissens das Übergehen von der Wahrscheinlichkeitsverteilung D zu einer Wahrscheinlichkeitsverteilung D', die die statistische Auffälligkeit der Sachverhalte, die als Vorwissen gegeben sind, nicht mehr enthält. Solch eine Transformation kann durch entsprechende Gewichtung der Beispiele erfolgen.

Wie oben beschrieben ist eine Regel $A \rightarrow B$ statistisch auffällig, für die $P(B|A)$ von $P(B)$ signifikant abweicht. Sei nun eine solche Regel als Vorwissen gegeben. Dann soll die statistische Auffälligkeit dieser Regel durch Neugewichtung der Beispiele und damit Übergang von der vorherigen Verteilung D zur Verteilung D' eliminiert werden, d.h. es soll bezüglich der neuen Verteilung D' gelten, dass

$$\Pr_{D'}(B|A) = \Pr_{D'}(B). \tag{10.1}$$

Dies bedeutet, dass bezüglich der Verteilung D' die Wahrscheinlichkeit, das durch B vorhergesagte Label unter der Prämisse A zu beobachten, gleich der A-Priori-Wahrscheinlichkeit, B zu beobachten, sein soll. Somit wäre die durch A beschriebene Subgruppe bzgl. D' nicht mehr statistisch auffällig und würde nicht mehr als interessante Subgruppe entdeckt. Prinzipiell soll die Verteilung bei der Umgewichtung der Beispiele jedoch nicht gänzlich verändert werden, sodass weitere Bedingungen gefordert werden müssen. So muss etwa gelten, dass die Wahrscheinlichkeit, ein Element der Subgruppe zu ziehen bzgl. der Verteilungen D und D' identisch sind, ebenso wie die A-Priori-Wahrscheinlichkeiten für das vorhergesagte Label. Somit muss gelten:

$$\Pr_{D'}(A) = \Pr_{D}(A) \tag{10.2}$$

$$\Pr_{D'}(B) = \Pr_{D}(B). \tag{10.3}$$

Weitere Gleichheiten, die gefordert werden müssen ergeben sich bzgl. der durch die Regel $A \rightarrow B$ induzierten Partitionen in der Menge der Beispiele (für nähere Informationen sei hier auf Scholz (2005b) verwiesen). Eine Transformation der Verteilung, die den genannten Bedingungen genügt, kann mittels der Multiplikation vorhandener Beispielgewichte (und damit der Sampling-Wahrscheinlichkeiten dieser Beispiele) mit dem Inversen des *Lifts* eines Beispiels geschehen. Hierzu sei zunächst der Lift einer Regel wie folgt definiert:

Definition 10.5 (Lift) *Sei $A \to B$ eine Regel. Dann ist der* Lift *der Regel gegeben durch*

$$\begin{aligned} \text{Lift}(A \to B) &:= \frac{\text{Precision}(A \to B)}{\Pr(B)} \\ &= \frac{\Pr(B|A)}{\Pr(B)}. \end{aligned}$$

Der Lift eines Beispiels ergibt sich, in dem in der Definition des Lifts einer Regel die jeweilige Partition bzgl. der Regel betrachtet wird, zu der das Beispiel gehört:

$$\text{Lift}(x, A \to B) := \begin{cases} \text{Lift}(A \to B) & \text{wenn } x \in A \cap \bar{B} \\ \text{Lift}(A \to \bar{B}) & \text{wenn } x \in A \cap \bar{B} \\ \text{Lift}(\bar{A} \to B) & \text{wenn } x \in \bar{A} \cap B \\ \text{Lift}(\bar{A} \to \bar{B}) & \text{wenn } x \in \bar{A} \cap \bar{B}. \end{cases}$$

Somit kann ein Übergang von der Verteilung D zu D' geschehen durch

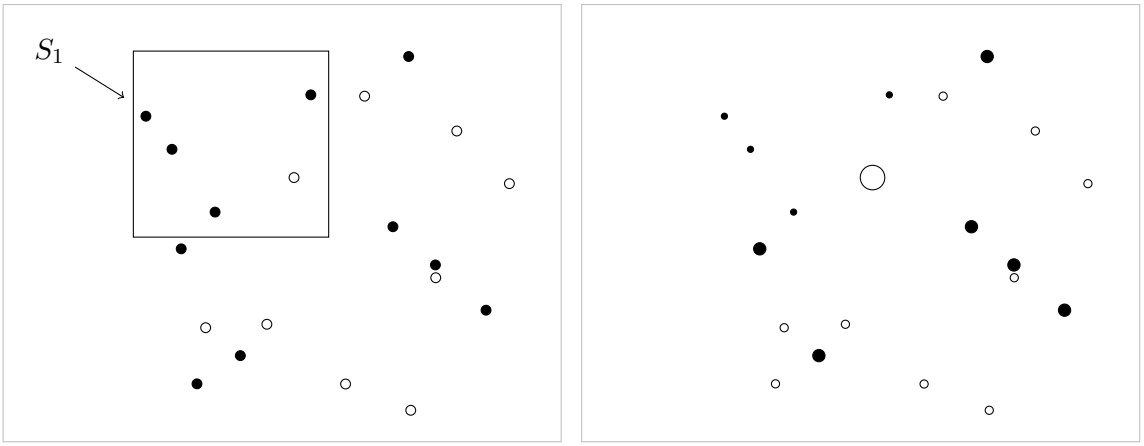
$$\Pr_{D'}(x) = \frac{\Pr_D(x)}{\text{Lift}(x, A \to B)}, \tag{10.4}$$

wobei der Lift hinsichtlich der alten Verteilung D zu berechnen ist. Einen Beweis, dass die oben genannten Bedingungen hinsichtlich der Transformation der Verteilungen durch Gleichung (10.4) erfüllt werden, gibt Scholz (2005b). Erfolgt eine technische Berücksichtigung der Wahrscheinlichkeiten für Beispiele durch Gewichte, dann erfolgt die Transformation des Gewichts $w_t(x)$ eines Beispiels $x \in E$ in Iteration t bzgl. eines als Vorwissen bekannten Regel $A \to B$ durch

$$w_{t+1}(x) = \frac{w_t(x)}{\text{Lift}(x, A \to B)}. \tag{10.5}$$

Die soeben beschriebene Vorgehensweise zum Heraussamplen von Vorwissen kann zur Entdeckung interessanter Subgruppen genutzt werden, indem iterativ interessante Modelle entdeckt werden, deren nun als Vorwissen gegebene Auffälligkeit mittels der beschriebenen Transformation aus den Daten eliminiert wird. Anschließend wird dann erneut ein interessantes Modell gelernt, und so weiter. Abbildung 10.3 veranschaulicht die beiden Komponenten des Prozesses exemplarisch.

Darüber hinaus sei kurz erwähnt, dass es möglich ist, durch Kombination der in den einzelnen Iterationen berechneten Lifts (für Beispiele) unter Instrumentalisierung der Unabhängigkeitsannahme von Naïve Bayes Vorhersagen für Beispiele zu treffen. Somit können die vom Knowledge-Based Sampling und einem Verfahren zur Regelinduktion erzeugten Regelmengen neben ihrer deskriptiven Funktion auch als Klassifikationsmodelle dienen und somit auch hinsichtlich ihrer Klassifikationsperformanz evaluiert werden. Da in dieser Arbeit hauptsächlich die deskriptive Funktion der Regelmengen von Bedeutung ist, sei für eine ausführliche Darstellung und Herleitung der Vorgehensweise zur Nutzung der Regelmengen zur Klassifikation von Beispielen abermals auf Scholz (2005b) verwiesen.



(a) Regelinduktion (bzw. Subgruppenentdeckung) (b) Umgewichtung

Abbildung 10.3: Funktionsweise des Knowledge-Based Sampling

10.5 Ergebnisse

Im Rahmen dieser Arbeit wurde das in Kapitel 10.4 beschriebene Knowledge-Based Sampling zur Entdeckung wenig überlappender Regeln angewendet. Als inneres Verfahren zur Induktion einer Regel und damit dem Finden eines lokalen Modells mit hoher Qualität wurde die in Kapitel 10.3 erläuterte Top-Down-Strategie verwendet. Allerdings wurde hierbei keine vollständige Suche durchgeführt, sondern diese Suche durch eine gezielte Vorgabe beschränkt. Dies geschah durch Begrenzung der Länge der Hornregeln auf zwei Atome im Körper. Dies hat neben der Beschränkung der Suche (und damit der Laufzeit des Algorithmus) den Vorteil, dass die Regeln nicht zu komplex und damit unverständlich werden. Als Qualitätsfunktion, anhand derer die Regeln bewertet wurden, wurde die Funktion Weighted Relativ Accuracy verwendet. In RapidMiner entspricht dieser Aufbau der Anwendung des Regellerners `BestRuleInduction` innerhalb des Meta-Operators `BayesianBoosting`, dessen Parameter rescale_label_priors hierzu auf true und allow_marginal_skews auf false gesetzt werden müssen. Des Weiteren wurde die Anzahl der Iterationen der Durchführung des Knowledge-Based Samplings und damit die Anzahl der gefundenen lokalen Modelle anhand des Parameters iterations auf 20 begrenzt.

Exemplarisch werden im Folgenden Ergebnisse der durchgeführten Experimente für drei der zehn Lernaufgaben ausführlich besprochen. Die Ergebnisse der Experimente für die restlichen sieben Lernaufgaben sind in tabellarischer Form in Anhang A zu finden. Die aus der Anwendung des Knowledge-Based Samplings für die Lernaufgabe mit dem Label `LFSC_Working_Not-working` resultierenden Regeln sind in Verbindung mit ihrer Weighted Relative Accuracy sowie mit ihrer Coverage, d.h. ihrer relativen Abdeckung der Beispielmenge, in Tabelle 10.1 aufgelistet. Auf technischer Ebene fällt bei der Betrachtung dieser Ergebnisse generell auf, dass die Qualität (die Weighted Relative Accuracy) für die Regeln über die Iterationen meist abnimmt. Ausnahmen hiervon sind durch die Umgewichtung der Beispiele zur Veränderung der Verteilung und Eliminierung des statistischen Auffälligkeit des Vorwissens bedingt. Ein Abnehmen der Coverage ist dagegen nicht auszumachen, was insofern trivialerweise nicht verwunderlich ist, als dass die Qualitätsfunktion

Weighted Relative Accuracy auch den Bias als Komponente enthält und Regeln dementsprechend auch danach bewertet. Als zunächst fragwürdige Anomalie fallen außerdem die Regel (SEX=*weiblich*) ∧ (FAMSTD=*verh. zus.*) ⇒ *positive*, die Ergebnis der ersten Iteration war, und die Regel (SEX=*weiblich*) ∧ (FAMSTD=*verh. zus.*) ⇒ *negative* als Ergebnis der 17. Iteration auf. Diese beiden Modelle scheinen sich zu widersprechen, da sie für die gleiche Subgruppe unterschiedliche Konklusionen folgern. Ein Widerspruch besteht hier jedoch nicht. Es ist vielmehr zu beachten, dass beim Knowledge-Based Sampling eine gefundene Regel immer die statistische Auffälligkeit einer Subgruppe unter Berücksichtigung des Vorwissens, welches in vorhergehenden Iterationen entdeckt wurde, beschreibt. Die Regeln können und sollen daher zum einen nicht gleichrangig, zum anderen jedoch auch nicht für sich und damit alleinig betrachtet werden. Sie müssen stattdessen stets unter Berücksichtigung der zuvor gefundenen Regeln interpretiert werden.

Auf inhaltlicher Ebene fallen zuallererst die Regeln auf, die eine erhöhte Tendenz der Altersgruppe der 59- bis 63-jährigen, in den Zustand nichterwerbstätig überzugehen, beschreiben. Hierdurch wird der schon in den Globalmodellen, die in Kapitel 9.4 beschrieben

Tabelle 10.1: KBS-Regeln für LFSC_Working_Not-working

Regel	WRAcc	Coverage
(SEX=*weiblich*) ∧ (FAMSTD=*verh. zus.*) ⇒ *positive*	0,0059	0,2417
(PSBIL=*Hauptschulabschluss*) ⇒ *positive*	0,0049	0,3999
(IMMMIGRATED=*nicht immigriert*) ∧ (HHCHILDREN=*Ja*) ⇒ *negative*	0,0039	0,3042
(SEX=*weiblich*) ∧ (LOC1989=*West Germany*) ⇒ *positive*	0,0033	0,3429
(IMMIGRATED=*nicht immigriert*) ∧ (FAMSTD=*ledig*) ⇒ *negative*	0,0030	0,2695
(AUSB=*Berufsausbildung*) ⇒ *negative*	0,0032	0,4655
(SEX=*maennlich*) ∧ (HHGR=*2*) ⇒ *positive*	0,0026	0,1450
(AGE=*60*) ⇒ *positive*	0,0022	0,0114
(ERLJOB=*Ja*) ⇒ *negative*	0,0020	0,5208
(SEX=*weiblich*) ∧ (WIEDERVEREINIGUNG=*Nein*) ⇒ *positive*	0,0019	0,1220
(AGE=*63*) ⇒ *positive*	0,0018	0,0054
(LOC1989=*West Germany*) ∧ (PSBIL=*Hauptschulabschluss*) ⇒ *negative*	0,0017	0,3606
(AGE=*62*) ⇒ *positive*	0,0016	0,0070
(SEX=*weiblich*) ∧ (HHCHILDREN=*Ja*) ⇒ *positive*	0,0016	0,1431
(IMMIGRATED=*nicht immigriert*) ∧ (OEFFD=*Nein*) ⇒ *negative*	0,0015	0,6331
(AUSB=*Studium*) ⇒ *negative*	0,0015	0,1208
(SEX=*weiblich*) ∧ (FAMSTD=*verh. zus.*) ⇒ *negative*	0,0016	0,2434
(AGE=*61*) ⇒ *positive*	0,0014	0,0086
(AGE=*59*) ⇒ *positive*	0,0014	0,0145
(WACHSTUM_L3=*mittel*) ∧ (ZYKLUS=*Aufschwung*) ⇒ *positive*	0,0014	0,3350

wurden, angedeutete Einfluss des Alters auf den Übergang von der Erwerbstätigkeit in die Nichterwerbstätigkeit konkretisiert. Somit wird die Intuition, die eine erhöhte Übergangstendenz aufgrund etwa von Frühverrentung vermuten lässt, bestätigt. Auch weitere Regelmäßigkeiten, die aus den dargestellten Regeln abzuleiten sind, erscheinen weitestgehend intuitiv erklärbar. Die anderen Regeln sind teilweise weniger intuitiv bzw. vielfach höchstens hypothetisch und damit spekulativ erklärbar. Nichtsdestotrotz weisen sie auf bestehende Auffälligkeiten in den Daten hin, die interessant sein könnten aber einer weiteren Untersuchung bedürften. Als Beispiel hierfür ist etwa das Modell zu nennen, welches für Frauen in der Zeit vor der Wiedervereinigung eine erhöhte Wahrscheinlichkeit des Überganges in die Nichterwerbstätigkeit beschreibt.

Tabelle 10.2: KBS-Regeln für `LFSC_Working_Unemployed`

Regel	WRAcc	Coverage
(`IMMIGRATED`=*nicht immigr.*) ∧ (`LOC1989`=*West Germany*) ⇒ *negative*	0,0071	0,7314
(`ERLJOB`=*Ja*) ⇒ *negative*	0,0044	0,5215
(`LOC1989`=*West Germany*) ∧ (`FAMSTD`=*verh. zus.*) ⇒ *negative*	0,0038	0,4790
(`IMMIGRATED`=*nicht immigr.*) ∧ (`PSBIL`=*Hauptschulabschl.*) ⇒ *positive*	0,0042	0,3746
(`OEFFD`=*Ja*) ⇒ *negative*	0,0030	0,2529
(`ERWZEIT`=*1*) ⇒ *positive*	0,0028	0,1160
(`ERWZEIT`=*0*) ⇒ *positive*	0,0023	0,0607
(`SEX`=*maennlich*) ∧ (`HHCHILDREN`=*Ja*) ⇒ *negative*	0,0022	0,2106
(`ZYKLUS`=*Abschwung*) ⇒ *positive*	0,0019	0,2208
(`BETR`=*GE 2000*) ⇒ *negative*	0,0019	0,2217
(`IMMIGRATED`=*nicht immigriert*) ∧ (`AUTONO`=*niedrig*) ⇒ *positive*	0,0017	0,2343
(`ORTKINDH`=*Grossstadt*) ⇒ *positive*	0,0017	0,2259
(`AUTONO`=*sehr niedrig*) ⇒ *positive*	0,0018	0,1538
(`ERLJOB`=*Nein*) ⇒ *negative*	0,0020	0,3244
(`ORTKIND1`=*Ja, immer noch*) ∧ (`HHCHILDREN`=*Nein*) ⇒ *negative*	0,0017	0,3045
(`BULA`=*Baden-Wuerttemberg*) ⇒ *negative*	0,0015	0,1409
(`LOC1989`=*West Germany*) ∧ (`ERWZEIT`=*3*) ⇒ *positive*	0,0016	0,0550
(`ERWZEIT`=*2*) ⇒ *positive*	0,0016	0,0870
(`IMMIGRATED`=*nicht immigr.*) ∧ (`TYPHH`=*(Ehe-)Paar o. K.*) ⇒ *positive*	0,0016	0,2256
(`PSBIL`=*Abitur*) ⇒ *negative*	0,0015	0,1713

Für den Übergang von der Erwerbstätigkeit in die Arbeitslosigkeit wurden ebenfalls solche Regeln gefunden, die intuitiv nachvollziehbar sind, und solche, die eventuell einer eingehenderen und auf den jeweiligen Sachverhalt konzentrierte Analyse zur Klärung ihrer Bedeutung bedürfen. Die gefundenen Regeln sind in Tabelle 10.2 dargestellt. Analog zu den bei der letzten Lernaufgabe relevanten Regeln zum Alter der Untersuchungseinheiten haben in der nun betrachteten Lernaufgabe die Regeln, in denen der Körper lediglich Atome für das Attribut `ERWZEIT` enthält, augenscheinlich eine hohe Relevanz. So haben Personen, die erst kurz (d.h. null bis drei Jahre) zu einem Betrieb gehören, eine durchschnittlich höhere Wahrscheinlichkeit arbeitslos zu werden. Auch dies ist eine Bestätigung bzw. Konkretisierung des Ergebnisses von Naïve Bayes. Daneben gibt es einige weitere Regeln, die ebenso intuitive Vermutungen bestätigen, etwa eine erhöhte Wahrscheinlichkeit zum Übergang in die Arbeitslosigkeit im Abschwung bzw. im Fall einer sehr niedrigen Autonomie bei der Ausübung der beruflichen Tätigkeit. Auch die niedrigere Wahrscheinlichkeit eines Übergangs in die Arbeitslosigkeit bei Personen, die im öffentlichen Dienst arbeiten, bzw. bei Personen aus Baden-Württemberg ist vermutlich lediglich ein Beleg für bereits bekannte Tatsachen. Interessanter sind dagegen eher etwa die Regeln, die Aussagen hinsichtlich der Bildung machen. So wird zum einen eine erhöhte Wahrscheinlichkeit des Übergangs in die Arbeitslosigkeit bei Personen mit Hauptschulabschluss, eine verminderte Wahrscheinlichkeit dagegen bei Personen mit Abitur durch die Regeln hervorgehoben. Interessant ist zudem sicherlich die verminderte Wahrscheinlichkeit des Übergangs in die Arbeitslosigkeit bei Untersuchungseinheiten, die in einer Firma mit mehr als 2.000 Mitarbeitern arbeiten.

Die den Analysekomplex zum Thema Arbeitslosigkeit ergänzenden Ergebnisse, die den Übergang von der Arbeitslosigkeit in die Erwerbstätigkeit beschreiben, sind in Tabelle 10.3 abgebildet. Auch diese Ergebnisse bestätigen bzw. verfeinern im wesentlichen die Beobachtungen, die bereits auf Basis der globalen Klassifikationsmodelle gemacht wurden. So ist

Tabelle 10.3: KBS-Regeln für LFSC_Unemployed_Working

Regel	WRAcc	Coverage
(FAMSTD=*verh. zus.*) ∧ (HHCHILDREN=*Nein*) ⇒ *negative*	0,0360	0,3209
(PSBIL=*Hauptschulabschluss*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *negative*	0,0231	0,3837
(FAMSTD=*ledig*) ⇒ *positive*	0,0173	0,2596
(CORIGIN=*Deutschland*) ∧ (FAMSTD=*verh. zus.*) ⇒ *positive*	0,0204	0,4091
(SEX=*weiblich*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0132	0,4349
(PBBIL02=*trifft nicht zu*) ∧ (PBBIL01=*trifft nicht zu*) ⇒ *negative*	0,0117	0,2601
(AGE=*60*) ⇒ *negative*	0,0103	0,0463
(ORTKINDH=*Grossstadt*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0113	0,2252
(ORTKINDH=*Auf dem Lande*) ⇒ *negative*	0,0103	0,3334
(AGE=*59*) ⇒ *negative*	0,0102	0,0447
(WACHSTUM_L2=*niedrig*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *positive*	0,0097	0,1892
(SEX=*maennlich*) ∧ (SUMKIDS=*0*) ⇒ *negative*	0,0089	0,3945
(IMMIGRATED=*nicht immigriert*) ∧ (SEX=*maennlich*) ⇒ *positive*	0,0081	0,4165
(HHGR=*1*) ⇒ *negative*	0,0080	0,1947
(IMMIGRATED=*nicht immigriert*) ∧ (HHGR=*2*) ⇒ *negative*	0,0086	0,3000
(SEX=*weiblich*) ∧ (SUMKIDS=*0*) ⇒ *positive*	0,0076	0,1247
(AGE=*58*) ⇒ *negative*	0,0074	0,0391
(WACHSTUM_L2=*mittel*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *positive*	0,0071	0,3791
(BULA=*Bayern*) ⇒ *positive*	0,0071	0,1121
(LOC1989=*East Germany*) ∧ (FAMSTD=*verh. zus.*) ⇒ *positive*	0,0075	0,1859

etwa ein klarer negativer Alterseffekt (der im Übrigen schon in Abbildung 7.5(e) zu erkennen ist) auf den Übergang von arbeitslos nach erwerbstätig bei Untersuchungseinheiten im Alter von 58 bis 60 Jahren zu verzeichnen. Bestätigt wird außerdem der negative Einfluss geringer Bildung auf den Übergang, etwa durch die in der Zeit nach der Wiedervereinigung erhöhte Wahrscheinlichkeit von Personen mit Hauptschulabschluss, nicht erwerbstätig zu werden. Auch für Personen, die weder einen Hochschulabschluss noch überhaupt einen berufsbildenden Abschluss besitzen, ist eine solche erhöhte Wahrscheinlichkeit zu beobachten. Ungewöhnlicher und damit möglicherweise interessanter als diese Resultate sind die Regeln, die sich auf die familiäre Situation beziehen. Hier ist zum einen etwa das Ergebnis zu nennen, dass Männer ohne eigene Kinder eine verringerte, Frauen ohne eigene Kinder dagegen eine gesteigerte Wahrscheinlichkeit des Übergangs in die Erwerbstätigkeit haben. Abschließend sei auf die Regel hingewiesen, die eine erhöhte Wahrscheinlichkeit des Übergangs für in Bayern lebende Personen und damit einen regionalen Effekt beschreibt.

Zum Vergleich der mittels des Knowledge-Based Sampling erlernten Regelmengen mit den in Kapitel 9.4 evaluierten Globalmodellen hinsichtlich ihrer prädiktiven Performanz wurden auch die Vorhersagegenauigkeiten dieser Regelmengen gemessen. Die Ergebnisse dieser Messung sind in Tabelle 10.4 dargestellt. Wie leicht zu erkennen ist, weichen die Vorhersageperformanzen der Regelmengen nur unwesentlich von denen des Default-Lerners ab und erzielen definitiv keine besseren Werte als die der zuvor evaluierten globalen Modelle. Allerdings ist dies im Angesicht der Tatsache, dass in dieser Arbeit vor allem die Intention des Findens deskriptiver Modelle im Vordergrund steht, eher zweitrangig.

Zusammenfassend ist bzgl. der Intention dieser Arbeit zu bemerken, dass die durch das Knowledge-Based Sampling gelernten Regelmengen sowohl auf intuitiv klare Sachverhalte als auch einige eher ungewöhnliche Auffälligkeiten in den Daten hinweisen.

Tabelle 10.4: Performanz der Regelmengen von KBS. Accuracy auf Trainingsdaten sowie durchschnittliche Accuracy und Standardabweichung der Accuracy bei einer 10-fachen Kreuzvalidierung auf Trainingsdaten. Relative Häufigkeit der Default-Klasse als Referenz.

Label	Default	Training	Kreuzvalidierung	
		Acc.	Avg. Acc.	(Std.dev.)
LFSC_Working_Not-working	97,415%	97,342%	97,329%	(0,138%)
LFSC_Working_Unemployed	96,757%	96,758%	96,757%	(0,133%)
LFSC_Jobbing_Not-working	88,386%	87,575%	87,254%	(1,171%)
LFSC_Jobbing_Unemployed	96,391%	96,274%	96,106%	(0,941%)
LFSC_Unemployed_Working	71,956%	72,080%	71,321%	(1,419%)
LFSC_Unemployed_Jobbing	96,165%	96,165%	96,164%	(0,725%)
LFSC_Training_Working	72,470%	72,549%	72,240%	(1,438%)
LFSC_Training_Jobbing	87,523%	87,523%	87,537%	(1,197%)
LFSC_Parenthood_Working	77,467%	78,103%	77,070%	(3,395%)
LFSC_Parenthood_Jobbing	92,956%	92,956%	92,845%	(1,785%)

11 Zusammenfassung und Fazit

Im Rahmen dieser Arbeit wurden Daten des Sozio-ökonomischen Panels anhand von Verfahren aus den Bereichen Data Mining und maschinelles Lernen in Bezug auf eine beispielhaft gewählte Analyseaufgabe aus dem Themenkomplex der Arbeitslosigkeit analysiert. Die Analyseaufgabe bestand darin, Einflussfaktoren für Übergänge zwischen Arbeitsmarktzuständen zu finden. Um diese Analyseaufgabe erfolgreich bearbeiten zu können, wurden zunächst Daten aus dem umfangreichen Datensatz des SOEP extrahiert und diese gemäß der Vorgaben der späteren Analyseverfahren vorverarbeitet. Dies bildete den ersten Teil dieser Arbeit. Hierzu wurde zunächst die Software PanelX zur Extraktion von Daten aus großen Paneldatensätzen implementiert. Obschon im Rahmen dieser Arbeit im Wesentlichen eine Fokussierung der Implementierung auf die Daten des SOEP erfolgte, so gewährleistet das modulare Design von PanelX doch eine leichte Erweiterbarkeit und Anpassbarkeit auf andere Datensätze sowie andere Ein- und Ausgabeformate der Daten. Zudem garantiert PanelX als einzige dem Autor bekannte Lösung vollkommene Unabhängigkeit von kommerziellen Softwareprodukten bei der Extraktion von Daten, sodass bei Verwendung einer ebenfalls frei verfügbaren Datenanalysesoftware wie etwa RapidMiner sämtliche Analyseschritte ohne die Nutzung kommerzieller Produkte auskommen. Neben der Software zur Extraktion der Paneldaten wurden für die freie Data-Mining-Suite RapidMiner Operatoren zur speziellen Vorverarbeitung von Paneldaten entwickelt, die im sogenannten Panel-Plugin zur Verfügung stehen. In Verbindung mit dem flexiblen Operatorkonzept von RapidMiner, welches eine Vielzahl von Kombinationsmöglichkeiten der Vorverarbeitungsschritte erlaubt, bieten sich damit umfangreiche Möglichkeiten zur Auf- und Vorbereitung insbesondere sozio-ökonomischer Paneldaten für die spätere Analyse. Zu den diesbezüglich durchführbaren Schritten zählen im Einzelnen etwa die Bereinigung von Attributwerten, die Konstruktion neuer Attribute bzw. Wertereihen auf Basis vorliegender Paneldaten, die Transformation und Aggregation von Wertereihen und das Ändern der Repräsentation von Paneldaten.

Für die exemplarische Anwendungsaufgabe der Analyse von Arbeitsmarktzustandsübergängen wurden mit Hilfe der geschaffenen Werkzeuge erfolgreich Daten aus dem SOEP-Datensatz extrahiert und gemäß der Anforderung vorverarbeitet, dass gebräuchliche Datenanalyseverfahren aus dem Bereich des Data Mining bzw. des maschinellen Lernens auf den vorverarbeiteten Daten angewendet werden konnten. Die dazu notwendige Erstellung des Prozesses zur adäquaten Vorbereitung der Daten für die spätere Analyse war trotz der Erleichterung, die mit der Nutzung von PanelX sowie dem Panel-Plugin einherging, ein komplexes und zeitraubendes Unterfangen. Dies ist zum größten Teil dem Fakt geschuldet, dass es sich bei den SOEP-Daten sowohl um reale als auch sehr umfangreiche Daten handelt. Die teilweise unzureichende Datenqualität erforderte beispielsweise in Bezug auf Inkonsistenzen bei der Benennung der Attributwerte umfangreiche Bereinigungsmaßnahmen.

Den zweiten Fokus der Arbeit bildete die Anwendung bislang äußerst selten auf sozio-ökonomischen Daten eingesetzter Datenanalyseverfahren aus dem Gebiet des maschinellen Lernens. Für die Bearbeitung der Analyseaufgabe zur Identifikation von Einflüssen auf Arbeits-

marktzustandsübergänge wurden gängige und im Bereich des Data Mining häufig verwendete Verfahren ausgewählt und auf den Daten angewandt. Nach einer einführenden deskriptiven Analyse der Paneldaten auf Basis einer Aggregation der Daten zu Zeitreihen, wurden auf den vorverarbeiteten Daten somit erstens Verfahren zur Attributgewichtung angewendet, um Einflüsse von Attributen auf die betrachteten Arbeitsmarktzustandsübergänge zu quantifizieren. Zweitens erfolgte danach die Anwendung von Naïve Bayes und eines Entscheidungsbaumlerners. Diese bekannten Klassifikationsverfahren sollten Globalmodelle erstellen, die eine möglichst gute Beschreibung der Gesamtheit der vorgelegten Daten und bestehender, globaler Zusammenhänge liefern. Um vom Globalmodell abweichendes Verhalten von Subgruppen, d.h. lokale Muster, zu entdecken, wurde drittens das Knowledge-Based Sampling zur Subgruppenentdeckung auf den Daten erprobt.

11.1 Interpretation der Ergebnisse

Die aus der Anwendung der genannten Verfahren resultierenden Analyseergebnisse sind auf mehreren Ebenen zu interpretieren. Zum einen sollten die Ergebnisse hinsichtlich ihrer inhaltlichen Aussagen, d.h. ihrer Folgerungen in Bezug auf die Anwendungsdomäne, interpretiert werden. Zum anderen muss die Performanz der Verfahren aus der Perspektive des Data Minings, dem die Verfahren entstammen, bewertet werden. Aus diesen beiden Komponenten können schließlich Schlüsse bezüglich der Anwendbarkeit der hier eingesetzten Verfahren für die betrachtete Analyseaufgaben und damit auch für ähnliche Analyseaufgaben gezogen werden.

Auf vornehmlich inhaltlicher Ebene ergeben sich folgende Schlussfolgerungen: Die durch einfache Aggregation der Paneldaten entstehenden Zeitreihen beschreiben teilweise deutlich sowohl Perioden- als auch Alterseffekte. In Bezug auf den Zeitraum von 1984 bis 2004, den die vorliegenden Daten umfassen, existiert zum einen eine Rückläufigkeit des Wechsels in die Nichterwerbstätigkeit. Daneben ist ein Rückgang ebenfalls bei den Übergängen in die Erwerbstätigkeit zu verzeichnen, demgegenüber steht jedoch eine gegensätzliche, ansteigende Entwicklung bei den Übergängen in die Nebenerwerbstätigkeit. Besonders deutlich ist zudem der scheinbar stark konjunkturabhängige Verlauf des Wechsels von der Erwerbstätigkeit in die Arbeitslosigkeit sowie ein starker Anstieg der Wahrscheinlichkeit, von der Nebenerwerbstätigkeit in die Arbeitslosigkeit überzugehen. In Bezug auf Alterseffekte sind vor allem die erhöhte Wahrscheinlichkeit für den Übergang von erwerbstätig nach arbeitslos bei 55- bis 60-Jährigen zu nennen. Einen starken Einfluss des Alters auf die Übergänge aus der Arbeitslosigkeit belegt zudem die Zeitreihe zum Übergang von arbeitslos nach erwerbstätig, die eine mit zunehmendem Lebensalter stark abnehmende Wahrscheinlichkeit für diesen Übergang angibt. Deutlich erkennbar ist zudem die mit zunehmendem Lebensalter steigende Wahrscheinlichkeit, in den Zustand nichterwerbstätig zu wechseln. Diese Zunahme der Wahrscheinlichkeit beginnt bei Wechseln aus der Erwerbstätigkeit jedoch erst ab einem Alter von etwa 60 Jahren und steigt dann rasant. Die Zunahme der Übergangswahrscheinlichkeit bei Wechseln aus der Nebenerwerbstätigkeit verläuft dagegen in etwa linear zum Lebensalter.

Im Gegensatz zu den beschriebenen Zeitreihen sind die Ergebnisse der Verfahren zur Gewichtung der in die Analyseaufgaben einbezogenen Attribute teilweise wenig deutlich und in ihren Aussagen nicht vollkommen eindeutig. So ist etwa der Informationsgewinn des Attributes, welches das Alter der Personen erfasst, von allen Attributen für die Lernaufgabe

bzgl. des Übergangs von arbeitslos nach erwerbstätig am höchsten. Das ebenfalls verwendete Verfahren Relief gewichtet dieses Attribut jedoch beispielsweise sehr niedrig. Dennoch deutet eine hohe Gewichtung mancher Attribute zumindest einen nicht zu vernachlässigenden Einfluss dieser Attribute an. Insgesamt ist die Aussagekraft der Attributgewichte jedoch begrenzt, da nur der Einfluss einzelner Attribute auf die Zielattribute betrachtet wird. Somit können aus der alleinigen Betrachtung der Attributgewichte keine klaren Schlüsse hinsichtlich der Wichtigkeit der Attribute gezogen werden.

Die Ergebnisse des Klassifikationsverfahrens Naïve Bayes sind insofern eindeutiger und damit aussagekräftiger, als dass sie Einflüsse einiger Attribute auf die Zustandsübergänge als Unterschiede in den bedingten Verteilungen der Attributwerte in den einzelnen Klassen (Durchführung des Zustandsübergangs gegenüber Nichtdurchführung) erkennen lassen. Bezüglich des Übergangs in die Nichterwerbstätigkeit lässt sich aus dem Naïve-Bayes-Modell beispielsweise eine klare Abhängigkeit vom Alter erkennen, die die Erkenntnisse aus der Betrachtung der Aggregatzeitreihen bestätigt. Des Weiteren ist eine erhöhte Wahrscheinlichkeit für den Übergang bei Frauen sowie bei Personen, die wenig autonomen Handlungsspielraum im Beruf besitzen, auszumachen. Auch eine hohe Anzahl leiblicher Kinder wirkt sich verstärkend auf die Wahrscheinlichkeit des Übergangs aus. Beim Übergang von der Erwerbstätigkeit in die Arbeitslosigkeit spielt vor allem die Dauer der Betriebszugehörigkeit eine Rolle: neuere Mitarbeiter werden eher entlassen. Erkennbar ist zudem ein verstärkender Effekt auf die Wahrscheinlichkeit des genannten Übergangs, wenn ein geringeres Wirtschaftswachstum zu verzeichnen ist. Überraschenderweise lässt sich der Einfluss des Konjunkturzyklus mit Hilfe von Naïve Bayes nicht eindeutig nachweisen. Die deutlichsten Ergebnisse finden sich beim Übergang in die Erwerbstätigkeit. Ein positiver Effekt auf den Übergang kann vor allem bei einem geringeren Alter sowie höherer Bildung verzeichnet werden. Des Weiteren ist auffällig, dass sich Kinder im Haushalt positiv auf den Übergang auswirken, eine höhere Anzahl leiblicher Kinder jedoch negativ. Weiterhin ist es für Männer scheinbar leichter, nach Arbeitslosigkeit in die Erwerbstätigkeit überzugehen.

Im Gegensatz zu Naïve Bayes bieten die gelernten Entscheidungsbäume wenig konkrete Anhaltspunkte in Bezug auf Attributeinflüsse. Durch die Anordnung der Attribute, anhand derer die Beispielmenge aufgespalten wurde, lässt sich teilweise allenfalls ein hoher Informationsgewinn dieser Attribute ablesen, wirklich aussagekräftige Modelle resultierten jedoch nicht. Indizien für den Einfluss von Attributen sind lediglich manchen Teilbäumen zu entnehmen, beispielsweise ebenfalls im Übergang von arbeitslos nach erwerbstätig in Bezug auf das Attribut, welches die Hochschulbildung erfasst. Demnach wirkt sich etwa eine abgeschlossene Hochschulbildung positiv auf den genannten Übergang aus. Aufgrund der vorherigen Einschränkung der Beispielmenge in höheren Ebenen des Baumes ist diese Aussage allerdings auf Personen mit einem Höchstalter von 46 Jahren beschränkt und somit eventuell nur partiell gültig. Dies ist jedoch kein Nachteil des Modells. Vielmehr verdeutlicht dies, dass das Finden eines global geltenden Zusammenhangs zwischen Eigenschaften von Untersuchungseinheiten und ihrer Tendenz, Arbeitsmarktzustandsübergänge durchzuführen, nahezu unmöglich ist. Dies motiviert die Bildung von lokalen Modellen, die Zusammenhänge nicht global, d.h. für alle betrachteten Untersuchungseinheiten, sondern für Subgruppen beschreiben.

Eine Bildung von lokalen Modellen erfolgte im Rahmen dieser Arbeit durch Anwendung des Knowledge-Based Sampling zur Subgruppenentdeckung. Die Ergebnisse der mittels des Knowledge-Based Sampling durchgeführten Subgruppenentdeckung bestätigen bzw. konkretisieren manche der zuvor genannten Ergebnisse. Beim Übergang in die Nichter-

werbstätigkeit wird z.B. eine erhöhte Wahrscheinlichkeit eines Übergangs bei 59- bis 63-Jährigen ebenso deutlich erkannt wie eine erhöhte Wahrscheinlichkeit bei Frauen. Bei Übergängen in die Arbeitslosigkeit wird auch das Ergebnis bestätigt, dass Arbeitnehmer mit einer kurzen Betriebszugehörigkeit eine höhere Wahrscheinlichkeit haben, arbeitslos zu werden. Das gleiche gilt für Personen mit geringerer Bildung sowie wenig qualifizierten Arbeitsstellen. Bei dem Übergang von der Arbeitslosigkeit in die Erwerbstätigkeit besteht ebenfalls ein Zusammenhang bzgl. des Alters: ältere Personen haben eine niedrigere Wahrscheinlichkeit, wieder erwerbstätig zu werden. Dies trifft ebenfalls für wenig gebildete Personen zu.

Zusammenfassend ergibt sich hinsichtlich der inhaltlichen Ergebnisse ein ambivalentes Bild. Auf der einen Seite bestätigen und belegen einige Ergebnisse partiell bereits bekannte Tatsachen bzw. unterstützen gängige Vermutungen. Dies legt nahe, dass die angewendeten Verfahren sehr wohl in der Lage sind, Auffälligkeiten und Regelmäßigkeiten in den Daten zu erkennen. Allerdings sind auf der anderen Seite manche Ergebnisse insofern wenig aussagekräftig, als dass sie entweder keine sinnvolle oder viele, d.h. willkürliche Interpretationen, zulassen. Zusammengenommen bieten die Modelle damit lediglich Indizien für bestehende Zusammenhänge. Diesbezüglich unterscheiden sich die Globalmodelle in ihren Aussagen auch wenig von den lokalen Modellen, als dass Modelle beider Typen nur auf eher lokale Auffälligkeiten hinzuweisen scheinen. Somit können die Modelle in ihrer Gesamtheit keine plakativen und unzweifelhaften Beschreibungen der Zusammenhänge liefern.

Diese Schlussfolgerung wird auf einer zweiten Ebene auch von den prädiktiven Performanzen der Verfahren, die bei deren Einsatz gemessen wurden, untermauert. So ist keines der Verfahren herausragend performant und in der Lage, die Daten nahezu vollständig genau zu beschreiben. Vielmehr erfasst kein Modell die positive Klasse in den Lernaufgaben hinreichend gut. So ist die Performanz der Lernverfahren bei einer Kreuzvalidierung in keinem Fall besser als die Performanz des Referenzmodells, welches immer nur die häufigste Klasse vorhersagt. Dies belegt zum einen die Schwierigkeit der Lernaufgaben, zum anderen ist daraus zu folgern, dass kein erlerntes Modell die Daten zumindest im prädiktiven Kontext besser beschreibt als ein Modell, welches nur die häufigste Klasse, d.h. hier stets die keinem Übergang entsprechende negative Klasse, vorhersagt. Für die tatsächliche Beschreibungskraft der einzelnen Modelle im deskriptiven Kontext folgt daraus ihre nicht objektive Messbarkeit. Die deskriptive Akkuratheit der Modelle ist demnach lediglich von Experten der Anwendungsdomäne sicher einschätzbar.

Auf Basis der erfolgten Beobachtungen auf inhaltlicher Ebene sowie auf Ebene der Performanzbewertung der Modelle ist schließlich zu folgern, dass einerseits manche Modelle partiell Anzeichen für bestehende Zusammenhänge aufdecken, etwa die anhand des Knowledge-Based Sampling erlernten Regeln bzw. ansatzweise die von Naïve Bayes errechneten bedingten Verteilungen. Die Modelle sind jedoch keineswegs in der Lage, diese Indizien zu einem schlüssigen, aussagekräftigen und ohne Vorkenntnisse zu interpretierenden, Gesamtbild zu vereinigen. Die Verfahren bleiben damit unter der gewählten Vorgehensweise klar hinter dem Anspruch zurück, aus der großen, wenig eingeschränkten Menge realer, sozio-ökonomischer Daten leicht verwertbares Wissen zu extrahieren.

Dieses Fazit allein ist jedoch nicht hinreichend, um die Anwendbarkeit der Verfahren auf derartigen Daten vollständig in Frage zu stellen. Vielmehr stellt sich die Frage, ob unter der offenen, in dieser Arbeit verfolgten Vorgehensweise und unter Berücksichtigung der dadurch kaum beschränkten Komplexität der SOEP-Daten die dargestellten Resultate nicht bereits ein gutes Ergebnis sind. Zu bedenken ist hierbei, dass für die betrachtete Aufgabenstellung etablierte, ökonometrische Modellierungen häufig mit einer sehr viel restriktiveren Vorge-

hensweise in Bezug auf die Einschränkung der analysierten Daten, etwa hinsichtlich der Selektion der in die Analyse einbezogenen Attribute sowie ihrer Ausprägungen, einhergehen. Trotz des Faktes, dass die erzielten Ergebnisse inhaltlich bislang eher unbefriedigend sind, ist damit eine faire Beurteilung der Leistungsfähigkeit der Verfahren in Bezug auf die Aufgabenstellung noch nicht vollständig möglich. Zur Ermöglichung einer solchen Bewertung der Anwendbarkeit der Verfahren wären Modifikationen des Analyseprozesses etwa in Bezug auf die Offenheit der Vorgehensweise notwendig, um weitere Schlüsse hinsichtlich des Zusammenspiels aus Extraktion, Vorverarbeitung und Analyse der Daten zu ziehen und daraus unter Umständen eine gesicherteres Fazit in Bezug auf das Anwendungspotential der verwendeten Verfahren abzuleiten.

11.2 Mögliche Modifikationen und Erweiterungen

Bei Betrachtung der erzielten Analyseergebnisse spricht vieles dafür, dass die Vorgehensweise in der Gestaltung des Analyseprozesses zu offen war. So erscheinen die Daten durch die geringe Beschränkung in höchstem Maß heterogen, was allem Anschein nach der Grund dafür ist, dass die erlernten Modelle nicht in der Lage sind, wirklich aussagekräftig zu abstrahieren bzw. zu aggregieren. Auf der in dieser Arbeit geschaffenen Grundlage können nicht zuletzt mittels der implementierten Werkzeuge jedoch leicht Modifikationen und Erweiterungen des Analyseprozesses direkt erfolgen, die die offene Vorgehensweise hinsichtlich verschiedener Aspekte einschränken: Einführend könnte erstens etwa eine gezieltere und damit restriktivere Extraktion beispielsweise durch Expertenwissen als relevant bekannter Attribute vorgenommen werden. Alternativ bzw. ergänzend könnte zweitens zudem eine gröbere Einteilung bzw. eine vorherige Aggregation der Attributwerte während der Vorverarbeitung erfolgen. Hiermit würde eine eventuell zu hohe Diversität bei Attributen mit sehr vielen vorkommenden Ausprägungen innerhalb der betrachteten Menge von Untersuchungseinheiten beseitigt. Beispielhaft sei hier etwa auf die Vercodung der Branche, in der Personen erwerbstätig sind, hingewiesen, die etwa im SOEP sehr feingliedrig ist und die aufgrund ihrer hohen Differenziertheit bei der durchgeführten Analyse dazu beigetragen haben könnte, dass die Bildung eines klaren Modells nicht möglich war. Drittens mag sich unter Umständen auch eine dedizierte Diskretisierung numerischer Attribute wie des Alters, wie sie auch bei einigen ökonometrischen Modellierungen teilweise praktiziert wird (siehe beispielsweise Bachmann (2005)), bei manchen Verfahren als vorteilhaft erweisen. Als vierten Punkt soll auch die Konstruktion von Merkmalen, möglichst ebenfalls basierend auf Expertenwissen, vorgeschlagen werden. Denkbar wäre diesbezüglich beispielsweise die Konstruktion eines Attributes, welches die Dauer des vorherigen Zustands in die Erklärung der Zustandsübergänge einbezieht. Letztlich propagieren alle Maßnahmen sowohl eine erhöhte Verzahnung von Datenselektion, Datenvorverarbeitung und Datenanalyse als auch eine vermehrte Rückkoppelung bei iterativer Vorgehensweise innerhalb des Prozesses. Zum Nachteil für den Datenanalysten bedingen die Maßnahmen allerdings eine weitere Intensivierung der Vorverarbeitung und damit eine abermalige Steigerung des bereits in dieser Arbeit dokumentierten hohen Aufwands.

11.3 Ausblick

Neben den in Kapitel 11.2 vorgeschlagenen Modifikationen und Erweiterungen, die speziell versuchen, die Ergebnisse bezüglich der in dieser Arbeit betrachteten Lernaufgaben zu verbessern, könnten weitere Aspekte der Vorgehensweise (bei grundsätzlicher Beibehaltung der Analyseaufgabe) modifiziert werden. Diesbezüglich wäre etwa die Behandlung der Analyseaufgabe unter Verwendung monatlicher Daten zu nennen, die mit dem SOEP zwar möglich, für diese Arbeit jedoch in Kapitel 4.2 ausgeschlossen wurde. Des Weiteren sollten neuere Versionen der SOEP-Daten verwendet werden. So sind zum jetzigen Zeitpunkt Daten bis einschließlich 2006 verfügbar. Diese erweitern den Beobachtungszeitraum somit um zwei weitere Jahre und schließen einen Großteil einer weiteren, konjunkturellen Aufschwungphase ein. Zudem könnten damit etwa auch Effekte der im Jahr 2005 erfolgten Umsetzung der sogenannten Hartz-IV-Reform einbezogen und untersucht werden. Auf methodischer Ebene wäre außerdem ein Vergleich der aus der Anwendung von maschinellen Lernverfahren resultierenden Ergebnisse mit Ergebnissen von Verfahren, die bereits für diese Anwendungsdomäne etabliert sind, interessant. In Bezug auf die Ergebnisse dieser Arbeit entspräche dies etwa einem Vergleich mit Ergebnissen der logistischen Regression in Verbindung mit geeigneten ökonometrischen Paneldatenmodellen.

Eine Möglichkeit, die in dieser Arbeit durchgeführten Analyseaufgaben inhaltlich zu ergänzen, wäre eine Beschreibung bzw. Vorhersage der Verweildauern in Zuständen mit maschinellen Lernverfahren. Dies ist etwa als Gegenstück zu der ökonometrischen Modellierung von Verweildauern und Bestimmung von Einflüssen auf diese durch die Cox-Regression anzusehen.

In Bezug auf die Paneldatenanalyse wäre zudem eine Anwendung aus dem Bereich des Data Mining stammender Verfahren interessant, die speziell zur Analyse zeitlicher Daten entwickelt wurden. Hier ist etwa die Erkennung sequentieller Muster zu nennen. Eine sozio-ökonomische Anwendung dessen wäre beispielsweise die Erkennung prototypischer Lebensverläufe.

A Ergebnisse des Knowledge-Based Sampling

Als Ergänzung zu den ausführlichen Betrachtungen in Kapitel 10.5 dokumentieren die folgenden Tabellen die bislang nicht vorgestellten Ergebnisse der Anwendung des Knowledge-Based Samplings für die restlichen Lernaufgaben.

Tabelle A.1: KBS-Regeln für `LFSC_Jobbing_Not-working`

Regel	WRAcc	Coverage
(`FAMSTD`=*ledig*) ⇒ *negative*	0,0381	0,4117
(`IMMIGRATED`=*nicht immigr.*) ∧ (`PSBIL`=*Hauptschulabschl.*) ⇒ *positive*	0,0220	0,3628
(`SEX`=*weiblich*) ∧ (`LOC1989`=*West Germany*) ⇒ *positive*	0,0138	0,4515
(`WACHSTUM_L1`=*mittel*) ∧ (`ORTKINDH`=*Auf dem Lande*) ⇒ *positive*	0,0121	0,1704
(`WACHSTUM_L2`=*mittel*) ∧ (`FAMSTD`=*ledig*) ⇒ *positive*	0,0118	0,2145
(`SEX`=*weiblich*) ∧ (`FAMSTD`=*verh. zus.*) ⇒ *positive*	0,0110	0,2742
(`SEX`=*maennlich*) ∧ (`HHGR`=*1*) ⇒ *positive*	0,0110	0,0895
(`IMMIGRATED`=*nicht immigriert*) ∧ (`PBBIL01`=*trifft nicht zu*) ⇒ *negative*	0,0092	0,4405
(`IMMIGRATED`=*nicht immigriert*) ∧ (`HHGR`=*3*) ⇒ *negative*	0,0081	0,2093
(`IMMIGRATED`=*nicht immigriert*) ∧ (`SUMKIDS`=*0*) ⇒ *negative*	0,0076	0,5547
(`INFLATION`=*niedrig*) ∧ (`PBBIL02`=*trifft nicht zu*) ⇒ *negative*	0,0086	0,4316
(`WACHSTUM_L2`=*mittel*) ∧ (`WACHSTUM_L3`=*mittel*) ⇒ *positve*	0,0080	0,2697
(`LOC1989`=*West Germany*) ∧ (`ZYKLUS`=*Abschwung*) ⇒ *positive*	0,0076	0,1655
(`ORTKINDH`=*Mittlere Stadt*) ⇒ *positive*	0,0078	0,1492
(`AGE`=*62*) ⇒ *positive*	0,0071	0,0243
(`SEX`=*weiblich*) ∧ (`ORTKINDH`=*Grossstadt*) ⇒ *positive*	0,0066	0,0976
(`AGE`=*61*) ⇒ *positive*	0,0059	0,0222
(`LOC1989`=*West Germany*) ∧ (`HHCHILDREN`=*Ja*) ⇒ *positive*	0,0071	0,2369
(`AGE`=*63*) ⇒ *positive*	0,0066	0,0195

Tabelle A.2: KBS-Regeln für LFSC_Jobbing_Unemployed

Regel	WRAcc	Coverage
(IMMIGRATED=*nicht immigriert*) ∧ (LOC1989=*West Germany*) ⇒ *negative*	0,0115	0,7644
(PBBIL01=*Lehre*) ⇒ *positive*	0,0066	0,3467
(WACHSTUM_L1=*mittel*) ∧ (CORIGIN=*Deutschland*) ⇒ *negative*	0,0068	0,4365
(WACHSTUM_L3=*mittel*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *positive*	0,0047	0,4283
(PSBIL=*Abitur*) ⇒ *negative*	0,0042	0,2112
(WACHSTUM_L2=*mittel*) ∧ (SEX=*maennlich*) ⇒ *positive*	0,0046	0,2284
(WACHSTUM_L2=*mittel*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0038	0,4550
(SEX=*weiblich*) ∧ (ORTKINDH=*Auf dem Lande*) ⇒ *negative*	0,0035	0,2063
(LEGISLATUR=*14*) ⇒ *positive*	0,0036	0,2750
(AGE=*57*) ⇒ *positive*	0,0032	0,0128
(SEX=*weiblich*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *positive*	0,0030	0,4416
(ORTKINDH=*Grossstadt*) ⇒ *negative*	0,0031	0,1872
(WACHSTUM_L1=*sehr niedrig*) ⇒ *positive*	0,0030	0,1330
(LOC1989=*West Germany*) ∧ (HHCHILDREN=*Nein*) ⇒ *negative*	0,0031	0,5765
(WACHSTUM_L2=*mittel*) ∧ (TYPHH=*Paar+2 K. GT 16*) ⇒ *positive*	0,0032	0,0447
(WACHSTUM=*mittel*) ∧ (HHCHILDREN=*Nein*) ⇒ *negative*	0,0035	0,3063
(IMMIGRATED=*nicht immigriert*) ∧ (PBBIL01=*Lehre*) ⇒ *positive*	0,0036	0,3364
(BILZEIT=*9*) ⇒ *positive*	0,0035	0,1371
(IMMIGRATED=*nicht immigriert*) ∧ (BULA=*Bayern*) ⇒ *negative*	0,0030	0,1469
(WACHSTUM_L2=*mittel*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0031	0,4541

Tabelle A.3: KBS-Regeln für LFSC_Unemployed_Jobbing

Regel	WRAcc	Coverage
(IMMIGRATED=*nicht immigriert*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *positive*	0,0051	0,6704
(LEGISLATUR=*14*) ⇒ *posivite*	0,0031	0,2256
(FAMSTD=*ledig*) ⇒ *negative*	0,0023	0,2639
(ORTKINDH=*Auf dem Lande*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *negative*	0,0024	0,2702
(WACHSTUM_L1=*mittel*) ∧ (ORTKINDH=*Auf dem Lande*) ⇒ *positive*	0,0025	0,1665
(INFLATION=*niedrig*) ∧ (SEX=*maennlich*) ⇒ *positive*	0,0022	0,2560
(PSBIL=*Abitur*) ⇒ *positive*	0,0019	0,0744
(PBBIL01=*Lehre*) ⇒ *positive*	0,0021	0,5067
(WACHSTUM_L3=*mittel*) ∧ (PBBIL01=*trifft nicht zu*) ⇒ *positive*	0,0022	0,1496
(ORTKINDH=*Grossstadt*) ∧ (LOC1989=*West Germany*) ⇒ *negative*	0,0019	0,1765
(IMMIGRATED=*nicht immigr.*) ∧ (BULA=*Nordrhein-Westfalen*) ⇒ *positive*	0,0019	0,1540
(BULA=*Niedersachsen*) ⇒ *positive*	0,0016	0,0907
(ORTKIND1=*Ja, immer noch*) ∧ (BILZEIT=*9*) ⇒ *positive*	0,0016	0,0645
(PBBIL01=*Lehre*) ⇒ *positive*	0,0023	0,5069
(ORTKINDH=*Grossstadt*) ∧ (ORTKIND1=*Ja, immer noch*) ⇒ *negative*	0,0017	0,1583
(ORTKINDH=*Auf dem Lande*) ∧ (FAMSTD=*verh. zus.*) ⇒ *negative*	0,0017	0,2137
(LOC1989=*West Germany*) ∧ (FAMSTD=*ledig*) ⇒ *negative*	0,0020	0,1822
(WACHSTUM_L3=*sehr hoch*) ∧ (ORTKIND1=*Ja, immer noch*) ⇒ *positive*	0,0017	0,0616
(WACHSTUM_L3=*niedrig*) ∧ (FAMSTD=*ledig*) ⇒ *positive*	0,0019	0,0550
(IMMIGRATED=*nicht immigriert*) ∧ (ZYKLUS=*Abschwung*) ⇒ *negative*	0,0017	0,1760

Tabelle A.4: KBS-Regeln für LFSC_Training_Working

Regel	WRAcc	Coverage
(PSBIL=*Abitur*) ⇒ *negative*	0,0140	0,2662
(LOC1989=*West Germany*) ∧ (WIEDERVEREININGUNG=*Ja*) ⇒ *negative*	0,0110	0,4930
(LOC1989=*West Germany*)∧(PSBIL=*Ohne Abschl. verlassen*) ⇒ *negative*	0,0107	0,2056
(BILZEIT=*0*) ⇒ *negative*	0,0101	0,1593
(PSBIL=*Hauptschulabschluss*) ⇒ *positive*	0,0074	0,1005
(ORTKINDH=*Auf dem Lande*) ∧ (FAMSTD=*ledig*) ⇒ *positive*	0,0060	0,1573
(ORTKIND1=*Ja, immer noch*) ∧ (PBBIL01=*trifft nicht zu*) ⇒ *negative*	0,0070	0,4251
(WACHSTUM_L3=*mittel*) ∧ (LOC1989=*West Germany*) ⇒ *positive*	0,0054	0,3556
(LOC1989=*West Germany*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0056	0,7900
(SEX=*maennlich*) ∧ (HHGR=*4*) ⇒ *negative*	0,0048	0,1613
(HHGR=*3*) ⇒ *negative*	0,0052	0,2824
(HHGR=*5*) ⇒ *negative*	0,0050	0,1279
(IMMIGRATED=*nicht immigriert*) ∧ (AGE=*17*) ⇒ *positive*	0,0052	0,2176
(WACHSTUM_L3=*mittel*) ∧ (SEX=*weiblich*) ⇒ *negative*	0,0050	0,2174
(SEX=*weiblich*) ∧ (AGE=*20*) ⇒ *positive*	0,0048	0,0397
(LOC1989=*West Germany*) ∧ (AGE=*18*) ⇒ *positive*	0,0049	0,1253
(IMMIGRATED=*nicht im.*) ∧ (TYPHH=*Paar+2 K. LE u. GT 16*) ⇒ *negative*	0,0043	0,1094
(SEX=*weiblich*) ∧ (HHCHILDREN=*Nein*) ⇒ *negative*	0,0052	0,3083
(WACHSTUM_L1=*mittel*) ∧ (SEX=*weiblich*) ⇒ *positive*	0,0047	0,2230
(WACHSTUM=*niedrig*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0040	0,1353

Tabelle A.5: KBS-Regeln für LFSC_Training_Jobbing

Regel	WRAcc	Coverage
(IMMIGRATED=*nicht immigriert*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *positive*	0,0111	0,594
(SEX=*maennlich*) ∧ (SUMKIDS=*0*) ⇒ *positive*	0,0082	0,4998
(WACHSTUM_L3=*mittel*) ∧ (FAMSTD=*ledig*) ⇒ *positive*	0,0081	0,3960
(ORTKIND1=*Ja, immer noch*) ∧ (SUMKIDS=*0*) ⇒ *positive*	0,0075	0,4897
(INFLATION=*mittel*) ∧ (ZYKLUS=*Aufschwung*) ⇒ *positive*	0,0067	0,1547
(WACHSTUM_L1=*mittel*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *positive*	0,0067	0,2965
(PSBIL=*Abitur*) ⇒ *positive*	0,0054	0,2657
(WACHSTUM_L1=*mittel*) ∧ (HHCHILDREN=*Nein*) ⇒ *negative*	0,0054	0,2960
(PSBIL=*Hauptschulabschluss*) ⇒ *negative*	0,0039	0,1023
(WACHSTUM=*mittel*) ∧ (ORTKIND1=*Nein*) ⇒ *positive*	0,0037	0,0659
(IMMIGRATED=*nicht immigriert*) ∧ (FAMSTD=*ledig*) ⇒ *positive*	0,0037	0,8336
(WACHSTUM=*mittel*) ∧ (HHCHILDREN=*Nein*) ⇒ *negative*	0,0035	0,3016
(INFLATION=*niedrig*) ∧ (HHCHILDREN=*Nein*) ⇒ *positive*	0,0037	0,2689
(WACHSTUM_L3=*niedrig*) ∧ (FAMSTD=*ledig*) ⇒ *positive*	0,0036	0,1839
(SEX=*maennlich*) ∧ (WIEDERVEREINIGUNG=*Nein*) ⇒ *negative*	0,0034	0,1847
(WACHSTUM_L2=*mittel*) ∧ (PSBIL=*Abitur*) ⇒ *negative*	0,0033	0,1075
(WACHSTUM_L1=*hoch*) ∧ (FAMSTD=*ledig*) ⇒ *positive*	0,0030	0,1279
(IMMIGRATED=*nicht immigriert*) ∧ (ZYKLUS=*Wendepunkt*) ⇒ *negative*	0,0038	0,1840
(WACHSTUM_L2=*mittel*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *positive*	0,0034	0,2395
(AGE=*22*) ⇒ *positive*	0,0027	0,0502

Tabelle A.6: KBS-Regeln für LFSC_Parenthood_Working

Regel	WRAcc	Coverage
(SEX=*weiblich*) ∧ (Legislatur=*13*) ⇒ *negative*	0,0188	0,3098
(WACHSTUM_L2=*mittel*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0178	0,4135
(INFLATION=*sehr niedrig*) ∧ (CORIGIN=*Deutschland*) ⇒ *positive*	0,0157	0,3983
(ORTKIND1=*Nein*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0135	0,2781
(ORTKINDH=*Auf dem Lande*) ∧ (PBBIL02=*trifft nicht zu*) ⇒ *negative*	0,0116	0,2904
(LOC1989=*West Germany*) ∧ (FAMSTD=*verh. zus.*) ⇒ *negative*	0,0106	0,5781
(WACHSTUM=*sehr niedrig*) ∧ (ZYKLUS=*Abschwung*) ⇒ *negative*	0,0101	0,1464
(PSBIL=*Abitur*) ⇒ *positive*	0,0097	0,1733
(SEX=*weiblich*) ∧ (PSBIL=*Realschulabschluss*) ⇒ *positive*	0,0112	0,4620
(INFLATION=*sehr niedrig*) ∧ (SUMKIDS=*2*) ⇒ *positive*	0,0095	0,1575
(HHGR=*3*) ⇒ *positive*	0,0112	0,4466
(ORTKINDH=*Kleinstadt*) ∧ (ORTKIND1=*Ja, immer noch*) ⇒ *negative*	0,0089	0,1078
(WACHSTUM_L1=*mittel*) ∧ (SEX=*weiblich*) ⇒ *negative*	0,0096	0,4511
(WACHSTUM=*mittel*) ∧ (HHCHILDREN=*Ja*) ⇒ *positive*	0,0081	0,4258
(WACHSTUM_L3=*mittel*) ∧ (PBBIL01=*trifft nicht zu*) ⇒ *negative*	0,0070	0,1337
(WACHSTUM_L2=*hoch*) ∧ (ORTKIND1=*Nein*) ⇒ *positive*	0,0068	0,0606
(ORTKIND1=*Nein*) ∧ (HHCHILDREN=*Ja*) ⇒ *negative*	0,0097	0,3223
(ORTKIND1=*Ja, immer noch*) ∧ (PBBIL01=*Lehre*) ⇒ *negative*	0,0082	0,2613
(ORTKINDH=*Kleinstadt*) ∧ (AGE=*29*) ⇒ *positive*	0,0068	0,0181
(WACHSTUM_L3=*hoch*) ∧ (FAMSTD=*verh. zus.*) ⇒ *negative*	0,0062	0,1680

Tabelle A.7: KBS-Regeln für LFSC_Parenthood_Jobbing

Regel	WRAcc	Coverage
(INFLATION=*niedrig*) ∧ (LOC1989=*West Germany*) ⇒ *positive*	0,0156	0,1894
(IMMIGRATED=*nicht immigr.*) ∧ (LOC1989=*West Germany*) ⇒ *positive*	0,0087	0,6298
(WACHSTUM_L2=*mittel*) ∧ (ORTKIND1=*Nein*) ⇒ *positive*	0,0063	0,1588
(ORTKIND1=*Ja*) ∧ (WIEDERVEREINIGUNG=*Ja*) ⇒ *positive*	0,0069	0,4621
(HHGR=*3*) ⇒ *positive*	0,0057	0,4427
(IMMIGRATED=*nicht immigriert*) ∧ (SUMKIDS=*1*) ⇒ *negative*	0,0053	0,4160
(WACHSTUM=*sehr niedrig*) ∧ (CORIGIN=*Deutschland*) ⇒ *positive*	0,0053	0,1929
(WACHSTUM=*mittel*) ∧ (FAMSTD=*verh. zus.*) ⇒ *positive*	0,0065	0,3590
(WACHSTUM_L1=*mittel*) ∧ (BILZEIT=*12*) ⇒ *positive*	0,0055	0,1827
(LOC1989=*West Germany*) ∧ (LEGISLATUR=*13*) ⇒ *negative*	0,0054	0,2272
(WACHSTUM_L2=*mittel*) ∧ (CORIGIN=*Deutschland*) ⇒ *positive*	0,0051	0,3974
(ORTKINDH=*Mittlere*) ∧ (Stadt=*FAMSTD*) ⇒ *verh. zus.*	0,0053	0,1337
(WACHSTUM_L3=*mittel*) ∧ (HHGR=*3*) ⇒ *positive*	0,0051	0,2116
(IMMIGRATED=*nicht immigriert*)∧(PBBIL01=*Berufsfachschule*) ⇒ *positive*	0,0040	0,379
(WACHSTUM=*mittel*) ∧ (LOC1989=*West Germany*) ⇒ *negative*	0,0038	0,3239
(WACHSTUM=*mittel*) ∧ (FAMSTD=*verh. zus.*) ⇒ *positive*	0,0038	0,3599
(WACHSTUM=*mittel*) ∧ (CORIGIN=*Deutschland*) ⇒ *negative*	0,0041	0,3970
(WACHSTUM=*mittel*) ∧ (TYPHH=*Paar+1 K. LE 16*) ⇒ *positive*	0,0039	0,1872
(SUMKIDS=*1*) ⇒ *negative*	0,0039	0,4708
(ORTKINDH=*Grossstadt*) ∧ (SUMKIDS=*1*) ⇒ *positive*	0,0041	0,0980

B RapidMiner-Darstellung des Vorverarbeitungsprozesses

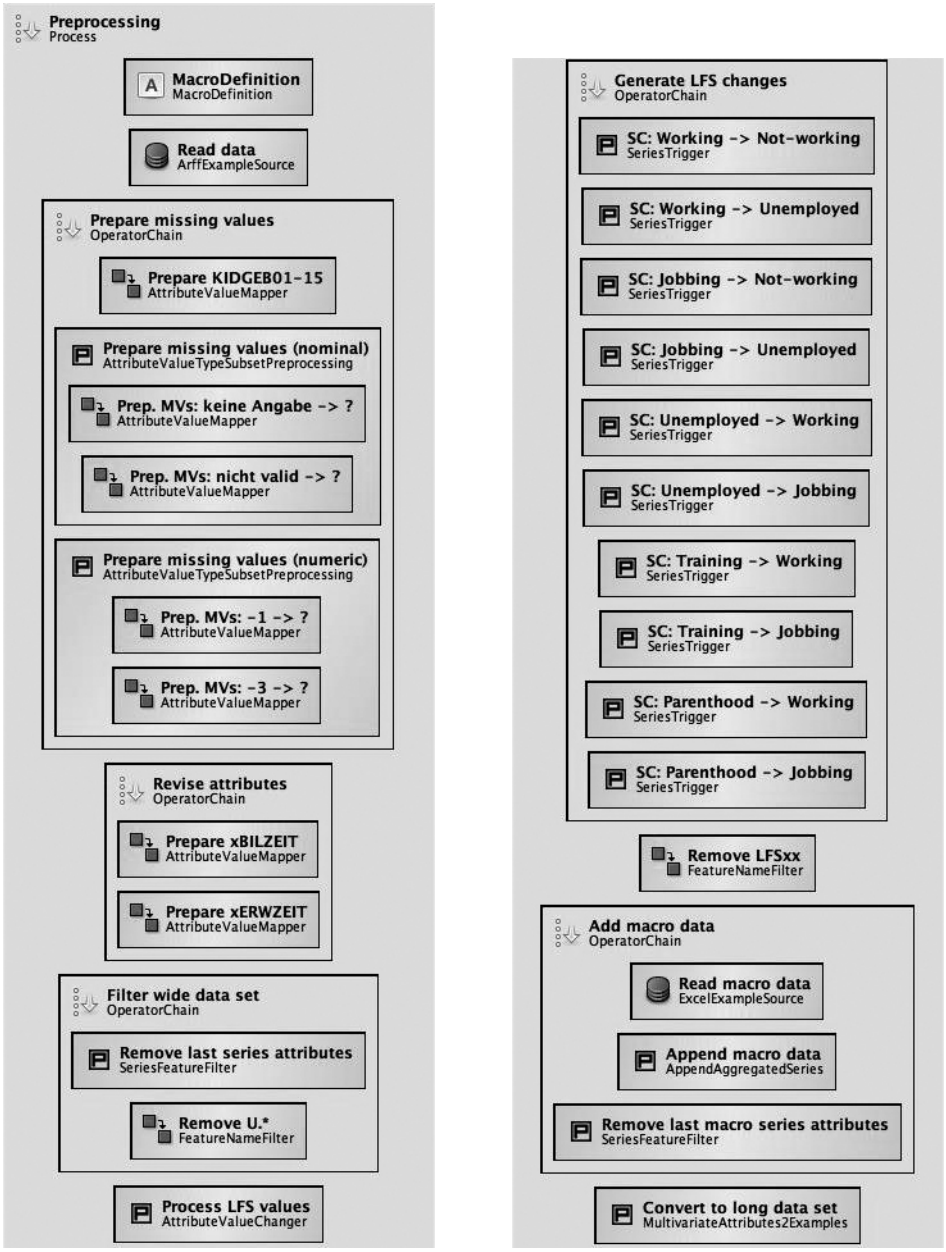


Abbildung B.1: RapidMiner-Darstellung des Vorverarbeitungsprozesses (Teil 1)

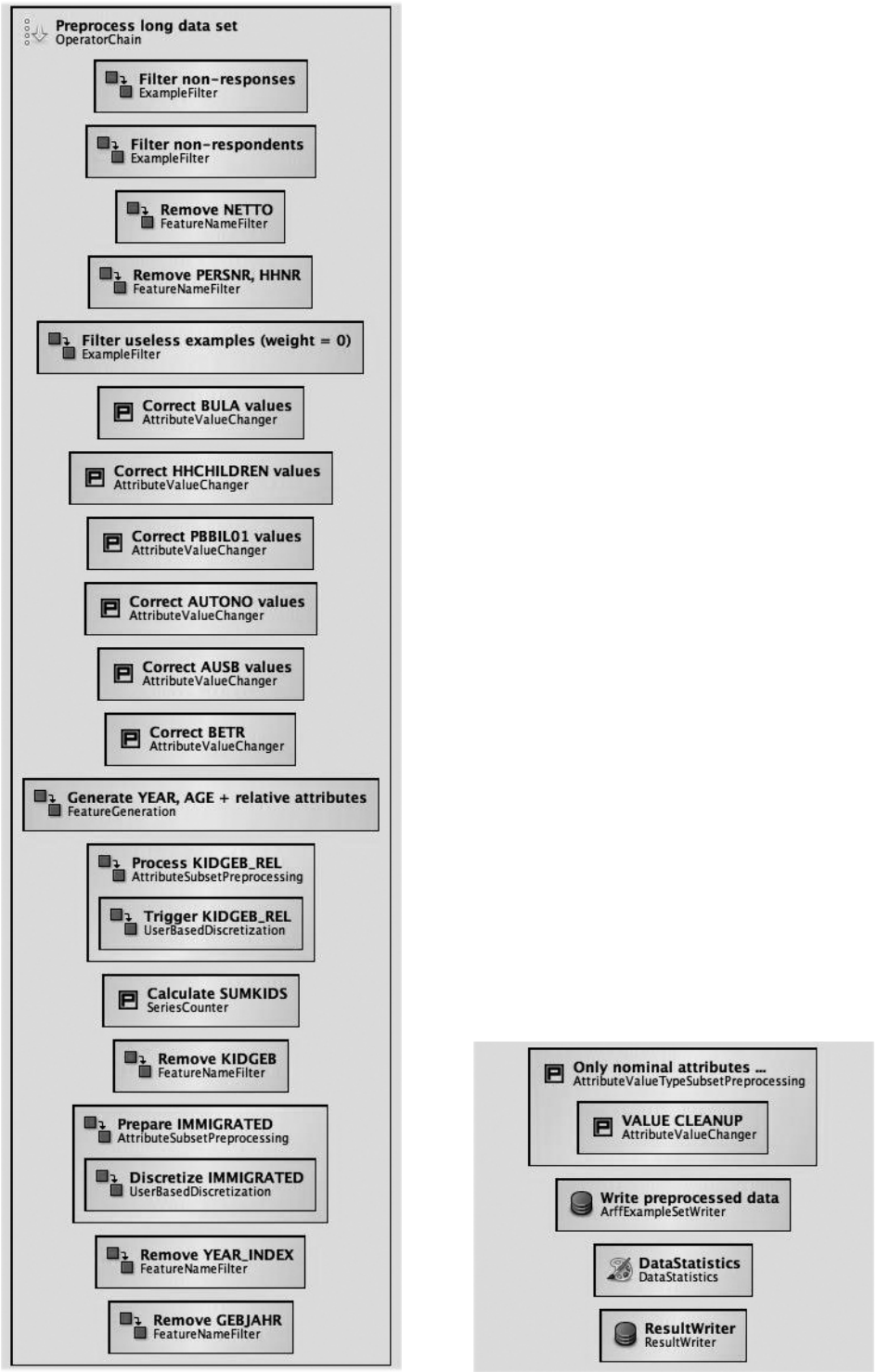


Abbildung B.2: RapidMiner-Darstellung des Vorverarbeitungsprozesses (Teil 2)

C Panel-Plugin für RapidMiner

Im Folgenden erfolgt eine kurze Beschreibung der Operatoren für RapidMiner, die als Teil des Panel-Plugins im Rahmen dieser Arbeit entwickelt wurden. Diese Beschreibung resümiert zum einen kurz die Funktion der Operatoren und dient als Referenz zu deren Nutzung, indem die Parameter der Operatoren erläutert werden.

`AppendAggregatedSeries`

Verknüpft zwei ExampleSets gemäß der Transformation in Abbildung 6.1. Das erste (vorher zu ladende) ExampleSet beinhaltet Mikrodaten, das zweite die aggregierten Makrodaten, die für jede Untersuchungseinheit des ersten ExampleSets gelten sollen (vgl. auch die Beschreibung der Transformation in Kapitel 6.2 zum genauen Aufbau der ExampleSets, die Eingabe des Operators sind).

`AttributeValueChanger`

Erlaubt analog zum Operator `AttributeValueMapper` die Abbildung von Werten auf andere Werte in Attributen. Über den Parameter attributes müssen diese Attribute spezifiziert werden. Hierbei sind reguläre Ausdrücke zur Angabe mehrerer Attribute erlaubt. Die alten sowie neuen Werte, auf die die alten Werte abgebildet werden sollen, sind in der Parameterliste value_mappings anzugeben.

`ConditionedFeatureGeneration`

Erzeugt ein Attribut und weist diesem Attribut Werte nach dem Zutreffen gegebener Bedingungen zu. Der Name des zu erzeugenden Attributs muss durch den Parameter attribute_name gegeben sein. Der Parameter value_type gibt den Typ des resultierenden Attributs an. Alle standardmäßig in RapidMiner verfügbaren Attributtypen sind hier wählbar. In der Parameterliste values können Werte angegeben werden, die bei Zutreffen der zugehörigen über den Parameter conditions gegebenen Bedingungen dem Attribut zugewiesen werden. Die Bedingungen werden für jedes Beispiel der Reihe nach überprüft. Der Wert, der zu der ersten zutreffenden Bedingung gehört wird dem Attribut zugewiesen. Trifft keine der Bedingungen zu, dann wird dem Attribut der Wert zugewiesen, der anhand des Parameter default_value spezifiziert wird.

`Examples2Attributes`

Konvertiert ein ExampleSet in Long-Format in das Wide-Format. Der Parameter id_attribute gibt das Attribut an, welches die Untersuchungseinheiten indiziert. Der Parameter index_attribute spezifiziert das Attribut, welches die Zeitpunkte erfasst, zu denen die Beispiele im ursprünglichen ExampleSet gehören.

`IntegerDiscretization`

Erlaubt die Diskretisierung numerischer Attribute derart, dass die numerischen Werte auf ihren ganzzahligen Anteil abgebildet oder auf eine ganze Zahl gerundet werden (per Parameter mode als Optionen cut bzw. round einstellbar). Über den Parameter attribute_type ist anzugeben, ob das resultierende Attribut vom Typ integer oder nominal sein soll.

`MultivariateAttributes2Examples`

Konvertiert ein ExampleSet in Wide-Format in das Long-Format. Über die Parameterliste series werden dem Operator einerseits die Wertereihen durch Spezifikation anhand regulärer Ausdrücke im Parameter attributes bekannt gemacht, andererseits werden diesen Wertereihen Namen zugewiesen, die im resultierenden ExampleSet in Long-Format als Attributnamen für die konvertierten Wertereihen dienen. Die angegebenen Wertereihen müssen alle die gleiche Länge haben, die regulären Ausdrücke also jeweils gleich viele Attribute erfassen. Weiterhin ist im Parameter index_attribute ein Name für das Attribut im resultierenden ExampleSet anzugeben, welches die Zeitpunkte angibt, für die das jeweilige Beispiel im Long-Format zutrifft. Der boolsche Parameter keep_missings gibt an, ob Zeitpunkte, die in den multivariaten Wertereihen des Wide-Datensatzes nur fehlende Werte besaßen, noch im resultierenden ExampleSet vorkommen oder aber gelöscht werden sollen.

`MultivariateSeriesAdjustment`

Richtet die als multivariate Wertereihen gegebenen Beobachtungen der einzelnen Untersuchungseinheiten an einem Ereigniszeitpunkt aus. Die auszurichtenden Wertereihen sind über die Parameterliste series zu spezifizieren. Ihnen ist ein Name zu geben, der für die resultierenden, ausgerichteten Wertereihen verwendet wird. Des Weiteren müssen die zu den ursprünglichen Wertereihen gehörenden Attribute als reguläre Ausdrücke im Parameter attributes angegeben werden. Der Parameter adjustment_index spezifiziert das Attribut, welches für jede Untersuchungseinheit den Zeitpunkt des Ereignisses, an dem die Beobachtungen ausgerichtet werden sollen, angibt. Als Parameter series_start_index ist ein Offset anzugeben, der den Zeitpunkt der ersten Beobachtung der ursprünglichen Wertereihen indiziert. Der Parameter unknown_value gibt den zu verwendenden fehlenden Wert an. Der Parameter remove_original_series gibt an, ob die ursprünglichen Wertereihen aus dem ExampleSet entfernt werden sollen.

`MultivariateSeriesAggregation`

Aggregiert eine multivariate Wertereihe eines ExampleSets (im Wide-Format). Die Wertereihe muss als regulärer Ausdruck im Parameter attributes gegeben sein. Über den Parameter aggregation_type erlaubt der Operator die Einstellung der Verhaltensweise: es können entweder die absoluten Häufigkeiten (Option frequencies) oder die relativen Häufigkeiten (Option frequency ratios) aller zu einem Zeitpunkt vorkommenden Werte berechnet werden. Alternativ können bei Wahl der Optionen value bzw. value ratio nur gezielt die absoluten bzw. relativen Häufigkeiten des über den Parameter value zu spezifizierenden Wertes berechnet werden. Bei Wahl der Optionen unknown bzw. unknown ratio können stattdessen die absoluten bzw. relativen Häufigkeiten fehlender Werte gezählt werden. Der Parameter use_weights gibt an, ob Querschnittsgewichte verwendet werden sollen. Die Attribute,

die die Querschnittsgewichte enthalten, müssen in diesem Fall über einen regulären Ausdruck im Parameter weights spezifiziert werden (die Anzahl der Attribute muss dabei mit der Anzahl der Attribute der obig verwendeten Wertereihe übereinstimmen). Der Parameter exclude_missings gibt an, ob bei Berechnung relativer Häufigkeiten fehlende Werte in die Berechnung der relativen Anteile der Werte einbezogen werden sollen oder nicht. Der Parameter take_percentage erlaubt, die relativen Häufigkeiten wahlweise als Prozentzahlen ausgeben zu lassen.

`SeriesCounter`

Ermöglicht die Konstruktion von Attributen, die Vorkommnisse von spezifizierten Werten bzw. Wertwechseln in multivariaten Wertereihen zählen. Der Parameter counter_mode steuert die diesbezügliche Verhaltensweise. Die Option occurance bedeutet die Zählung auftretender Werte, die Option change die Zählung auftretender Wertwechsel. Die Parameter param_1 sowie param_2 spezifizieren die zu zählenden Attributwerte bzw. Ausgangs- und Zielwert des zu zählenden Attributwertwechsels. Der Parameter attributes erwartet die Angabe eines regulären Ausdrucks, der die Attribute der Wertereihe spezifiziert. Im Parameter counter_name muss die Angabe des Namens für das resultierende Attribut für den Zähler erfolgen.

`SeriesFeatureFilter`

Entfernt Attribute aus Wertereihen aus einem ExampleSet (in Wide-Format). Die Parameterliste series_attributes erlaubt die Angabe von Wertereihen über reguläre Ausdrücke und über den Parameter remove zudem für jede Wertereihe die Angabe, welche Attribute der jeweiligen Wertereihe gelöscht werden sollen. Alternativ können alle Attribute (Option all) der Wertereihe, die ersten (Option first) oder letzten (Option last) Attribute der Wertereihe entfernt werden. Die Anzahl der Attribute, die bei Wahl einer der letzten beiden Optionen am Anfang bzw. Ende der Wertereihe entfernt werden sollen, kann über den (allerdings für alle Wertereihen geltenden) Parameter remove_count spezifiziert werden.

`SeriesNameChanger`

Bennent die Attribute einer Wertereihe eines ExampleSets (in Wide-Format) um. Die Attribute der Wertereihe können über den Parameter attributes durch einen regulären Ausdruck spezifiziert werden. Die Angabe des neuen Namens der Wertereihe erfolgt durch den Parameter new_series_name. Die Attribute erhalten dann den so spezifizierten Namen kombiniert mit ihrer laufenden Nummern innerhalb der Wertereihe.

`SeriesTrigger`

Erzeugt binominale Attribute, die angeben, ob Werte bzw. Wertwechsel in einer multivariaten Wertereihe vorhanden sind. Der Parameter trigger_mode bestimmt durch Wahl der alternativen Optionen occurance oder change, ob Werte oder Wertwechsel getriggert werden sollen. Der Parameter exhaustive gibt an, ob nur ein Attribut erzeugt werden soll, welches erfasst, ob der Wert bzw. der Wertwechsel in der Wertereihe mindestens einmal vorkommt, oder ob mehrere Attribute erzeugt werden sollen, die für jeden Zeitpunkt bzw. Wechsel der Wertereihe diese Information erfassen. Der Parameter trigger_name bestimmt den Namen

des resultierenden Attributes bzw. der resultierenden Wertereihe. Der Parameter attributes erwartet die Angabe der Attribute der Wertereihe als regulären Ausdruck. Die Parameter param_1 und param_2 spezifizieren die als positiv zu triggernden Werte bzw. den Wertwechsel. Der Parameter negative_changes erlaubt die Spezifikation, welche Fälle von Attributwechseln als negativ getriggert werden sollen. Wählbar sind die Optionen all (alle Fälle außer dem spezifizierten Wechsel), conditional (alle Fälle, in denen der Ausgangswert mit dem spezifizierten übereinstimmt, der Zielwert allerdings nicht) oder no_change (die Fälle wie bei conditional, bei denen zusätzlich der Ausgangs- mit dem Zielwert übereinstimmt). Der Parameter invert bestimmt ein invertieren der getriggerten Bedingung, der Parameter delete_old führt zu einem Löschen der untersuchten Wertereihe.

Literaturverzeichnis

Adriaans, P. und Zantinge, D. (1996). *Data Mining*. Addison Wesley, Harlow.

Althoff, S. (1993). *Auswahlverfahren in der Markt-, Meinungs- und empirischen Sozialforschung*. Nr. 19 in Reihe Sozialwissenschaften. Centaurus-Verlagsgesellschaft, Pfaffenweiler.

Bachmann, R. (2005). Labour Market Dynamics in Germany: Hirings, Separations, and Job-to-Job Transitions over the Business Cycle. Discussion Paper 2005-045, SFB 649: Economic Risk, Berlin.

Baltagi, B. H. (2003). *Econometric Analysis of Panel Data*. John Wiley & Sons, 2. Auflage.

Batista, G. E. A. P. A. und Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. In: Abraham, A., del Solar, J. R., und Köppen, M. (Hrsg.): *Soft Computing Systems - Design, Management and Applications*, Band 87 der Reihe *Frontiers in Artificial Intelligence and Applications*, Seiten 251–260. IOS Press.

Baßeler, U., Heinrich, J., und Utecht, B. (2002). *Grundlagen und Probleme der Volkswirtschaft*. Schäffer-Poeschel, Stuttgart, 17., überarbeitete Auflage.

Biewen, M. und Wilke, R. A. (2005). Unemployment Duration and the Length of Entitlement Periods for Unemployment Benefits: Do the IAB Employment Subsample and the German Socio-Economic Panel Yield the Same Results? Discussion Paper 05-05, ZEW Mannheim.

Breiman, L., Friedman, J. H., Olshen, R. A., und Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.

Burkhauser, R. V., Butrica, B. A., Daly, M. C., und Lillard, D. R. (2000). The Cross-National Equivalent File: A product of cross-national research. In: Becker, I., Ott, N., und Rolf, G. (Hrsg.): *Soziale Sicherung in einer dynamischen Gesellschaft. Festschrift für Richard Hauser zum 65. Geburtstag*. Campus, Frankfurt/New York.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Diekmann, A. (2007). *Empirische Sozialforschung: Grundlagen, Methoden, Andwendungen*. Rowohlts Enzyklopädie. Rowohlt Taschenbuch Verlag, Reinbek, 18. Auflage.

Domingos, P. und Pazzani, M. J. (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. In: *International Conference on Machine Learning*, Seiten 105–112. Morgan Kaufmann.

Duda, R. O. und Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

Fayyad, U., Piatetsky-Shapiro, G., und Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17:37–54.

Fürnkranz, J. (2005). From Local to Global Patterns: Evaluation Issues in Rule Learning Algorithms. In: Morik et al. (2005), Seiten 20–38.

Galler, H. P. (1987). Zur Längsschnittgewichtung des Sozio-ökonomischen Panels. Nr. 2 in Sozio-ökonomische Daten und Analysen für die Bundesrepublik Deutschland, Seiten 295–317. Campus, Frankfurt.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall, 5. Auflage.

Grzymala-Busse, J. W. und Hu, M. (2000). A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In: *Rough Sets and Current Trends in Computing*, Seiten 378–385.

Haisken-DeNew, J. (2001). A Hitchhikers's Guide to the World's Household Panel Data Sets. *The Australian Economic Review*, 34(3):356–366.

Haisken-DeNew, J. P. und Frick, J. R. (2005). *Desktop Companion to the German Socio-Economic Panel (SOEP)*. DIW Berlin.

Haisken-DeNew, J. P. und Hahn, M. (2006). *PanelWhiz: A Flexible Modularized Stata Interface for Accessing Large Scale Panel Data Sets*.

Hand, D. J. (2002). Pattern Detection and Discovery. In: Hand et al. (2002), Seiten 1–12.

Hand, D. J., Adams, N. M., und Bolton, R. J. (Hrsg.): (2002). *Pattern Detection and Discovery*, Band 2447 der Reihe *Lecture Notes in Computer Science*. Springer, Berlin/Heidelberg.

Hanefeld, U. (1987). *Das Sozio-ökonomische Panel: Grundlagen und Konzeption*. Nr. 1 in Sozio-ökonomische Daten und Analysen für die Bundesrepublik Deutschland. Campus, Frankfurt.

Hansen, J. (1982). *Das Panel: zur Analyse von Verhaltens- und Einstellungswandel*, Band 39 der Reihe *Beiträge zur sozialwissenschaftlichen Forschung*. Westdeutscher Verlag, Opladen.

Hartung, J., Elpelt, B., und Klösener, K.-H. (2005). *Statistik: Lehr- und Handbuch der angewandten Statistik*. Oldenbourg, München/Wien, 14., unwesentlich veränderte Auflage.

Heilemann, U. und Münch, H. J. (2007). Cyclical Accuracy of Macroeconomic Forecasts - A Multivariate Perspective. In: *27th International Symposium on Forecasting (ISF), June 24 to 28, 2007*, New York.

Hsiao, C. (2003). *Analysis of Panel Data*. Nr. 34 in Econometric Society Monographs. Cambridge University Press, 2. Auflage.

Kalton, G. (1983). *Introduction to Survey Sampling*. Nr. 07-035 in Sage University series on Quantitative Application in the Social Sciences. Sage Publications, Beverly Hills/London.

Kira, K. und Rendell, L. A. (1992). A Practical Approach to Feature Selection. In: Sleeman, D. und Edwards, P. (Hrsg.): *Proceedings of the 9th International Conference on Machine Learning*, Seiten 249–256, San Mateo. Morgan Kaufmann.

Kleinbaum, D. G. und Klein, M. (2002). *Logistic Regression: A Self-Learning Text*. Statistics for Biology and Health. Springer, New York, 2. Auflage.

Kluve, J., Schaffner, S., und Schmidt, C. M. (2005). Labor Force Status Dynamics in the German Labor Market: Individual Heterogeneity and Cyclical Sensivity. Technischer Bericht, RWI Essen.

Kononenko, I. (1994). Estimation attributes: Analysis and extensions of RELIEF. In: *European Conference on Machine Learning*, Band 784 der Reihe *Lecture Notes on Computer Science*, Seiten 171–182, Berlin/Heidelberg. Springer.

Kreienbrock, L. (1993). *Einführung in die Stichprobenverfahren*. Oldenbourg, München, 2. Auflage.

Kroh, M. und Spieß, M. (2006). Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) (1984 until 2005). Data Documentation 15, DIW Berlin, Berlin.

Lobo, O. O. und Numao, M. (1999). Ordered Estimation of Missing Values. In: *PAKDD '99: Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*, Seiten 499–503, London. Springer.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., und Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seiten 935–940, New York. ACM.

Mitchell, T. M. (1999). Machine Learning and Data Mining. *Communications of the ACM*, 42(11):30–36.

Morik, K., Boulicaut, J.-F., und Siebes, A. (Hrsg.): (2005). *Local Pattern Detection*, Band 3539 der Reihe *Lecture Notes in Artificial Intelligence*. Springer, Berlin/Heidelberg.

Murthy, S. K. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2(4):345–389.

Mátyás, L. und Sevestre, P. (Hrsg.): (1996). *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, Band 33 der Reihe *Advanced Studies in Theoretical and Applied Econometrics*. Kluver Academic Publishers, Dordrecht, 2., revidierte Auflage.

Neubäumer, R. (1982). *Die Eigenschaften verschiedener Stichprobenverfahren bei wirtschafts- und sozialwissenschaftlichen Untersuchungen*. Nr. 365 in Europäische Hochschulschriften: Reihe 5, Volks- und Betriebswirtschaft. Peter Lang, Frankfurt.

Pfanzagl, J. (1971). *Theory of Measurement*. Physica-Verlag, Würzburg, 2., revidierte Auflage.

Pyle, D. (1999). *Data preparation for data mining.* Morgan Kaufmann Publishers, San Francisco.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Quinlan, J. R. (1989). Unknown attribute values in induction. In: *Proceedings of the sixth international workshop on Machine learning*, Seiten 164–168, San Francisco, CA, USA. Morgan Kaufmann.

Quinlan, J. R. (1993). *C4.5: programs for machine learning.* Morgan Kaufmann, San Francisco.

Roberts, F. S. (1979). *Measurement Theory - with Applications to Decisionmaking, Utility and the Social Sciences*, Band 7 der Reihe *Encyclopedia of Mathematics and its Applications.* Addison-Wesley, Reading.

Rokach, L. und Maimon, O. (2005). Top-Down Induction of Decision Trees Classifiers - A Survey. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 35(4):476–487.

Rubin, D. R. (1976). Inference in missing data. *Biometrika*, 63(3):581–592.

Sauermann, J. (2005). Registrierte Arbeitslosigkeit oder Erwerbslosigkeit: Gibt es das bessere Konzept? *Wirtschaft im Wandel*, Seiten 104–108.

Schafer, J. L. und Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2):147–177.

Scheffer, T. und Wrobel, S. (2000). A Sequential Sampling Algorithm for a General Class of Utility Criteria. In: *KDD '00: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seiten 330–334, New York. ACM.

Scheffer, T. und Wrobel, S. (2002). Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling. *Journal of Machine Learning Research*, 3:833–862.

Scherf, M. und Brauer, W. (1997). Feature Selection by Means of a Feature Weighting Approach. Forschungsberichte Künstliche Intelligenz FKI-221-97, Institut für Informatik, Technische Universität München.

Schnell, R. (1986). *Missing-Data Probleme in der empirischen Sozialforschung.* Dissertation, Ruhr-Universität Bochum.

Scholz, M. (2005a). Knowledge-Based Sampling for Subgroup Discovery. In: Morik et al. (2005), Seiten 171–189.

Scholz, M. (2005b). Sampling-Based Sequential Subgroup Mining. In: Grossman, R. L., Bayardo, R., Bennett, K., und Vaidya, J. (Hrsg.): *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '05)*, Seiten 265–274, New York. ACM.

SOEP-Gruppe (2004). *Dokumentation zum SOEP-Datensatz: Datenlieferung zur Welle U.*

Spiess, M. (2004). *Analyse von Längsschnittdaten mit fehlenden Werten: Grundlagen, Verfahren und Anwendungen*. Universität Bremen.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684):677–680.

Sun, Y. und Li, J. (2006). Iterative RELIEF for Feature Weighting. In: *Proceedings of the 23rd International Conference on Machine Learning*.

Velleman, P. F. und Wilkinson, L. (1994). Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. In: Borg, I. und Mohler, P. (Hrsg.): *Trends and Perspectives in Empirical Social Research*, Seiten 161–177. De Gruyter, Berlin.

von Rosenbladt, B. (2004). SOEP 2004. Methodenbericht zum Befragungsjahr 2004 (Welle 21) des Sozio-oekonomischen Panels, SOEP-Gruppe, TNS Infratest Sozialforschung, München.

Winkelmann, L. und Winkelmann, R. (1998). Why Are the Unemployed So Unhappy? Evidence from Panel Data. *Economica*, 65:1–15.

Wrobel, S. (1997). An Algorithm for Multi-Relational Discovery of Subgroups. In: Komorowski, J. und Zytkow, J. M. (Hrsg.): *PKDD '97: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, Band 1263 der Reihe *Lecture Notes on Computer Science*, Seiten 78–87, Berlin/Heidelberg. Springer.

Index

Danksagung

Zu Dank verpflichtet bin ich zuallererst Prof. Dr. Katharina Morik und Ingo Mierswa für die stets freundliche und kompetente Betreuung sowie für Alles, was ich von ihnen über Data Mining und maschinelles Lernen lernen durfte. Ebenfalls bedanken möchte ich mich bei Prof. Dr. Gert G. Wagner für die Möglichkeit, die SOEP-Gruppe am DIW Berlin zu besuchen. Des Weiteren bedanke ich mich bei Christian Schmitt für eine erste, umfassende Einführung in das SOEP und bei Christin Schäfer, die mir von ihren Erfahrungen der Analyse sozioökonomischer Daten berichtete. Danken möchte ich außerdem Dr. Ronald Bachmann und Sandra Schaffner für hilfreiche Kritik und Isabell Hoffmann für das Korrektur Lesen einiger Kapitel. Ein ganz besonderer Dank gebührt meinen Eltern Detlef und Renate, die mich stets in meiner Arbeit unterstützt und in meinen Plänen bestärkt haben. Zum Schluß möchte ich Renate dafür danken, dass sie für mich da ist und überdies diese Arbeit Korrektur las.

Dortmund, den 11. März 2008

Tobias Malbrecht