

Graphische Modellierung in der Kundenzufriedenheitsanalyse

Dietrich Eherler

Lehrstuhl für Statistik, Wirtschaftswissenschaftliche Fakultät
Friedrich-Schiller-Universität Jena
Carl-Zeiss-Str. 3, 07743 Jena
e-mail: d.eherler@wiwi.uni-jena.de

Abstract

Mit der graphischen Modellierung im Rahmen der Abhängigkeitsanalyse ist seit Anfang der 90er Jahre eine neue Methodik der multivariaten Datenanalyse entstanden, welche insbesondere auch als Data-Mining Instrument an Bedeutung gewinnt. Dieser Beitrag berichtet über eine Anwendung der graphischen Modellierung im Bereich der Kundenzufriedenheitsuntersuchung. Im einzelnen sollen Daten zur Einschätzung verschiedener Merkmale von Mobiltelefonen aus einer Kundenbefragung analysiert werden. Bei der Generierung graphischer Modelle ist die Feststellung von marginalen und bedingten Unabhängigkeitsbeziehungen von entscheidender Bedeutung. Die hierfür zur Verfügung stehenden SAS-Prozeduren zur Untersuchung nominaler und respektive ordinaler Daten kommen hierfür zur Anwendung und die Resultate der verschiedenen Ansätze werden verglichen.

1 Einleitung

Ziel jeder Modellbildung ist es, ein problemadäquates Abbild der Wirklichkeit zu schaffen mit Hilfe dessen tiefere Einsichten in den betrachteten Sachverhalt gewonnen werden sollen. Die graphische Modellierung bedient sich dabei eines sehr intuitiven Darstellungsmediums und ermöglicht, komplexe, multidimensionale Problemstellungen visuell zu repräsentieren in einer übersichtlichen Art und Weise. Die Untersuchung von Einflußfaktoren von Kundenzufriedenheit bzw. Präferenzen, die schließlich zu Kaufentscheidungen führen, stellen ein ideales Feld der graphischen Modellierung im Rahmen der Abhängigkeitsanalyse dar. Von Interesse ist die Kenntnis der relevanten Merkmale sowie deren gegenseitige Beeinflussung. Für beide Fragestellungen liefert die Methodik Antworten. Darüber hinaus können, sofern entsprechende Identifizierbarkeitskriterien erfüllt sind, Ursache-Wirkungszusammenhänge zur Evaluation von Effekten externer Eingriffe ausgenutzt werden, um somit Vorhersagen über aktive Steuerungsmaßnahmen mit Techniken des Marketing auf ihre potentielle Wirksamkeit hin zu untersuchen.

Im Anschluß an diese Einleitung folgt eine kurze Darlegung des theoretischen Fundaments auf dem die graphische Darstellung abhängigkeitsanalytischer Modelle beruht. Der Abschnitt 3 stellt drei Konstruktionsprinzipien von Graphen vor. Statistische Aspekte und deren Realisierung in SAS beinhaltet der 4. Abschnitt. Ergebnisse der Untersuchung von Daten zur Einschätzung von Merkmalen von Mobiltelefonen werden im Abschnitt 5 präsentiert. Den Abschluß dieses Beitrages bildet eine Zusammenfassung und ein Ausblick.

2 Abhängigkeitsanalyse und graphische Modellierung

Mit dem Ansatz der graphischen Darstellung von abhängigkeitsanalytischen Modellen ist seit Anfang der 90er Jahre eine neue Methodik der multivariaten Datenanalyse in Erscheinung getreten (Spirtes et al.; 1993, Edwards, 1995). Eine Definition für Abhängigkeitsanalyse gibt Fayyad et al. (1996, S. 15). Er bezeichnet diese als einen Ansatz der statistischen Modellbildung mit dem Ziel, signifikante Abhängigkeiten zwischen Variablen des untersuchten Bereiches zu beschreiben.

2.1 Abhängigkeitsmodell

Ein abhängigkeitsanalytisches Modell kann in zweierlei Formen auftreten. Zunächst kann es implizit in Form einer gemeinsamen Wahrscheinlichkeitsverteilung \mathbf{P} einer Menge von Zufallsvariablen $V = \{X_1, \dots, X_n\}$ gegeben sein. Mit einer solchen Verteilung \mathbf{P} sind sämtliche strukturellen Interdependenzen numerisch spezifiziert und verifizierbar. Ist jedoch die Zielstellung eine graphische Darstellung dieser Abhängigkeiten, so eignet sich dazu die zweite, strukturelle bzw. qualitative Darstellung eines abhängigkeitsanalytischen Modells.

Definition: (Pearl, 1988, S. 88)

Gegeben sei eine endliche Menge V von Zufallsvariablen und X, Y, Z disjunkte Teilmengen von V . Ein Abhängigkeitsmodell M ist dann definiert als eine Liste von bedingten Unabhängigkeitsbeziehungen der Gestalt $I(X, Z, Y)$ mit der Bedeutung, X ist unabhängig von Y gegeben Z bezüglich einer Verteilung \mathbf{P} von V .

Bemerkung 2.1:

1. Jede Wahrscheinlichkeitsverteilung einer endlichen Menge von Zufallsvariablen V induziert ein Abhängigkeitsmodell, da für jedes Triplet (X, Y, Z) von disjunkten Teilmengen von V überprüft werden kann, ob $I(X, Z, Y)$ gilt.
2. Ist M das von einer gemeinsamen Verteilung von V abgeleitete Abhängigkeitsmodell, so liefert M eine qualitative Darstellung der Struktur von \mathbf{P} bzgl. marginaler und bedingter Unabhängigkeit.
3. Im Rahmen dieses Beitrags werden nur diskrete Zufallsvariablen betrachtet, um die Schreibweisen nicht unnötig zu komplizieren.

2.2 Graphen

Zur Abbildung inhaltlich bedeutsamer Abhängigkeiten haben sich gerichtete azyklische Graphen als besonders geeignet herauskristallisiert. Wesentliches Merkmal gerichteter azyklischer Graphen sind eindeutige Eltern-Kind-Beziehungen bezüglich ihrer Knoten.

Ein den qualitativen probabilistischen Unabhängigkeitsbeziehungen entsprechender graphischer Separationsbegriff ist der folgende:

Definition: (Pearl, 1988, S. 117)

Gegeben sei ein gerichteter azyklischer Graph $G(V, E)$. Seien X und Y zwei Knoten von G und sei Z eine Teilmenge der Knoten mit $X, Y \notin Z$. Dann wird X durch Z von Y d-separiert (Symbol: $\langle X | Z | Y \rangle$), 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 Z , oder
2. W besitzt nicht aufeinanderzulaufende Kanten $\rightarrow W \rightarrow$, $\leftarrow W \rightarrow$ oder $\leftarrow W \leftarrow$ und W gehört zu Z .

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

Zur Konstruktion von Graphen aus gegebenem Datenmaterial spielt folgende Eigenschaft der d-Separation eine wichtige Rolle.

Lemma 2.1: (Eherler/Friedel, 1999)

Gegeben sei ein gerichteter azyklischer Graph G . Sind zwei Knoten X, Y notwendigerweise durch eine nicht leere Menge Z d-separiert, so sind X und Y auch d-separiert durch eine Menge Z' mit $Z' \subseteq [\text{Adj}(X) \cup \text{Adj}(Y)]$, wobei $\text{Adj}(X)$ und $\text{Adj}(Y)$ die Adjazenzmengen von X bzw. Y im zu G zugehörigen ungerichteten Graphen sind.

2.3 Graphen und Abhängigkeitsmodelle

Mit dem d-Separationskriterium auf graphischer Seite und der Eigenschaft der Unabhängigkeit auf probabilistischer Seite stehen zwei qualitative Begriffe zur Verfügung, für welche es nun das Ziel ist, diese miteinander zu verknüpfen.

Definition: (vgl. Pearl, 1988, S. 93)

Gegeben sei ein Abhängigkeitsmodell M und ein gerichteter azyklischer Graph $G(V, E)$ über eine endliche Variablenmenge V . Seien X, Y, Z disjunkte Teilmengen von V .

(i) Der gerichtete azyklische Graph $G(V, E)$ ist eine Unabhängigkeitsabbildung (I-Map) für das Abhängigkeitsmodell M , falls gilt:

$$\langle X | Z | Y \rangle \Rightarrow I(X, Z, Y)$$

(ii) Der gerichtete azyklische Graph $G(V, E)$ ist eine Abhängigkeitsabbildung für das Abhängigkeitsmodell M , falls gilt

$$I(X, Z, Y) \Rightarrow \langle X | Z | Y \rangle$$

(iii) Der gerichtete azyklische Graph $G(V, E)$ ist eine perfekte Abbildung (perfekter Graph) für das Abhängigkeitsmodell M , falls gilt

$$I(X, Z, Y) \Leftrightarrow \langle X | Z | Y \rangle.$$

3 Konstruktion von Graphen

Wie Graphen mit wahrscheinlichkeitstheoretischen Implikationen konstruiert werden können, soll anhand von drei Prinzipien dargestellt werden. Es werden dabei jeweils verschiedene Annahmen getroffen und in Abhängigkeit der Annahmen ist dann auch die Aussagekraft des resultierenden Graphen.

3.1 Konstruktion aus substantiellen Annahmen

Aus substanzwissenschaftlichen Überlegungen sei es möglich, ein rekursives System von im allgemeinen nichtlinearen Gleichungen anzugeben der Gestalt

$$X_i = f_i(\mathbf{PA}_i, U_i) \quad 1 \leq i \leq n$$

mit $\mathbf{PA}_i \subseteq V \setminus \{X_i\}$ und U_i paarweisen unabhängigen Störtermen, welches die betrachteten Variablen miteinander in Verbindung setzt. Die Idee dabei ist, daß durch die Funktion f_i diejenigen Variablen aus der Menge V , welche einen direkten Einfluß auf X_i haben, spezifiziert werden. Sonstige Einflüsse, die außerdem auf X_i einwirken, werden durch den Störterm U_i modelliert.

Der Graph zum obigen Gleichungssystem wird durch gerichtete Kanten der Knoten von \mathbf{PA}_i nach X_i gebildet. Der resultierende Graph hat dann aufgrund der Annahmen der Rekursivität und der Unabhängigkeit der Störterme die Eigenschaft, azyklisch gerichtet und eine I-Map zur gemeinsamen Verteilung \mathbf{P} von V zu sein (Pearl, 1996, S. 27), für welche gilt

$$\mathbf{P}(V) = \mathbf{P}(X_1=x_1, \dots, X_n=x_n) = \prod_{i=1}^n \mathbf{P}(X_i=x_i | \mathbf{PA}_i=pa_i).$$

3.2 Konstruktion aus gemeinsamer Verteilung bei gegebener Ordnung

Gegeben sei die Menge V von Zufallsvariablen mit einer durch die Indizierung fest vorgegebenen Ordnung und deren gemeinsame Verteilung \mathbf{P} . Für jede Variable X_i seien minimale Mengen $\mathbf{PA}_i \subseteq \{X_1, \dots, X_{i-1}\}$ ($\mathbf{PA}_0 = \emptyset$) bestimmt mit der Eigenschaft

$$\mathbf{P}(X_i=x_i | X_1=x_1, \dots, X_{i-1}=x_{i-1}) = \mathbf{P}(X_i=x_i | \mathbf{PA}_i=pa_i)$$

Der Graph, welcher wiederum Kanten von den Elementen von \mathbf{PA}_i nach X_i enthält, hat die Eigenschaft, gerichtet azyklisch und eine I-Map zur gemeinsamen Verteilung \mathbf{P} von V zu sein (Pearl, 1988, S. 119).

3.3 Konstruktion aus Unabhängigkeitsbeziehungen

Die Eigenschaft der Mengen \mathbf{PA}_i im Abschnitt 3.2 implizieren, daß $I(X_i, \mathbf{PA}_i, \{X_1, \dots, X_{i-1}\} \setminus \mathbf{PA}_i)$ gilt. Zur Konstruktion des Graphen geht aber essentiell die vorgegebene Ordnung mit ein. Eine Annahme, in der die Ordnung der Variablen nicht gefordert wird, ist, daß zu der betrachteten gemeinsamen Verteilung \mathbf{P} ein perfekter Graph existiert. Unter dieser Annahme haben Spirtes et al. (1993) einen Algorithmus vorgestellt, der es ermöglicht, in polynomieller Zeit (abhängig vom Grad des Graphen) einen Graphen aus Unabhängigkeitsbeziehungen zu konstruieren. Dieser sogenannte PC Algorithmus ist im Rahmen eines Softwarepakets Tetrad II/III implementiert (Scheines et al., 1994).

Von entscheidender Bedeutung ist dabei die Eigenschaft der d-Separation, daß bereits benachbarte Knoten im Graphen zu dieser Trennungseigenschaft (Lemma 2.1) und somit in der Wahrscheinlichkeitsebene zu bedingter Unabhängigkeit beitragen.

Der Algorithmus besteht aus zwei Phasen, einer Kantengenerierungs- und einer Orientierungsphase. In der Kantengenerierungsphase wird ein Abhängigkeitsmodell aufgestellt, indem in einem ineinander greifenden Prozeß eine Kante zwischen X und Y ausgehend von einem vollständigen ungerichteten Graphen entfernt wird, falls eine Beziehung $I(X, \mathbf{Z}, Y)$ verifiziert wurde. Die Bedingungs Menge \mathbf{Z} ist eine Teilmenge der noch zu X und Y adjazenten Knoten im aktuellen Graphen. Die Orientierungsphase richtet Objekte der Gestalt $X-W-Y$ zu $X \rightarrow W \leftarrow Y$, falls W nicht in der Bedingungs Menge \mathbf{Z} enthalten ist, welche $I(X, \mathbf{Z}, Y)$ erfüllt, und falls gilt $\neg I(X, W, Y)$ (vgl. Definition der d-Separation).

Resultat dieses Algorithmus ist ein Muster, welches diejenigen gerichteten azyklischen Graphen repräsentiert, die aufgrund von Unabhängigkeitsbeziehungen nicht unterscheidbar sind. Solche Graphen heißen markoväquivalent (Scheines et al., 1994, S. 33-35).

4 Statistische Aspekte

Wie aus den vorherigen Abschnitten ersichtlich wird, spielen zur graphischen Modellierung von Abhängigkeitsmodellen marginale und bedingte Unabhängigkeitsbeziehungen die zentrale Rolle. In der hier untersuchten Fragestellung liegen die Daten von Einschätzungen verschiedener Merkmale von Mobiltelefonen zugrunde. Diese sind von ordinalem Meßniveau. Für ordinal skalierte Daten stehen als Assoziationsmaße zunächst Rangkorrelationen zur Verfügung, welche das Meßniveau explizit ausnützen, oder z.B. das klassische Chi-Quadrat, wobei hierfür die Ränge keine Berücksichtigung finden.

4.1 Die Daten

Das Datenmaterial stammt aus einer repräsentativen Umfrage von Kaufinteressenten von Mobiltelefonen. Insgesamt stehen 326 Datensätze zur Verfügung, die Auskunft über die

Wichtigkeit (bewertet auf einer Skala von 1 (unwichtig) bis 5 (sehr wichtig)) folgender Faktoren von Mobiltelefonen geben.

Faktor	Kürzel
Produktqualität	<i>A</i>
Preisniveau	<i>B</i>
hoher Sicherheitsstandard	<i>C</i>
Haltbarkeit/Verschleiß	<i>D</i>
Benutzerfreundlichkeit	<i>E</i>
Reparaturservice	<i>F</i>
Bekannte Marke/gutes Image	<i>G</i>
Ausreichendes Katalog und Infomaterial	<i>H</i>
Garantiezeit	<i>I</i>

Tabelle 1: Untersuchte Faktoren

Darüber hinaus wurde befragt, inwieweit einzelne Produzenten von Mobiltelefonen diesen Anforderungen genügen. Zur Darstellung der Idee der graphischen Modellierung beschränkt sich die Betrachtung hier auf die ersten Umfrageergebnisse.

4.2 Die nominale Sichtweise

Werden die Datensätze nominal skaliert betrachtet, so bietet sich als Test auf marginale Unabhängigkeit der in SAS implementierte Chi-Quadrat Unabhängigkeitstest bzw. der Likelihood-Quotienten Unabhängigkeitstest an. Beide sind über die PROC FREQ unter der Option CHISQ im TABLE Statement abrufbar. Einen allgemeineren und multivariaten Ansatz bietet die Mantel-Haenzel Methodik (z.B. SAS-Institute, 1995, S. 106; SAS Institute, 1996). Auch diese ist innerhalb SAS in der Prozedur FREQ realisiert. Die Programmzeilen haben die folgende Struktur:

```
proc freq data=Datei;
table Xi*Xj/cmh;      *Test auf marginale Unabhängigkeit;
table Xk*Xi*Xj/cmh;  *Test auf bedingte Unabhängigkeit I(Xi,Xk,Xj);
table Xl*Xk*Xi*Xj/cmh; *Test auf bedingte Unabhängigkeit I(Xi,{Xk,Xl},Xj);
run;
```

Der Vorgehensweise des Algorithmus folgend werden sämtliche Tests auf Unabhängigkeit durchgeführt. Kann die Nullhypothese auf marginale bzw. bedingte Unabhängigkeit $I(X,Z,Y)$ nicht abgelehnt werden, wird die entsprechende Kante zwischen diesen beiden Variablen X und Y aus dem Graphen entfernt. Eine Liste der einzelnen nicht abgelehnten Nullhypothesen wird in Abschnitt 5.1 wiedergegeben.

Wiederum nach der Vorschrift des Algorithmus werden nach Beendigung der Kantengenerierungsphase möglichst viele Kanten orientiert. Weitere Tests sind dazu erforderlich, ob für Variablen eines Objektes der Gestalt $X-W-Y$ im noch ungerichteten Graphen die Bedingungen $I(X,Z,Y)$ mit $W \notin Z$ und $\neg I(X,W,Y)$ gelten, und somit eine Orientierung zu $X \rightarrow W \leftarrow Y$ erfolgen kann.

4.3 Die ordinale Sichtweise

Durch die nominale Sichtweise der Daten geht Information verloren, welche in der Ordnung der eingeschätzten Wichtigkeit der Faktoren steckt. Ferner leiden die Tests auf bedingte Unabhängigkeit für nominal skalierte Daten an ihrer geringen Machtfunktion (SAS Institute,

1990, S. 348). Dagegen muß einschränkend erwähnt werden, daß Assoziationsmaße, welche für ordinal skalierte Variablen konzipiert sind, lediglich auf monotone Zusammenhänge untersuchen und nur in Spezialfällen auf Unabhängigkeit der untersuchten Variablen geschlossen werden kann (Büning/Trenkler, 1993, S. 239). Dies schränkt die Aussagekraft der Graphen auf monotone Assoziationen ein, sofern nicht anders, z.B. inhaltlich begründbar ist, daß auf Unabhängigkeit im probabilistischen Sinn geschlossen werden kann.

4.3.1 Der Spearman'sche Rangkorrelationskoeffizient

Der Rangkorrelationskoeffizient von Spearman berechnet sich wie der gewöhnliche Korrelationskoeffizient von Pearson, nur daß statt der Stichprobenwerte selbst die Ränge dieser verwendet werden. Formal dargestellt bei Vorliegen einer Stichprobe ist der Koeffizient von Spearman in der folgenden Definition.

Definition: (vgl. Gibbons/Chakraborti, 1992, S. 357)

Seien $R_i = R(X_i)$ und $R_j = R(Y_j)$ für $i, j = 1, \dots, n$ die jeweiligen Ränge einer Stichprobe vom Umfang m zu den Variablen X und Y . Dann heißt

$$R_{XY} = \frac{\sum_{i=1}^m (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^m (R_i - \bar{R})^2 \sum_{i=1}^m (S_i - \bar{S})^2}}$$

mit $\bar{R} = \bar{S} = \frac{m+1}{2}$ der Rangkorrelationskoeffizient von X und Y nach Spearman.

Sind die Zufallsvariablen X_1, \dots, X_n gegeben, so ist die symmetrische Matrix mit den Einträgen der paarweisen Spearman'schen Rangkorrelationskoeffizienten

$$\mathbf{R} = \begin{pmatrix} 1 & R_{X_1, X_2} & \cdots & R_{X_1, X_n} \\ R_{X_2, X_1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & R_{X_{n-1}, X_n} \\ R_{X_n, X_1} & \cdots & R_{X_n, X_{n-1}} & 1 \end{pmatrix}$$

definiert als die Spearman'sche Rangkorrelationsmatrix. Mit ihrer Hilfe ist der partielle Spearman'sche Rangkorrelationskoeffizient definiert als

$$R_{ij/\{1, \dots, n\} \setminus \{i, j\}} = \frac{|\mathbf{R}|}{|\mathbf{R}_{ij}|}$$

wobei $|\mathbf{R}|$ die Determinante von \mathbf{R} ist und $|\mathbf{R}_{ij}|$ die Determinante der Matrix, die durch Streichung der i -ten Zeile und j -ten Spalte aus \mathbf{R} entsteht.

In SAS ist ein Test auf verschwindende (partielle) Spearman'sche Rangkorrelationskoeffizienten implementiert und in der PROC CORR verfügbar. Die Programmzeilen haben die folgende Struktur:

```
PROC CORR SPEARMAN DATA = DATEI;
VAR Xi, Xj;
PARTIAL Z ⊆ {X1, ..., Xn} \ {Xi, Xj};
RUN;
```

Im Falle von Bindungen (Büning/Trenkler, 1994, S. 43), d.h. die Ränge sind nicht eindeutig jeder Stichprobenvariablen zuzuordnen, verwendet SAS Durchschnittsränge (SAS Institute, 1990, S. 221). Dadurch ändert sich aber die Verteilung der Teststatistik unter der Nullhypothese, welche nur im Fall keiner Bindungen exakt angegeben wird (Gibbons/Chakraborti, 1992, S. 367). Deshalb werden neben den standardmäßig mit SAS bestimmten Spearman'schen Rangkorrelationskoeffizienten auch noch die auf die Definition zurückgreifende Pearson'sche Korrelationskoeffizienten der Rohdaten berechnet, welche als Ränge vorliegen. Auch hier bietet SAS Tests auf verschwindende (partielle) Korrelationskoeffizienten. Die approximative Verteilungsaussage der Teststatistik gilt auch bei dieser Betrachtungsweise (z.B. Scheines et al., 1994, S. 232).

```
PROC CORR DATA = DATEI;
VAR Xi,Xj;
PARTIAL Z  $\subseteq$  {X1, ..., Xn} \ {Xi, Xj};
RUN;
```

Zur Konstruktion eines gerichteten azyklischen Graphen wird völlig analog vorgegangen. Ein verschwindender (partieller) Rangkorrelationskoeffizient führt zur Löschung einer Kante in der Kantengenerierungsphase. Die Orientierung erfolgt ebenfalls unter Beachtung der oben beschriebenen Regeln.

4.3.2 Kendall's Tau

Ein weiteres Assoziationsmaß für ordinal skalierte Daten ist der Rangkorrelationskoeffizient nach Kendall (Gibbons/Chakraborti, 1992, S. 339). Für diesen ist zwar in analoger Weise wie für den Spearman'schen Koeffizienten ein partieller Koeffizient definiert (SAS Institute, 1996, S.222), jedoch gibt es für den Fall, bei dem die Bedingungsmenge die Mächtigkeit 1 übersteigt (Maghsoodloo /Pallos, 1981, S. 47), keine Aussagen, welcher Verteilung eine adäquate Teststatistik genügt. SAS bietet lediglich einen Test auf verschwindende marginale Korrelationskoeffizienten nach Kendall, so daß mit diesem Maß keine weiteren Untersuchungen durchgeführt wurden. Hier besteht also noch Forschungsbedarf.

5 Ergebnisse

Die Ergebnisse mit den verschiedenen Testprozeduren fallen sehr unterschiedlich aus, und es tritt auch die von den Entwicklern der Verfahren erkannte nicht unbedingt wünschenswerte Eigenschaft der Ergebnissensitivität von den Testentscheidungen zu Tage. Doch zunächst die konkreten Resultate.

5.1 Konstruierte Graphen

Die folgende Liste gibt das Abhängigkeitsmodell wieder und die zugehörigen p-Werte (bei einem Signifikanzniveau von $\alpha=5\%$) der nominalen Sichtweise.

$$\mathbf{M}_{\text{NOM}} = \{I(B, \emptyset, C) [p=0.101]; I(B, \emptyset, G) [p=0.070]; I(B, \emptyset, I) [p=0.067]; I(E, \emptyset, G) [p=0.068]; I(E, \emptyset, H) [p=0.050]; I(A, E, B) [p=0.086]; I(A, H, C) [p=0.065]; I(A, F, D) [p=0.224]; I(A, D, F) [p=0.713]; I(B, F, D) [p=0.994]; I(B, C, F) [p=0.108]; I(C, D, E) [p=0.057]; I(C, D, F) [p=0.503]; I(D, I, F) [p=0.316]; I(D, H, G) [p=0.147]; I(E, D, F) [p=0.215]; I(E, D, I) [p=0.993]; I(F, G, H) [p=0.215]; I(C, \{D, H\}, G) [p=0.051]; I(C, \{E, G\}, D) [p=0.082]; I(A \{C, G\}, H) [p=0.051]\}$$

Der resultierende Graph 1 hat dann die folgende Gestalt. Doppelt gerichtete Pfeile weisen auf latente Variablen hin, welche die Abhängigkeit der so miteinander verbundenen Variablen aufheben würde (Scheines et al, 1994, S. 91).

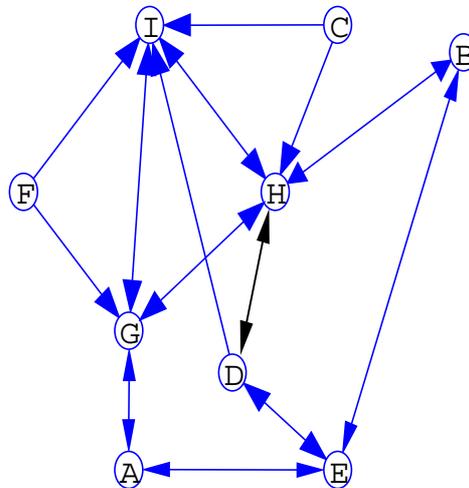


Abbildung 1: Graph 1

Bei Verwendung der standardmäßigen SAS-Routine für den Spearman'schen Rangkorrelationskoeffizienten ergibt sich aus der Rangkorrelationsmatrix R (p-Werte unter den jeweiligen Koeffizienten)

	A	B	C	D	E	F	G	H	I
A	1	0,148	0,194	0,216	0,239	0,200	0,203	0,212	0,147
	0	0,008	0,000	0,000	0,000	0,000	0,000	0,000	0,008
B	0,148	1	0,111	0,033	0,089	0,105	0,077	0,107	0,061
	0,008	0	0,044	0,555	0,111	0,059	0,168	0,054	0,273
C	0,194	0,111	1	0,186	0,201	0,134	0,182	0,162	0,133
	0,000	0,044	0	0,001	0,000	0,015	0,001	0,003	0,016
D	0,216	0,033	0,186	1	0,373	0,324	0,228	0,212	0,364
	0,000	0,555	0,001	0	0,000	0,000	0,000	0,000	0,000
E	0,239	0,089	0,201	0,373	1	0,236	0,106	0,145	0,195
	0,000	0,111	0,000	0,000	0	0,000	0,055	0,009	0,000
F	0,200	0,105	0,134	0,324	0,236	1	0,192	0,240	0,424
	0,000	0,059	0,015	0,000	0,000	0	0,001	0,000	0,000
G	0,203	0,077	0,182	0,228	0,106	0,192	1	0,356	0,255
	0,000	0,168	0,001	0,000	0,055	0,001	0	0,000	0,000
H	0,212	0,107	0,162	0,212	0,145	0,240	0,356	1	0,346
	0,000	0,054	0,003	0,000	0,009	0,000	0,000	0	0,000
I	0,147	0,061	0,133	0,364	0,195	0,424	0,255	0,346	1
	0,008	0,273	0,016	0,000	0,000	0,000	0,000	0,000	0

Tabelle 2: Rangkorrelationsmatrix nach Spearman

ebenfalls mit einem Signifikanzniveau von $\alpha=5\%$ der folgende (Korrelations-)Graph 2. Nicht alle Kanten können orientiert werden. Graph 2 repräsentiert also diejenigen gerichteten azyklischen Graphen, die bezüglich der d-Separation die gleichen Unabhängigkeitsbeziehungen implizieren.

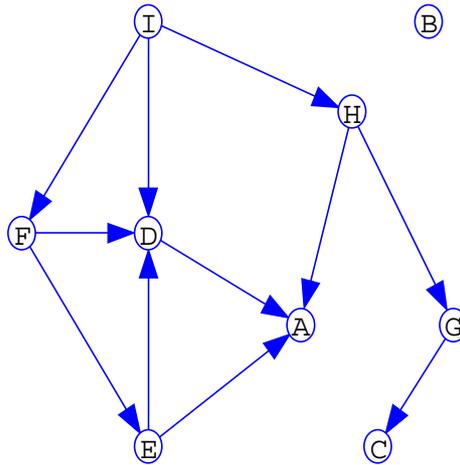


Abbildung 2: Graph 2

Werden die Daten, welche schon als Ränge vorliegen mit dem auf die Definition zurückgreifenden Pearson'schen Korrelationskoeffizienten untersucht, resultiert die Korrelationsmatrix

	A	B	C	D	E	F	G	H	I
A	1	0,124	0,163	0,251	0,189	0,194	0,206	0,204	0,182
B	0	1	0,109	0,041	0,106	0,136	0,079	0,093	0,083
C	0,025	0	1	0,463	0,057	0,014	0,154	0,092	0,133
D	0,163	0,109	0,165	1	0,278	0,319	0,235	0,201	0,388
E	0,003	0,050	0	0,003	1	0,003	0,000	0,004	0,002
F	0,251	0,041	0,165	0,278	0,319	1	0,235	0,201	0,388
G	0,000	0,463	0,003	0	0,000	0,000	1	0,000	0,000
H	0,189	0,106	0,166	0,278	0,319	0,221	0,114	1	0,120
I	0,001	0,057	0,003	0,000	0	0,000	0,040	0,030	1
A	0,194	0,136	0,110	0,319	0,221	1	0,216	0,237	0,434
B	0,000	0,014	0,048	0,000	0,000	0	0,000	0,000	0,000
C	0,206	0,079	0,207	0,235	0,114	0,216	1	0,357	0,262
D	0,000	0,154	0,000	0,000	0,040	0,000	0	0,000	0,000
E	0,204	0,093	0,158	0,201	0,120	0,237	0,357	1	0,339
F	0,000	0,092	0,004	0,000	0,030	0,000	0,000	0	0,000
G	0,182	0,083	0,168	0,388	0,132	0,434	0,262	0,339	1
H	0,001	0,133	0,002	0,000	0,017	0,000	0,000	0,000	0

Tabelle 3: Rangkorrelationsmatrix nach Definition

und daraus der (Korrelations-)Graph 3.

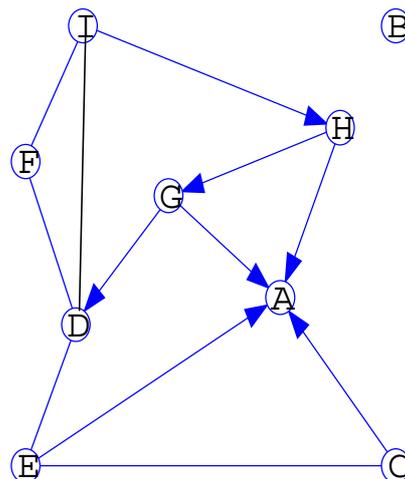


Abbildung 3: Graph 3

5.2 Vergleich

Als erstes kann bei einem Vergleich der konstruierten Graphen festgestellt werden, daß im Falle der nominalen Sichtweise mehr Kanten im Graphen enthalten sind. Dies ist auch konzeptionell erklärbar, werden doch bei dieser Sichtweise Abhängigkeiten betrachtet, welche die monotonen Abhängigkeiten miteinschließen.

Stabile Kanten, im Sinne, daß sie in allen Graphen erscheinen, sind Verbindungen zwischen den Faktoren Haltbarkeit und Benutzerfreundlichkeit, Garantiezeit und ausreichendes Prospektmaterial, bekannte Marke und ausreichendes Prospektmaterial, Garantiezeit und Reparaturservice, Produktqualität und Benutzerfreundlichkeit. Die Einschätzungen dieser Merkmale hängen also zusammen.

Ebenfalls bedeutsam für die Interpretation sind die Kanten, die nicht in den Graphen vorhanden sind bzw. wo aufgrund von d-Separationsbeziehungen auf (lineare) Unabhängigkeit geschlossen werden kann. Werden nur monotone Abhängigkeiten betrachtet, also die Graphen 2 und 3, so ist der Faktor Preisniveau völlig isoliert von den andern Faktoren. Weitere d-Separationsbeziehungen können aus den Graphen abgelesen werden.

5.3 Interpretation

Eine bemerkenswerte Bedeutung kommt Graphen zu, wenn sie als Darstellung eines kausalen Modells (Pearl, 1995) interpretiert werden. Dann ist es möglich, kontrafaktische Wahrscheinlichkeiten zu bestimmen, welche Aussagen geben, wie sich das betrachtete System unter externen Eingriffen verhalten würde. So könnte aus den Graphen 2 und 3 näherungsweise geschätzt werden, wie durch bewußte Manipulation des Prospektmaterial die Einschätzung der Produktqualität beeinflusst werden kann. Allerdings würde das den Rahmen der Betrachtungen hier sprengen. Der interessierte Leser findet in folgenden Quellen die entsprechenden Methoden beschrieben: Spirtes et. al. (1993), Pearl (1995), Galles/Pearl (1998).

6 Zusammenfassung

Dieser Beitrag gibt einen Einblick in neuere Ansätze der graphischen Modellierung von Abhängigkeitsmodellen. Es werden die theoretischen Grundlagen dargestellt um zu motivieren, wie sich einerseits wahrscheinlichkeitstheoretische Modelle mit der andererseits graphischen Darstellungen miteinander bedeutungsvoll verbinden lassen. Die Begriffe Unabhängigkeit und d-Separation spielen das Bindeglied. Soll nun ein graphisches Modell aus Daten gelernt wer-

den, so bedarf es bei dem in diesem Beitrag vorgestellten Verfahren der Durchführung von Unabhängigkeitstests. Diese wurden mit den SAS Prozeduren `FREQ` und `CORR` realisiert. Als Datenmaterial standen die Einschätzungen der Wichtigkeit verschiedener Faktoren bei Mobiltelefonen aus einer Umfrage zur Verfügung. Die Ergebnisse sind zumindest bei der nominalen Betrachtungsweise der Daten nicht sehr vielversprechend. Ein besseres Bild liefern die Untersuchung mit Rangkorrelationen. Hier muß aber einschränkend angemerkt werden, daß die Aussagekraft der Graphen auf monotone Zusammenhänge reduziert werden muß. Sind Graphen an und für sich schon ein sehr anschauliches und leicht kommunizierbares Darstellungsmedium, wird am Schluß dieses Beitrages noch darauf hingewiesen, welche weiteren Potentiale der Interpretation in Graphen stecken.

Literatur

- Bünig, H. ; Trenkler, G. (1994): Nichtparametrische statistische Methoden; 2. Aufl.; Berlin [u.a.]; de Gruyter.
- Edwards, D., Introduction to Graphical Modelling, Springer, New York, Berlin 1995.
- Eherler, D.; Friedel, S. (1999): Abhängigkeitsanalyse und graphische Modellierung - Das Programm Tetrad II und Anwendungsbeispiele auf Kapitalmarktdaten; erscheint als Diskussionspapier, Serie B (Volkswirtschaftslehre); Friedrich-Schiller-Universität Jena; Jena.
- Fayyad, U. et al. (ed.) (1996): Advances in Knowledge Discovery and Data Mining, Menlo Park.
- Galles, D.; Pearl, J. (1998): An Axiomatic Characterization of Causal Counterfactuals, Foundations of Science; 3(1); 151-182.
- Gibbons, J. D.; Chakraborti, S. (1992): Nonparametric Statistical Inference; 3. ed., rev. and expanded; New York [u.a.] : Dekker.
- Maghsoodloo, S.; Pallos, L. (1981): Asymptotic Behaviour of Kendall's Partial Rank Correlation Coefficient and Additional Quantil Estimates; Journal of Statistical Simulation, Vol. 13; S. 41-48.
- Pearl, J. (1988): Probabilistic Reasoning in Intelligent Systems; Morgan Kaufman Publishers; San Mateo, CA.
- 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.
- SAS Institute (1990): SAS Procedures Guide; Cary; NC.
- SAS Institute (1995): Categorical Data Analysis Using the SAS System; Cary; NC.
- SAS Institute (1996): SAS/STAT User's Guide, Volume 1; Cary; NC.
- Scheines, R.; Spirtes, P.; Glymour, C.; Meek, C. (1994): Tetrad II, Tools for Causal Modeling; Lawrence Erlbaum Association, Inc.; Hillsdale NJ.
- Spirtes, P.; Scheines, R.; Glymour, C. (1993): Causation Prediction and Search, Springer, New York.