Mundgesundheitsprognosen mit Bayesschen Netzwerken

D. Eherler, A. Borutta, T. Fischer

Lehrstuhl für Wirtschafts- und Sozialstatistik, Zentrum für Zahn-, Mund- und Kieferheilkunde
Wirtschaftswissenschaftliche Fakultät, Poliklinik für präventive Zahnheilkunde
Friedrich-Schiller-Universität Jena
Carl-Zeiss-Str. 3, Nordhäuser Str. 78
07743 Jena, 99089 Erfurt
Telefon: 03641 / 943304 bzw. 0361 / 7411301
eMail: d.eherler@wiwi.uni-jena.de, borutta@zmkh.ef.uni-jena.de,
fischer1@zmkh.ef.uni-jena.de

Abstract

Die Topologie Bayesscher Netzwerke beinhaltet auf intuitive Weise kausale Interpretationsmöglichkeiten. Arbeiten von Spirtes et al. (1993) und Pearl (1995) stellten den Zusammenhang einer kausalen Interpretation von solchen Netzwerken mit früheren kontrafaktischen Ansätzen von Rubin (1974) und Rosenbaum/Rubin (1983) für Kausalitätsuntersuchungen her. Der graphische Ansatz hat aber den Vorteil, wesentlich anwendungsorientierter zu sein. Mit Daten einer für das Bundesland Thüringen repräsentativen Mundgesundheitsstudie aus dem Jahre 1995 soll ein Beitrag geleistet werden, relevante Einflußgrößen auf die individuell wahrgenommene und objektive Mundgesundheit unter besonderer Berücksichtigung des individuellen Gesundheitsverhaltens mit Methoden der graphischen Modellierung zu ermitteln und zu analysieren, welche Auswirkungen auf die Mundgesundheit durch Verhaltensänderungen ausgelöst werden. Außerdem wird mit den von Kischka/Eherler (1999) angegeben Kriterien überprüft, ob die im Beispiel ermittelten Prognosen (kausale Effekte) bestimmter Verhaltensweisen auf Zielgrößen der Mundgesundheit unverzerrt (unkonfundiert) sind. Zur Lösung dieser Aufgaben ist die Verwendung elementarer SAS Prozeduren zur Verarbeitung kategorieller Daten hinreichend. Die bequeme Möglichkeit des Datenmanagements mit SAS liefert dabei große Hilfestellungen.

1 Einleitung

Das fundamentale Ziel vieler empirischer Untersuchungen ist es, die Auswirkungen von Verhaltensänderungen vorherzusagen. Dazu ist es erforderlich, die Verteilung der betrachteten Variablen auch unter der jeweiligen angewandten Verhaltensweise zu kennen (Spirtes et al. 1993, S. 201). Mit neueren Ansätzen der graphischen Modellierung kann in vielen Fällen die Bestimmung von Verteilungen unter definierten Verhaltensweisen aus Beobachtungsdaten erfolgen (Pearl 1996, S. 23). Dazu trägt die in gerichteten azyklischen Graphen porträtierte Information über (i) die qualitativen direkten und indirekten Beziehungen der betrachteten Variablen mittels eines Separationsbegriffes und (ii) die Richtung der Beeinflussung bei. Mit Daten einer Mundgesundheitsstudie aus dem Jahre 1995 für die Altersgruppe 35 bis 44 Jahre sollen die Auswirkungen von Verhaltensänderungen einerseits auf die subjektiv wahrgenommene Mundgesundheit und andererseits auf die objektiv diagnostizierte Mundgesundheit prognostiziert werden. Ein mit dem Softwarepaket TETRAD III (Scheines et al. 1996) generiertes graphisches Modell unter Berücksichtigung des konzeptionellen Design der Studie (Chen et al. 1986) dient dazu als Grundlage. Das innovative Element dieser Untersuchung liegt zum einen in der visuellen Darstellung der direkten Einflußgrößen auf interessierende Variablen, die auf dem generellen multivariaten Ansatz der Abhängigkeitsanalyse beruht. Zum anderen besteht die Möglichkeit, kausale Effekte von individuellen Verhaltensweisen auf Zielgrößen der Mundgesundheit zu bestimmen, ohne diese Verhaltensweisen bei allen Individuen beobachtet zu haben.

2 Graphische Modellierung

Die graphische Modellierung von probabilistischen Sachverhalten findet seit Mitte der 80er Jahre verstärkte Beachtung in der Statistik (z.B. Wermuth/Lauritzen 1983, Pearl 1988). Ein kurzer Abriß der theoretischen Grundlagen und die Bedeutung von Graphen für kausale Schlüsse sollen in diesem Abschnitt erfolgen.

2.1 Strukturgleichungsmodelle

Ein intuitiver Zugang zur kausalen Interpretation von Bayesschen Netzwerken geht auf Arbeiten von Pearl zurück. Ausgangspunkt ist ein Strukturgleichungsmodell, das wie folgt definiert ist.

Definition 1 (vgl. Galles/Pearl 1998)

Sei $(\Omega,\Sigma,\mathbf{P})$ ein Wahrscheinlichkeitsraum mit Zufallsvariablen $\boldsymbol{U}=\{U_1,...,U_n\}$ und $\boldsymbol{F}=\{f_1,...,f_n\}$ eine Menge meßbarer Funktionen. Ferner sei $\boldsymbol{V}=\{X_1,...,X_n\}$ eine Menge von Zufallsvariablen, die über das Gleichungssystem

$$X_i = f_i(\mathbf{PA}_i,U_i) \qquad i=1,...,n \qquad [1]$$

wohl definiert sind, wobei $\boldsymbol{PA}_i \subseteq \boldsymbol{V}/\{X_i\}$ die Menge der Variablen ist, von denen X_i funktional abhängt. Das Trippel $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$ erhält die Bezeichnung Strukturgleichungsmodell.

In einem Strukturgleichungsmodell werden die interessierenden Größen X_i in Abhängigkeit von ihren direkten Einflußgrößen $\boldsymbol{PA}_i$ über nicht näher spezifizierte Funktionen f_i dargestellt. Die Variablen U_i können als Störvariablen wie beispielsweise in einem gewöhnlichen Regressionsmodell aufgefaßt werden. Die über das Gleichungssystem [1] definierten Zufallsvariablen $\boldsymbol{V}=\{X_1,...,X_n\}$ besitzen eine gemeinsame Verteilung P auf dem Wahrscheinlichkeitsraum $(\Omega,\Sigma,\mathbf{P})$. Deren Ermittlung und deren Eigenschaften sind von zentralem Interesse. Weitere Hilfsmittel sollen dafür bereitgestellt werden. Dazu gehört die Darstellung eines Strukturgleichungsmodells als gerichteter Graph.

Definition 2

Sei $M=(U,F,V)$ ein Strukturgleichungsmodell. Der Graph $G(V,E)$ mit der Knotenmenge $V=\{X_1,...,X_n\}$ und den gerichteten Kanten $E=\{X_j \rightarrow X_i \mid X_j \in PA_i\}$ heißt zu M zugehöriger Graph.

Ein zu M zugehöriger Graph ist ein gerichteter azyklischer Graph, wenn das Gleichungssystem [1] rekursiv ist. Die graphische Darstellung von M stellt ein weitreichendes Inferenzinstrument dar, falls eine weitere Bedingung erfüllt ist.

Definition 3 (Pearl 1996)

Ein Strukturgleichungsmodell $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$ genügt der Markovbedingung, falls die Variablen U_i ($1 \le i \le n$) unabhängig sind.

Erfüllt ein Strukturgleichungsmodell diese Bedingung, so gilt der folgende Satz.

Satz 1 (Pearl 1988, Spirtes et al. 1993)

Sei $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$ ein Strukturgleichungsmodell, das die Markovbedingung erfüllt und sei der zugehörige gerichtete Graph $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ azyklisch. Dann sind äquivalent:

1. $P(X_1=x_1,...,X_n=x_n) = \prod_{i=1}^{n} P(X_i=x_i|\boldsymbol{PA}_i=\boldsymbol{pa}_i)$ [2]
2. $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ ist I-Map zur gemeinsamen Verteilung.
3. X_i ($1\leq i\leq n$) ist bedingt unabhängig von seinen Nichtnachfolgern gegeben seinen Eltern $\boldsymbol{PA}_i$ ($1\leq i\leq n$).

Die Aussagen des Satzes haben weitreichende Konsequenzen. Zunächst stellt der Graph und die gemeinsame Verteilung ein Bayessches Netzwerk dar, d.h. die gemeinsame Verteilung P kann bezüglich der Elternvariablen im Graph faktorisiert werden. Man beachte, daß durch die geforderte Rekursivität des Gleichungssystems [1] eine Ordnung der Variablen gegeben ist. Eine weitere Konsequenz ist, daß die Markovbedingung für Graphen (3) (Spirtes et al. 1993, S. 33) gilt. Diese hat aus kausalanalytischer Sicht eine sehr anschauliche Bedeutung. Sind die direkten Ursachen (hier die Elternknoten $\boldsymbol{PA}_i$) einer Variablen X_i unter Kontrolle, so gibt es keine weiteren systematischen Einflüsse auf eine Variable X_i und X_i ist somit isoliert bzw. bedingt unabhängig von allen anderen Variablen, die keine Nachfolger von X_i sind gegeben den Elternvariablen $\boldsymbol{PA}_i$.

Zur Interpretation der Aussage 2 des Satzes ist die Einführung eines Separationskonzeptes für gerichtete azyklische Graphen erforderlich, das auf Pearl (1988) zurückgeht.

Definition 4 (Pearl, 1988, S. 117)
Gegeben sei ein gerichteter azyklischer Graph $\mathbf{G}(\boldsymbol{V}, \boldsymbol{E})$. Seien X und Y zwei Knoten von $\mathbf{G}$ und sei $\boldsymbol{Z}$ eine Teilmenge der Knoten mit $X, Y \notin \boldsymbol{Z}$. Dann wird X durch $\boldsymbol{Z}$ von Y d-separiert (Symbol**:**$<X|\boldsymbol{Z}|Y>$), wenn gilt:
Für jeden ungerichteten Pfad zwischen X und Y existiert ein Knoten W auf dem Pfad mit

1. W besitzt aufeinander zulaufende Kanten $\rightarrow W\leftarrow$ und W sowie seine Nachfolger gehören nicht zu $\boldsymbol{Z}$,

   oder

2. W besitzt nicht aufeinanderzulaufende Kanten $\rightarrow W\rightarrow$**,** $\leftarrow W\rightarrow$ oder $\leftarrow W\leftarrow$ und W gehört zu $\boldsymbol{Z}$.

Zwei Mengen von Knoten $\boldsymbol{X}$ und $\boldsymbol{Y}$ werden durch eine Menge $\boldsymbol{Z}$ d-separiert, falls $\boldsymbol{Z}$ jeden Knoten $X \in \boldsymbol{X}$ von jedem Knoten $Y \in \boldsymbol{Y}$ d-separiert.

Die Eigenschaft eines Graphen, I-Map für eine Wahrscheinlichkeitsverteilung zu sein, besagt, daß aus jeder gültigen d-Separationseigenschaft $<X|\boldsymbol{Z}|Y>$ die entsprechende Unabhängigkeitsbeziehung (Symbol $X\perp Y|\boldsymbol{Z}$) folgt.

Neben dem hier dargestellten Zugang der Konstruktion von Bayesschen Netzwerken gibt es noch andere. Beispielsweise kann, bei gegebener gemeinsamer Wahrscheinlichkeitsverteilung P einer Menge von Variablen $\boldsymbol{V}=\{X_1,...,X_n\}$ in einer bestimmten z.B. durch die Indizierung gegebenen Reihenfolge, zu jeder Variablen X_i ($1\leq i\leq n$) eine minimale Menge $\boldsymbol{PA}_i$ bestimmt werden, so daß

$$P(X_i=x_i|\, X_1=x_1,...,\, X_{i-1}=x_{i-1}) = P(X_i=x_i|\boldsymbol{PA}_i=\boldsymbol{pa}_i)$$

gilt (vgl. Castillo et al. 1997, S. 240-1). Einen weiteren Zugang bieten sogenannte Abhängigkeitsmodelle bzw. Inputlisten (vgl. Pearl 1988, S. 119, Castillo et al. 1997, S. 185).

2.2 Kausale Effekte aus Strukturgleichungsmodellen

Zur Heranführung an Kausalitätsuntersuchungen mit Bayesschen Netzwerken soll in einem Zwischenschritt ein weiterer Verteilungsbegriff eingeführt werden.

Definition 5

Sei $\boldsymbol{V}=\{X_1,...,X_n\}$ eine Menge von Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega,\Sigma,\mathbf{P})$ mit gemeinsamer Verteilung und der durch die Indizierung bestimmten Reihenfolge. Ferner sei k ($1\leq k\leq n$) fest gewählt und $\bar{x} \in$ IR eine Realisation von X_k. Dann ist durch

$$\tilde{P}_{\bar{x}}\ (X_1=x_1,...,X_{k-1}=x_{k-1},X_{k+1}=x_{k+1},...,X_n=x_n) = \frac{P(X_1 = x_1, \ldots, X_k = \bar{x}, \ldots, X_n = x_n)}{P(X_k = \bar{x} \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1})} \qquad [3]$$

die gemeinsame Verteilung von $\{X_1,...,X_{k-1},X_{k+1},..,X_n\}$ unter der Festsetzung von X_k auf $\bar{x}$ definiert.

Lemma 1

Gegeben sei die Verteilung $\tilde{P}_{\bar{x}}$ von $\{X_1,...,X_{k-1},X_{k+1},..,X_n\}$ unter der Festsetzung von X_k auf $\bar{x}$. Dann gilt (Alle bedingten Wahrscheinlichkeiten seien wohldefiniert.):

1. $\tilde{P}_{\bar{x}}\ (X_1=x_1,...,X_{k-1}=x_{k-1},X_{k+1}=x_{k+1},...,X_n=x_n) = \prod_{i\neq k} P(X_i=x_i \mid X_1=x_1,...,X_{i-1}=x_{i-1})$
2. $\tilde{P}_{\bar{x}}\ (X_1=x_1,...,X_{k-1}=x_{k-1},X_{k+1}=x_{k+1},...,X_n=x_n)$
   $= P(x_1,..., x_{k-1},x_{k+1},..., x_n \mid X_k=\bar{x})P(x_1,...,x_{k-1})$
3. Ist X_k unabhängig von $\{X_1,...,X_{k-1}\}$, so entspricht die Verteilung $\tilde{P}_{\bar{x}}$ der bedingten Verteilung gegeben $X_k=\bar{x}$, d.h.

$$\tilde{P}_{\bar{x}}\ (x_1,..., x_{k-1},x_{k+1},..., x_n) = P(x_1,..., x_{k-1},x_{k+1},..., x_n \mid X_k=\bar{x})$$

Beweis:
Die Aussagen folgen aus dem Faktorisierungssatz und elementaren Umformungen für bedingte Wahrscheinlichkeiten.

Dieser Verteilungstyp, bei dem eine Variable auf einen bestimmten Wert festgesetzt wird, ergibt sich aus Strukturgleichungsmodellen oder Bayesschen Netzwerken, wenn eine Manipulation in einem noch zu spezifizierenden Sinn durchgeführt wird. Daher hat sich der Begriff manipulierte Verteilung für den Ausdruck $\tilde{P}_{\bar{x}}$ eingebürgert (Spirtes et al. 1993, Pearl 1995).

Definition 6 (Galles/Pearl 1998)
Sei $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$ ein Strukturgleichungsmodell. Eine Manipulation von X_k auf den festen Wert $\bar{x} \in$ IR besteht aus der Ersetzung der k-ten Gleichung durch die Identität $X_k \equiv \bar{x}$ und der Einsetzung von $\bar{x}$ in $\boldsymbol{pa}_j$, falls $X_k \in \boldsymbol{PA}_j$.

Graphisch läßt sich eine Manipulation dadurch darstellen, daß beim Knoten X_k alle Kanten entfernt werden, die auf X_k zulaufen. Sonst werden keine anderen Änderungen vorgenommen (Spirtes et al. 1993, S. 209).

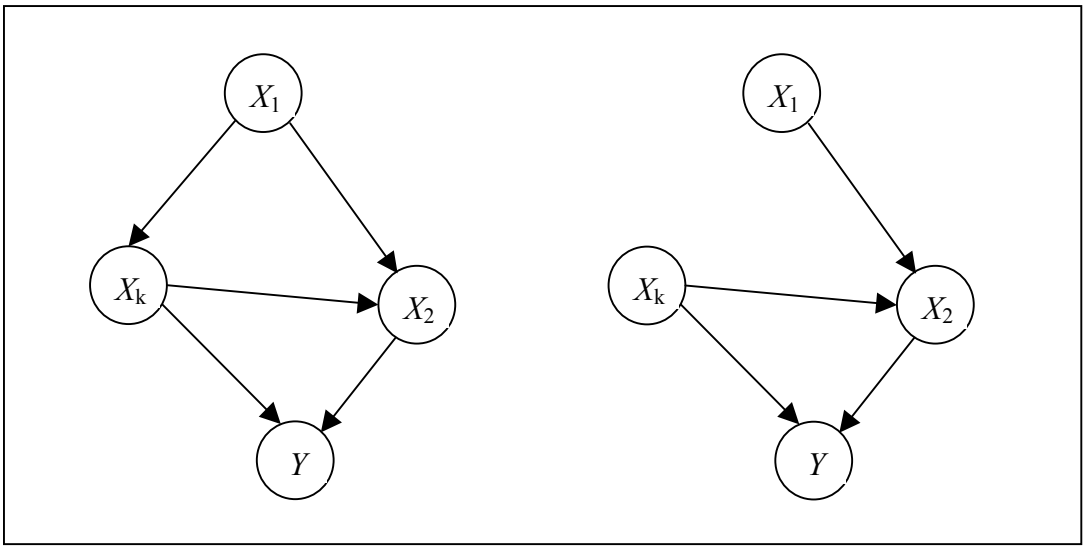

Abbildung 1: Ausgangsgraph und manipulierter Graph

Die Auswirkung einer solchen Manipulation auf die Verteilung der verbleibenden Variablen ist Inhalt des nächsten Satzes.

Satz 2 (Pearl 1995)
Sei $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$ ein Strukturgleichungsmodell, das die Markovbedingung erfüllt und sei der zugehörige gerichtete Graph $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ azyklisch. Es werde eine Manipulation von X_k auf den festen Wert $\bar{x}$ vorgenommen. Dann ist die Verteilung der verbleibenden Variablen gegeben durch die Verteilung unter der Festsetzung X_k auf den Wert $\bar{x}$ und läßt sich wie folgt faktorisieren

$$\tilde{\mathrm{P}}_{\bar{x}}(X_1=\mathrm{x}_1,\ldots,X_{k-1}=\mathrm{x}_{k-1},X_{k+1}=\mathrm{x}_{k+1},\ldots,X_n=\mathrm{x}_n) = \prod_{i\neq k}\mathbf{P}(X_i=x_i|\boldsymbol{PA}_i=\boldsymbol{pa}_i). \qquad [4]$$

Mit der Aussage des Satzes 2 liegt eine inhaltliche Bedeutung für die Verteilung unter einer Festsetzung nahe: Diese Verteilung beschreibt, wie die restlichen Variablen in einem System variieren, wenn für die Variable X_k nur noch der feste Wert $\bar{x}$ auftaucht. In der Realität wird diese Art von Verteilung im allgemeinen nicht beobachtbar sein. Jedoch innerhalb des durch die Manipulation erzeugten Systems ist die Verteilung $\tilde{\mathrm{P}}_{\bar{x}}$ trotz ihrer kontrafaktischen Natur bestimmbar[1]. Ist die Variable X_k auf verschiedene Werte „einstellbar“ im Sinne einer Behandlungsvariablen, so kann ein Effekt von X_k auf eine andere Variable im System wie folgt definiert werden.

Definition 7
Ein Effekt einer Variable X_k auf eine weitere Variable $Y\in V/\{X_k\}$ ist durch die Randverteilung $\tilde{\mathrm{P}}_{\bar{x}}$ bzgl. Y gegeben. Ein kausaler Effekt von X_k auf Y liegt vor, wenn sich die Randverteilungen bzgl. Y von $\tilde{\mathrm{P}}_{\bar{x}}$ und $\tilde{\mathrm{P}}_{\bar{x}'}$ für $\bar{x}\neq\bar{x}'$ unter den Manipulationen $X_k\equiv\bar{x}$ und $X_k\equiv\bar{x}'$ unterscheiden.

Ein Effekt beschreibt die Auswirkung einer durch $\bar{x}$ bestimmten Maßnahme auf eine Zielvariable Y. Die Verteilung von Y wird untersucht, wenn für die gesamte Population die Variable X_k lediglich den Wert $\bar{x}$ besitzt. Ein solcher Zustand könnte beispielsweise durch

[1] Weitere Ausführungen, insbesondere auch zum Zusammenhang dieses Ansatzes mit dem Rubinschen Ansatz befinden sich in Sprites et al. (1993) und Galles/Pearl (1998).

ein experimentelles Design erzeugt werden. Beschreibt jedoch ein Strukturgleichungsmodell einen Sachverhalt auf adäquate Weise, so kann auch aus einer Beobachtungsstudie ermittelt werden, wie eine Variable X_k eine Variable Y im oben definierten Sinne beeinflußt.

Zur Bestimmung der Effekte braucht nicht auf die gesamte gemeinsame Verteilung der verbleibenden Variablen zurückgegriffen werden. Ein graphisches Kriterium erlaubt häufig eine einfachere und schnellere Berechnung.

Satz 3 (Pearl 1995) Backdoor-Kriterium
Sei $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$ ein Strukturgleichungsmodell, das die Markovbedingung erfüllt und sei der zugehörige gerichtete Graph $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ azyklisch. Seien X_k und Y gegeben, $\bar{x}$ ein fester Wert von X_k und $\boldsymbol{Z}\subset\boldsymbol{V}/\{X_k,Y\}$ mit

(i) kein Element von $\boldsymbol{Z}$ ist Nachfolger von X_k

(ii) für jeden Pfad X_k zwischen und Y mit auf X_k zulaufende Kanten gilt: $\boldsymbol{Z}$ d-separiert X_k und Y.

Dann gilt:

$$\tilde{P}_{\bar{x}}(Y=y) = \sum_{z} P(Y=y|X_k=\bar{x},\boldsymbol{Z}=z)P(\boldsymbol{Z}=z). \qquad [5]$$

Die Formel des Backdoor-Kriteriums bewerkstelligt eine Adjustierung gemeinsamer Einflußfaktoren. Grundlage zur Anwendung des hier beschriebenen Ansatzes ist, daß ein Strukturgleichungsmodell **M** bzw. der zugehörige Graph bekannt ist, das bzw. der den betrachteten Sachverhalt adäquat abbildet und die zur Bestimmung von Effekten relevanten Variablen beobachtet sind.

2.3 Graphen aus Daten

Neben der Konstruktion von Graphen bzw. Strukturgleichungsmodellen aus substantiellen Überlegungen ist es auch möglich unter der Voraussetzung weiterer Annahmen, solche Modelle aus Daten zu generieren.

Definition 8 (Spirtes et al. 1993, S. 35)
Sei $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$ ein Strukturgleichungsmodell, das die Markovbedingung erfüllt und sei der zugehörige gerichtete Graph $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ azyklisch. Dann genügen $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ und die vom Modell **M** induzierte Verteilung P der Variablen $\boldsymbol{V}$ der Faithfulnessbedingung, falls d-Separationseigenschaften und Unabhängigkeitsbeziehungen übereinstimmen.

Meek (1995) hat unter schwachen Voraussetzungen gezeigt, daß diese Bedingung fast immer erfüllt ist. Das Konstruktionsverfahren Tetrad III (Scheines et al. 1996), das hier zur Anwendung kommt, geht so vor, daß Unabhängigkeitsbeziehungen in d-Separationseigenschaften übertragen werden und daraus ein partiell orientierter Graph konstruiert werden kann.

3 Datenbasis

Das in diesem Beitrag untersuchte Datenmaterial stammt aus einer repräsentativen Mundgesundheitsstudie für das Bundesland Thüringen aus dem Jahre 1995 (Borutta/Brocker, 1998). Konkret umfaßt der Datensatz 614 Beobachtungen von erwachsenen Probanden. Diese Studie gliederte sich in einen soziologischen und klinisch epidemiologischen Teil und folgte in der Durchführung dem in der Einleitung erwähnten Erkärungsmodell (Chen et al., 1986). Insbesondere werden soziodemographische Variablen sowie Einstellungen zur Mundgesundheit (Strukturblock), das Mundgesundheitsverhalten (Prozeßblock) neben dem subjektiv wahrgenommenen und objektiv bestimmten oralen Zustand (Ergebnisblock) berücksichtigt. Die untersuchten Variablen sind:

Strukturblock	
GESCHLECHT	Ausprägung: 0=weiblich, 1=männlich
SCHULBILDUNG	Ausprägung: 0=8 Jahre,1=mehr als 8 Jahre
EINSTELLUNG ZUR ZAHNÄRZTL. VERSORGUNG	Ausprägung: 0=positiv,1=eher negativ
Prozeßblock	
HÄUFIGKEIT DER ZWISCHENMAHLZEITEN	Ausprägung: 0=2 oder 3 mal, 1=1 mal ,2=selten oder nie
ZUCKERKONSUM	Ausprägung: 0=nein,1=ja
MUNDHYGIENE	Ausprägung: 0=mind. 2 mal, 1=mind 1 mal, 2=seltener
FLUORIDVERWENDUNG	Ausprägung: 0=ja,1=nein
ZAHNSEIDEBENUTZUNG	Ausprägung: 0=nein,1=ja
Ergebnisblock (objektiv (o) / subjektiv (s))	
(o) PARODONTALSTATUS (CPITN)	Ausprägung: 0,1,2,3,4
(o) KARIESVERBREITUNG (DMFT)	Ausprägung: 0=nicht größer 7, 1=[8,15], 2=[16-34]
(s) ZAHNFLEISCHBLUTEN	Ausprägung: 0=ja,1=nein
(s) ZAHNSCHMERZEN IM LETZTEN JAH R	Ausprägung: 0=ja,1=nein
(s) ENTZÜNDUNGEN IN DER MUNDHÖHLE	Ausprägung: 0=ja,1=nein

4 Untersuchung mit generierten Graphen

Der Entdeckungsalgorithmus von Tetrad III erlaubt die Berücksichtigung von Hintergrundwissen, welches hier in der Anordnung der Variablen nach ihrer Blockzugehörigkeit besteht. Zur Generierung des Graphen werden Unabhängigkeitstests durchgeführt. Das Signifikanzniveau wurde auf α=10% festgesetzt. In SAS implementierte Unabhängigkeitstests wurden in Eherler (1999) für andere Fragestellungen angewendet.

4.1 Struktur des generierten Graphen

Für den untersuchten Datensatz ergab sich der in Abbildung 2 dargestellte Graph. Direkt miteinander verbundene Variablen sind abhängig, und es besteht ein direkter Einfluß. Negative Einflüsse sind durch gestrichelte Kanten von positiven zu unterscheiden. Marginale und bedingte Unabhängigkeit lassen sich aus dem Graphen mit Hilfe des d-Separationskriteriums ablesen. Beispielsweise sind Geschlecht und Einstellung zur zahnärztlichen Versorgung marginal unabhängig oder das Empfinden von Zahnschmerzen bedingt unabhängig vom Geschlecht, falls die Mundhygiene und die Häufigkeit der Zwischenmahlzeiten bekannt sind. Bedingte Unabhängigkeit liegt auch zwischen der Kariesverbreitung und dem Zuckerkonsum vor, falls das Zahnputzverhalten und das Bildungsniveau als gegeben vorausgesetzt wird.

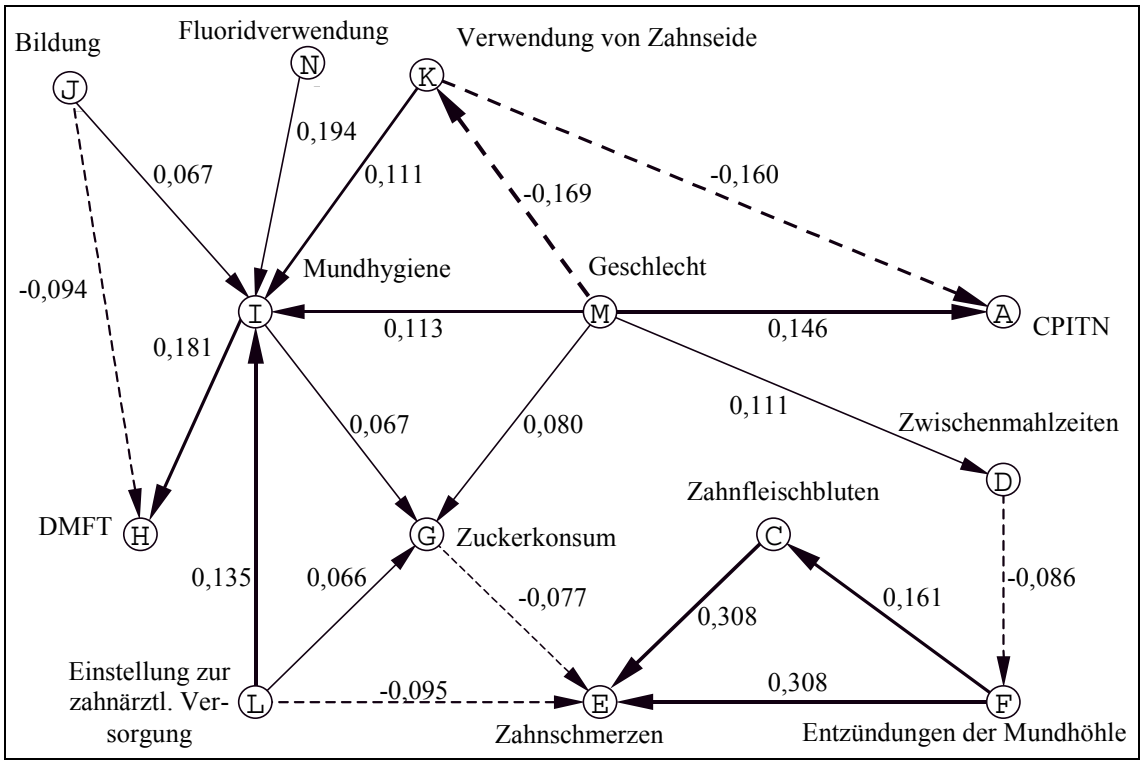

Abbildung 2: Generierter Graph

4.2 Quantifizierung der Struktur

Mit der in Gleichung [2] gegebenen Vorschrift kann die zugrundeliegende Verteilung der Variablen geschätzt werden. Das zugehörige SAS-Programm hat einen einfachen Aufbau:

```
PROC FREQ;
TABLE PA_i * X_i * DUMMY;   * Dummy-Variable mit einer Ausprägung;
```

Diese Struktur kann wiederum genutzt werden, um Fragestellungen wie in Gliederungspunkt 2.2 angerissen, zu bearbeiten. Ferner kann im Fall von binären Variablen ein der Korrelation ähnlicher Phi-Koeffizient (z.B. Hartung 1995, S. 446) bzw. bei den weiteren ordinalen Variablen der Spearmann'sche Rangkorrelationskoeffizient angegeben werden. Zu deren Ermittlung kommen die SAS-Prozduren (SAS 1990) FREQ und CORR zur Anwendung.

```
PROC FREQ;
TABLE X_j * X_i / CHISQ;    * Ermittlung des Phi-Koeffizienten;


PROC CORR SPEARMAN;
VAR X_k * X_l;  * Ermittlung des Rangkorrelationskoeffizienten;
```

Die Koeffizienten sind im Graphen eingetragen. Zwischen den Variablen Benutzung von Zahnseide und Parodontalstatus beträgt dieser -0,160. Bei der verwendeten Kodierung besagt dies, daß die Benutzung von Zahnseide den Parodontalstatus optimiert.

4.3 Wirksamkeit von Verhaltensänderungen

Zur Wirksamkeit von Verhaltensänderung können aus dem generierten Graphen (Abb. 2) beispielsweise folgende Fragestellungen bearbeitet werden:

- Welche Auswirkungen hat der Zuckerkonsum / die Mundhygiene / die Häufigkeit der Zwischenmahlzeiten auf die wahrgenommenen Zahnschmerzen?
- Welchen Effekt hat die Benutzung von Zahnseide auf den Parodontalstatus?

Eine erste Antwort auf diese Fragestellungen, welche allerdings nicht eine Adjustierung der gemeinsamen Ursachen berücksichtigt, ist die Betrachtung der geschätzten bedingten Wahrscheinlichkeiten. Für die Untersuchung der Auswirkung des Zuckerkonsums auf die empfundenen Zahnschmerzen ergäbe das einen Anteil der Population von 26,14% im Vergleich zu 19,1% derjenigen, die selten Zucker konsumieren.

Diese Anteile ändern sich jedoch, wird der konzeptionelle Ansatz der Manipulation verwendet. Bei dieser Betrachtungsweise entsprechen die resultierenden Anteile denen einer Population, bei der alle Individuen das durch die Manipulation vorgegebene Verhalten an den Tag legen würden. Mit dem von Pearl (1995) entwickelten Backdoor-Kriterium und denen im Graphen (Abb. 2) verwendeten Kürzel ergeben sich die folgenden Berechnungsvorschriften (alle Größen rechts sind beobachtet),

$$\tilde{P}_g(E=e) = \sum_{l,m} P(L=l,M=m)\, P(E=e|\, G=g,L=l,M=m) \quad [6]$$

$$\tilde{P}_i(E=e) = \sum_{l,m} P(L=l,M=m)\, P(E=e|\, I=i,L=l,M=m) \quad [7]$$

$$\tilde{P}_d(E=e) = \sum_{m} P(M=m)\, P(E=e|\, D=d,M=m) \quad [8]$$

$$\tilde{P}_k(A=a) = \sum_{m} P(M=m)\, P(A=a|\, K=k,M=m) \quad [9]$$

wobei die Variablen Geschlecht (M) und Einstellung zur zahnärztlichen Versorgung (L) eine Adjustierung der gemeinsamen Ursachen der manipulierten Variablen Zuckerkonsum (G), Mundhygiene (I) und Häufigkeit der Zwischenmahlzeiten (D) bzw. Benutzung von Zahnseide (K) und den Zielvariablen Zahnschmerzen (E) bzw. Parodontalstatus (A) bewerkstelligen.

G→E	0	1
0	0,1783	0,8217
1	0,2633	0,7367

I→E	0	1
0	0,2482	0,7518
1	0,3493	0,6507
2	0,3819	0,6181

D→E	0	1
0	0,2590	0,7410
1	0,2479	0,7522
2	0,2452	0,7552

Wäre der Zuckerkonsum (G) in der gesamten Population selten, ergäbe sich ein Populationsanteil mit Zahnschmerzen (E) von 17,83%. Dieser würde sich auf 26,33% erhöhen, falls die Population nur aus Individuen bestünde, welche häufig Zucker konsumierten. Die unter der Manipulation gefundenen Anteile unterscheiden sich nur geringfügig von den beobachteten in dieser Studie, wobei der positive Effekt einer Reduzierung des Zuckerkonsums auf das Zahnschmerzempfinden in der Beobachtungstudie geringer geschätzt [26,14%-19,1% = 7,04%] wird als beim manipulativen Ansatz [26,33%-17,83%=8,5%]. Analog sind die Effekte von Mundhygiene (I) und Häufigkeit der Zwischenmahlzeiten (D) auf die empfundenen Zahnschmerzen (E) zu interpretieren.

Würden alle Individuen ihr Zahnputzverhalten (I) optimieren, führte dies zu einer Verringerung des Anteils derjenigen, die Zahnschmerzen empfinden (E) von 38,19% auf 24,82% im Vergleich zu sehr nachlässiger Mundhygiene aller Individuen.

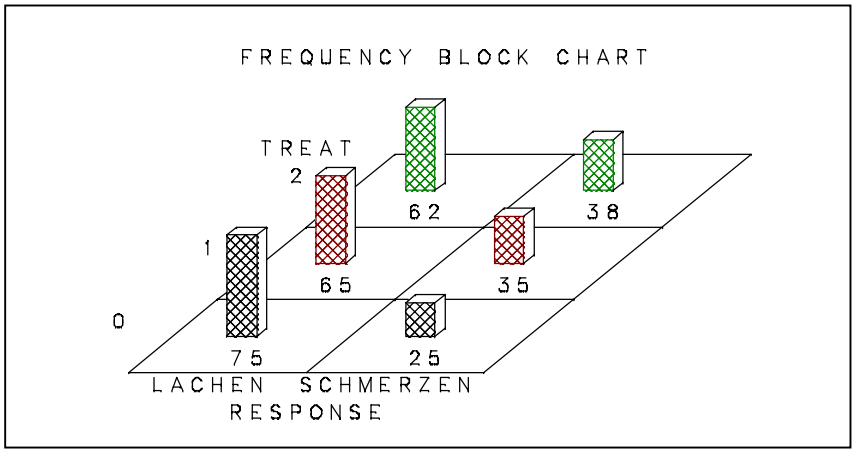

Abbildung 3: Treatment: Mundhygiene, Response: Zahnschmerzen (Anteile in %)

Nur geringen kausalen Einfluß auf die empfundenen Zahnschmerzen übt die Häufigkeit der Zwischenmahlzeiten (D) aus. Die Anteile unterscheiden sich nur gering.

Zur zweiten Fragestellung des Einflusses der Benutzung von Zahnseide auf den Parodontalstatus wird mit dem manipulativen Ansatz prognostiziert, daß eine Anwendung von Zahnseide der gesamten Population einen deutlich positiven Einfluß hinsichtlich dieses Kriteriums der Mundgesundheit hätte.

K→A	0	1	2
0	0,0206	0,0564	0,9231
1	0,0785	0,0677	0,8539

Dies würde sich in jeder betrachteten Kategorie bestätigen. Mit SAS können diese Ergebnisse wie folgt ermittelt werden:

```
PROC FREQ;
TABLE M*K*A*Dummy / OUT = SCHICHT_t;
WHERE K=x AND M = t;
RUN;
```

Für jede Treatmentausprägung x und jede Schicht bezüglich der Variable M wird die Randwahrscheinlichkeit $P(Y=y|K=x,M=t)$ ermittelt. Diese werden dann in einer Datei (`SCHI_all`) zusammengefaßt. In einer separaten Datei (`SCHI_gew`) werden die Schichtungsanteile ermittelt. Diese beiden Dateien enthalten die Information, um in einem DATA-Step die Effekte einer Manipulation zu bestimmen.

```
DATA EFFEKT;
MERGE SCHI_gew SCHI_all;
BY M;
SUMMAND = PERCENT * PERCENT1 / 10000; *Summation nach Formel [9];
P_TILDE_x + SUMMAND;
KEEP P_TILDE_x;
```

Die Variable `P_TILDE_x` enthält dann den Effekt der Manipulation $K \equiv x$.

4.4 Ausschluß der Existenz weiterer maßgeblicher Faktoren

Die hier durchgeführten Analysen besitzen zunächst Gültigkeit innerhalb des durch das Gleichungssystem spezifizierten Modells. Es können jedoch weitere Variablen, eventuell auch nicht beobachtete und nicht beobachtbare, existieren, die die ermittelten Effekte verzerren. Zum Beispiel können weitere konfundierende Variablen existieren, die in die Bestimmung eines Effektes mit einbezogen werden müssen. Beispielsweise müßte in Abb. 4 außer mit Variable *T* auch mit der Variablen *U* adjustiert werden.

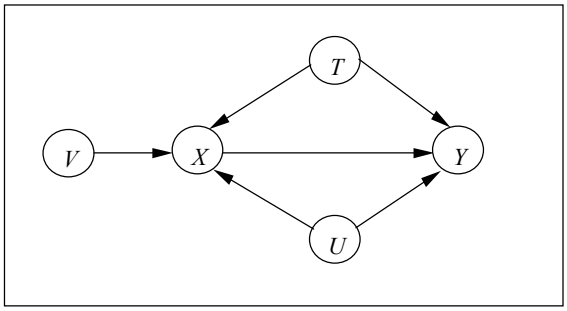

Abbildung 4: Beispielsgraph

In einem Diskussionspapier von Kischka/Eherler (1999) wurden Ansätze von Spirtes et al. (1993, S. 17-19) und Pearl (1998) verallgemeinert auf solche Situationen, in denen die Existenz von konfundierenden Variablen bekannt ist. Grundlage bildet der folgende Unkonfundiertheitsbegriff.

Definition 9
Gegeben sei ein Strukturgleichungsmodell $\mathbf{M}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$. Der Effekt eines geordneten Variablenpaares (X,Y) ist stabil unkonfundiert bzgl. einer Menge $\boldsymbol{T}$, falls

$$\tilde{P}_x(Y=y)=\sum_{\boldsymbol{t}} P(\boldsymbol{T}=\boldsymbol{t})P(Y=y|X=x,\boldsymbol{T}=\boldsymbol{t}) \qquad \text{für alle } x$$

für beliebige Störvariablen $\boldsymbol{U}$ und beliebige meßbare $\boldsymbol{F}$.

Der folgende Satz gibt ein hinreichendes Kriterium an, wann weitere konfundierende Variablen ausgeschlossen werden können.

Satz 4 (Kischka/Eherler 1999)
Gegeben sei ein nicht näher spezifiziertes Strukturgleichungsmodell $\mathbf{M'}=(\boldsymbol{U},\boldsymbol{F},\boldsymbol{V})$, der zugehörige Graph $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ sei azyklisch und $\mathbf{G}(\boldsymbol{V},\boldsymbol{E})$ und $P(\boldsymbol{V})$ genügen der Faithfulnessbedingung. Für ein Teilmodell $\mathbf{M}$ von $\mathbf{M'}$ über die Teilmenge $\boldsymbol{O}\subset\boldsymbol{V}$ seien (X,Y) stabil unkonfundiert bzgl. einer Menge $\boldsymbol{T}\subset \boldsymbol{O}\backslash\{X,Y\}$. Sei V eine weitere Variable von $\mathbf{M'}$ mit

1. Es existiert ein gerichteter Pfad von V nach X.
2. V ist nicht bedingt unabhängig von X gegeben $\boldsymbol{T}$
3. V ist bedingt unabhängig von Y gegeben $\boldsymbol{T}\cup\{X\}$

Dann gilt:
Es existieren keine weiteren konfundierenden Variablen in $\mathbf{M'}$ und der Effekt von X auf Y ist gegeben durch

$$\tilde{P}_x(Y=y)=\sum_{\boldsymbol{t}} P(\boldsymbol{T}=\boldsymbol{t})P(Y=y|X=x,\boldsymbol{T}=\boldsymbol{t}) \qquad \text{für alle } x.$$

Ist die Aufgabe gestellt, einen Effekt einer Treatmentvariablen X auf eine Responsevariable Y zu ermitteln, und sind bereits konfundierende Variablen $\boldsymbol{T}$ bekannt, kann nach einer weiteren Variablen V gesucht werden, die die Voraussetzungen (1)-(3) erfüllt. Die Existenz eines gerichteten Pfades stellt eine substantielle Annahme[2] dar, die weiteren Aussagen können statistisch überprüft werden. Gelingt dies, kann aus beobachteten Variablen ein kausal interpretierbarer Effekt bestimmt werden.

Exemplarisch soll Satz 4 auf zwei der hier betrachteten Problemstellungen angewendet werden. In obiger Studie wurde der Effekt für die Variablen Mundhygiene und empfundene Zahnschmerzen gemäß des Graphen (Abb. 2) ermittelt. Als konfundierende Variable wurden bereits die Einstellung zur zahnmedizinischen Versorgung und das Geschlecht berücksichtigt. Als weitere Variable V soll Fluoridverwendung betrachtet werden. Die erste Voraussetzung, daß Fluoridverwendung ein Vorgänger von der Mundhygiene ist, erscheint nicht unplausibel und mit der Abhängigkeit der beiden Variablen ist ein notwendiges Kriterium dafür erfüllt. Die beiden durchzuführenden Tests auf bedingte Unabhängigkeit ergeben mit SAS:

```
PROC FREQ;
TABLE L*M*N*I /CMH;          * Überprüfung der Bedingung 2;

PROC FREQ;
TABLE L*M*I*N*E /CMH;        * Überprüfung der Bedingung 3;
```

Eine Ausführung der Programmzeilen führt zu den Ausgaben

```
                SUMMARY STATISTICS
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic  Alternative Hypothesis    DF     Value     Prob
        3        General Association        2      44.368     0.001

                SUMMARY STATISTICS
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic  Alternative Hypothesis    DF     Value     Prob
        3        General Association        1       0.071     0.790
```

und somit zur Entscheidung, daß die entsprechende Nullhypothese auf bedingte Unabhängigkeit im ersten Fall abgelehnt werden kann und im zweiten Fall nicht. Es kann also, unter diesen Gegebenheiten festgehalten werden, daß der Effekt von Mundhygiene auf empfundene Zahnschmerzen gegeben ist durch den Ausdruck [7] bzw. wie in Abbildung 3 dargestellt.

Im Fall des durch Gleichung [9] bestimmten Effektes kann in der Studie keine derartige Variable V gefunden werden, die den Voraussetzungen in Satz 4 genügt. Die ermittelten Effekte sind in ihrer Aussagekraft dennoch in Bezug auf das betrachtete System gültig.

[2] Es gibt Möglichkeiten, die Existenz gerichteter Pfade auch aus Daten zu extrahieren (Eherler 2000)

5 Fazit

Mit der Weiterentwicklung der graphischen Modellierung wahrscheinlichkeitstheoretischer Sachverhalte ist ein neuer Ansatz der multivariaten Datenanalyse entstanden, welcher sich durch seine übersichtliche und klare Ergebnispräsentation auszeichnet (vgl. Helfenstein et al. 1999) Darüber hinaus stellen graphische Ansätze eine Methodik zur Verfügung, welche unter bestimmten Bedingungen die Berechnung kontrafaktischer Wahrscheinlichkeiten erlaubt. Diese bilden wiederum Grundlage für kausale Aussagen in der Statistik. Anhand der analysierten Mundgesundheitsparameter wurde mit dieser Methodik die Wirksamkeit individueller Verhaltensweisen auf die wahrgenommene Mundgesundheit bestätigt.

Literatur

Borutta, A.; Brocker, M. (1998): Oraler Gesundheitszustand in Beziehung zu relevanten Merkmalen der Persönlichkeit und ihres sozialen Umfeldes. Studiendesign, Stichprobenauswahl, Untersuchungsmethoden und biostatistische Auswertungsverfahren. In: Stoesser, L. (Hrsg.): Kariesdynamik und Kariesrisiko; S. 93-97; Quintessenz-Verl.; Berlin.

Castillo, E.; Gutiérrez, J.M.; Hadi, A.S. (1997): Expert Systems and Probabilistic Network Models, Springer; New York.

Chen, M. et al. (1986): Testing the Health Belief Model: LISREL analysis of Alternative Models of Causal Relationships Between Health Beliefs and Preventive Dental Behavior; Social Science Quarterly; 49; 45-60.

Eherler, D. (1999): Graphische Modellierung in der Kundenzufriedenheitsanalyse; in: C. Ortseifen (Ed.); Proceedings der 3. KSFE; S. 73-83; Ruprecht-Karls-Universität Heidelberg; Heidelberg.

Eherler, D. (2000): Causal Paths from Data; erscheint als Diskussionspapier.

Galles, D.; Pearl, J. (1998): An Axiomatic Characterization of Causal Counterfactuals; Foundations of Science; 3(1); 151-182.

Hartung, J. (1995): Statistik: Lehr- und Handbuch der angewandten Statistik; 10. Auflage; München; Oldenbourg.

Helfenstein, U.; Steiner, M. ;Menghini. G (1999): An Outline of Graphical Markov Models in Dentistry; in: Communitiy Dental Health; 16; 220-6.

Kischka, P.; Eherler, D. (1999): Causal Graphs and Unconfoundedness; Diskussionspapier; Wirtschaftswissenschaftliche Fakultät; Friedrich-Schiller-Universität Jena; Nr. 99/05; Jena.

Lauritzen, S.L. (1996): Graphical Models; Clarendon Press; Oxford.

Pearl, J. (1988): Probabilistic Reasoning in Intelligent Systems; Morgan Kaufman Publishers; San Mateo; CA.

Pearl, J. (1993): Aspects of Graphical Models connected to Causality; UCLA; Technical Report; TR-195-LL; Los Angeles.

Pearl, J. (1995): Causal Diagramms for Empirical Research. Biometrika; 82; December; 669-710.

Pearl, J. (1996): A Causal Calculus for Statistical Research; in: Fisher, D.; Lenz, H.J. (1996): Learning from Data; Springer; New York; LNS120.

Pearl, J (1998): Why there is no statistical test for confounding, why many think there is and why they are almost right; UCLA; Technical Report; R-256.

Rosenbaum, P.R.; Rubin, D. (1983): The Central Role of the Propensity Score in Obervational Studies for Causal Effects; Biometrica; 70,1; S. 41-55.

Rubin, D.B. (1974): Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies; Journal of Ed. Psychology; 66; 688-701.

SAS Institute (1990): SAS Procedures Guide; Cary; NC.

Scheines, R.; Spirtes, P.; Glymour, C.; Meek, C.; Richardson, T. (1996): Tetrad III; Tools for Causal Modeling; Lawrence Erlbaum Association; Inc.; Hillsdale NJ.

Spirtes, P.; Scheines, R.; Glymour, C. (1993): Causation, Prediction and Search; Springer; New York.

Wermuth, N.; Lauritzen, S. L. (1983): Graphical and Recursive Models for Contingency Tables; Biometrika; 70; 537-52.