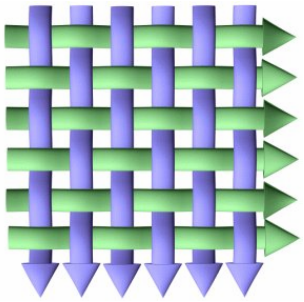
Sonderforschungsbereich 559
Modellierung großer Netze in der Logistik



Technical Report 05005
ISSN 1612-1376

Kriterien für die Kategorisierung statistischer Methoden im Rahmen eines Methodennutzungs-modells zur Informationsgewinnung in GNL

Teilprojekt M9:

Thomas Fender
Anne Krampe
Sonja Kuhnt

Universität Dortmund
Fachbereich Statistik
Institut für Mathematische Statistik und
industrielle Anwendungen
Vogelpothsweg 87

44221 Dortmund

Dortmund, den 27. Juli 2005

Inhalt

1 Einleitung

Die modellbasierte Analyse bzw. die Modellierung von großen Netzen der Logistik (GNL) basiert in der Regel auf geeigneten Informationen über reale Gegebenheiten, z.B. Ankunftszeiten, Produktionsmengen, etc.. Zur Gewinnung solcher Informationen vor allem für die Simulation und Optimierung von GNL soll im Rahmen des Projektes „Informationsgewinnung in großen Netzen der Logistik“ des SFB 559 eine methodenintegrierte und disziplinübergreifende Arbeitsumgebung bereitgestellt werden. Diese Arbeitsumgebung, das Methodennutzungsmodell zur Informationsgewinnung (Abbildung 1), steht in dem übergeordneten Vorgehensmodell zur Modellbildung für GNL (angelehnt an VDI 3633) nach der Erstellung eines konzeptionellen Modells und beinhaltet ein Vorgehensmodell, eine Taxonomie- und eine Methodenebene sowie eine verbindende Metainformationsschicht.

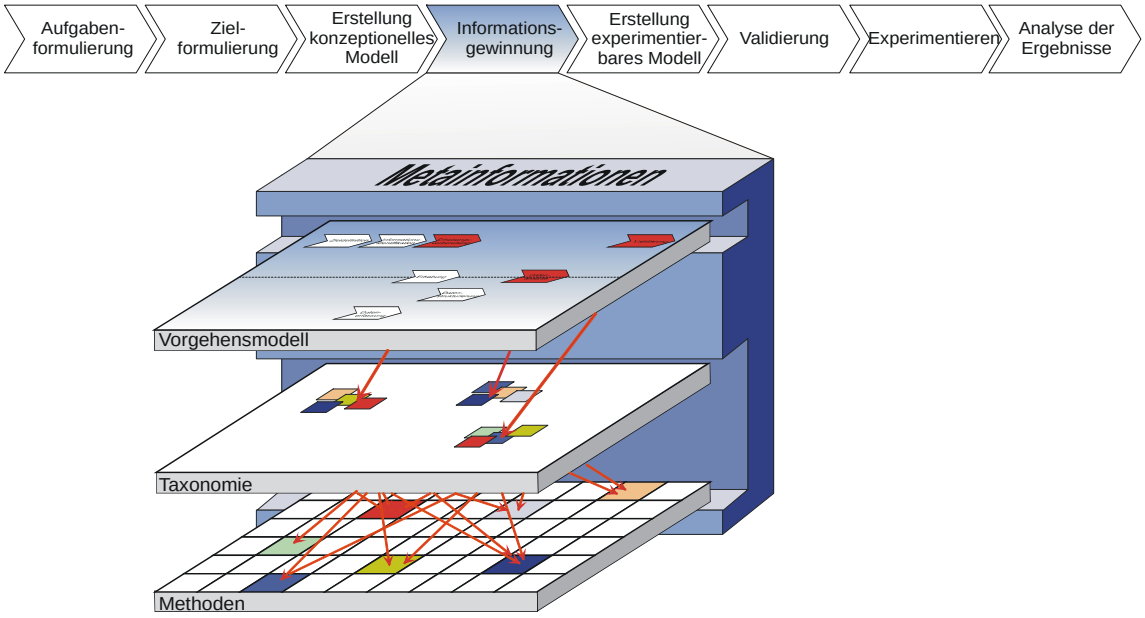


Abbildung 1: Methodennutzungsmodell

Das Vorgehensmodell (Abbildung 2) beschreibt die Prozessschritte von der Ableitung der Aufgabenstellung bis zur Validierung der Eingangsdaten für eine modellbasierte Analyse von GNL (Bernhard et al., 2005). Für die in den Prozessschritten relevanten Aufgabenstellungen werden in der Methodenebene die jeweils adäquaten Methoden bereitgestellt. Dabei müssen insbesondere in den Prozessschritten Erhebungsvorbereitung, Datenanalyse und Validierung aus einer Vielzahl von potentiellen Verfahren die geeigneten Methoden der Datenerhebung, Statistik und Visualisierung gewählt werden. Hierfür wird ein Methodenbaukasten zur Verfügung gestellt werden, in dem sowohl Methoden wie auch eine Taxonomie dieser Methoden enthalten sind.

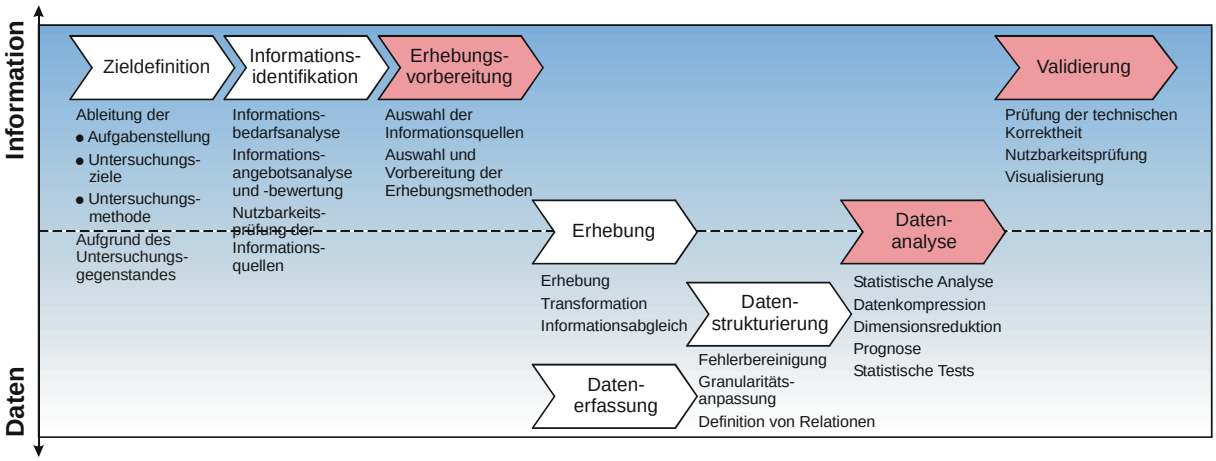


Abbildung 2: Vorgehensmodell

In diesem Beitrag werden für diese Taxonomie Kriterien erarbeitet, nach denen sich Methoden der Statistik klassifizieren lassen. Mit Bezug auf die Klassifizierungskriterien ähnliche Methoden können dann zu typischen Methodenkategorien zusammengefasst werden.

2 Klassifikationsschemata

Bei der Wahl einer geeigneten statistischen Methode für eine statistische Analyse von Daten ist eine Reihe von Aspekten zu beachten. Neben der generellen Fragestellung, die hinter der Analyse steht, gehört dazu die Art und Beschaffenheit der vorliegenden Daten. Primärer Zweck einer statistischen Datenanalyse ist immer die Erlangung bestimmter Informationen bezüglich der vorliegenden Problemstellung. Dabei lassen sich statistische Methoden nicht nur auf ein individuelles sachwissenschaftliches Problem anwenden, sondern auf bestimmte Arten von Fragestellungen. Die Methoden können dann noch danach unterschieden werden, ob und wie viele Variablen als erklärende Variablen (Einflussgrößen) bzw. als Zielvariablen (Zielgrößen) angesehen werden. Auch die Art der Daten in Bezug auf Skalenniveau, Vollständigkeit, Umfang etc. spielt eine Rolle bei der Wahl einer geeigneten statistischen Methode. Nachfolgend werden daher drei Aspekte für die Klassifikation statistischer Methoden herangezogen, die gemäß ihrer sinnvollen inhaltlichen und zeitlichen Reihenfolge wie folgt angeordnet werden: Information, Variablen-Symmetrie und das vorliegende Datenmaterial. Die damit erzielte Einteilung von statistischen Verfahren dient der pragmatischen Bereitstellung der Methoden im Methodennutzungsmodell und erhebt nicht den Anspruch alle Aspekte der Datenanalyse vollständig abzudecken. Es ist auch zu erwarten, dass es Verfahren geben wird, die nicht eindeutig einer Kategorie zugeordnet werden können und es somit zu Überschneidungen bei der Klassifizierung kommen kann.

2.1 Fokus Information

Unter dem Aspekt der Information werden statistische Methoden gemäß des übergeordneten interessierenden Untersuchungsziels kategorisiert. Auf Grund des mathematisch-statistischen Charakters aller Methoden ist zu beachten, dass für eine geeignete Kategorisierung eine exakte Formalisierung des Untersuchungsziels notwendig ist. Daraus ergibt sich in Anlehnung an Backhaus et al. (1996) die Unterscheidung in die drei Bereiche Strukturerkennung, -modellierung und -überprüfung. Diese Begriffe bzw. die damit verbundenen Methoden sind grundsätzlich nicht völlig voneinander getrennt zu behandeln. Vielmehr können und sollen sie sich in einer umfassenden adäquaten statistischen Analyse ergänzen und vervollständigen. Es sind aber auch Überschneidungen denkbar, so dass eine Methode, je nach Anwendung bzw. Untersuchungsziel, in mehrere Bereiche eingeordnet werden kann.

In der Strukturerkennung werden Verfahren zusammengefasst, deren primäres Ziel das Erkennen von Zusammenhängen bzw. Abhängigkeiten von Variablen oder Objekten ist. Vor Beginn der statistischen Untersuchung verfügt der Anwender über keinerlei Wissen bezüglich der Struktur des vorliegenden Datenmaterials. Im Unterschied dazu ist bei der Strukturmodellierung bereits eine vage Vorinformation über die Beschaffenheit (Struktur) der Daten bekannt. Die wesentliche Zielsetzung der Struktur modellierenden Verfahren ist es, die bereits vorhandenen Kenntnisse durch eine Beschreibung von Zusammenhängen zu spezifizieren und funktionale Beziehungen herzustellen. Wie bei der Modellierung besitzt der Anwender bei der Strukturüberprüfung vorab Kenntnisse über die Daten. Diese begründen sich durch sachlogische oder theoretische Überlegungen und beinhalten bereits genauere Annahmen über die Struktur der Daten. Daher besteht die Strukturüberprüfung aus Verfahren, deren zentrale Aufgabe die Prüfung bzw. Überprüfung von angenommenen Zusammenhängen und Strukturen ist.

Zur weiteren Verdeutlichung dieser Begriffe sowie der unterschiedlichen Zielsetzungen werden in Tabelle 1 einige Beispiele für die verschiedenen Bereiche angeführt.

Bereich	Ziel	Statistisches Verfahren
Strukturerkennung	Zusammenfassen von Objekten in Ähnlichkeitsklassen	Clusteranalyse
	Erkennen zugrunde liegender latenter Variablen	Faktorenanalyse
	Erkennen von Abhängigkeits-strukturen	(ungerichtete) graphische Modelle
Strukturmodellierung	Modellierung funktionaler Zusammenhänge	Regressions- / Kovarianzanalyse
	Modellierung funktionaler Zusammenhänge mit zeitlicher Komponente	Zeitreihenanalyse
	Modellierung von Verteilungen	Dichteschätzung
	Modellierung von Abhängigkeitsstrukturen	Kontingenztafelanalyse
Strukturüberprüfung	Einordnung von Objekten in vorgegebene Klassen	Diskriminanzanalyse
	Überprüfung von Hypothesen	Statistische Tests

Tabelle 1: Beispiele von Verfahren und Aufgabenstellungen bei der Strukturerkennung, -modellierung und -überprüfung

Der Fokus auf den Informationsgehalt bzw. die Informationsstrukturen bildet die Grundlage jeder Klassifikation statistischer Methoden, da erst nach der Entscheidung, welche Aufgabenstellung bearbeitet werden soll bzw. kann, die dann noch in Frage kommenden Methoden auf ihre Tauglichkeit bezüglich der untersuchten Variablen und der vorliegenden Daten überprüft werden können.

2.2 Fokus Variablen-Symmetrie

Statistische Verfahren der Datenanalyse werden auf Datensätze angewendet. Der eigentliche Vorgang der Datenerhebung geschieht im Vorgehensmodell (Abbildung 2) vor dem hier interessierenden Schritt der Datenanalyse und wird daher hier nicht näher diskutiert. Es wird davon ausgegangen, dass ein Datensatz vorliegt, der für ausgewählte Objekte Ausprägungen von Merkmalen enthält. Die Merkmale (oder Variablen) sind die eigentlich interessierenden Größen, welche für einzelne Objekte unterschiedliche Werte annehmen können. Zum Beispiel könnte für Läden einer Supermarktkette (Objekte) für ein bestimmtes Produkt die Anzahl verkaufter Stücke (Merkmal mit möglichen Ausprägungen 1,2,…) in einer gegebenen Woche festgehalten werden. Meistens werden mehrere Variablen zur Klärung einer bestimmten Fragestellung erhoben. Die Wahl einer geeigneten statistischen Methode hängt dann auch davon ab, ob diese Variablen symmetrisch oder nicht symmetrisch behandelt werden. Diese Einteilung wird sowohl von sachlogischen Aspekten der betrachteten Variablen als auch von der untersuchten Fragestellung beeinflusst. Bei

einer nicht-symmetrischen Behandlung der Variablen werden diese eingeteilt in erklärende Variablen (auch Kovariablen, unabhängige Variablen) und Zielvariablen (auch Zielgrößen, abhängige Variablen). Dabei wird aus sachlogischen Überlegungen davon ausgegangen bzw. es wird vermutet, dass die erklärenden Variablen den Ausgang der Zielvariablen (funktional) beeinflussen und die Art und Weise dieser Beeinflussung steht oft im Mittelpunkt der Fragestellung. Bei einer symmetrischen Behandlung hingegen werden alle Variablen gleich behandelt. Statistische Verfahren lassen sich danach einteilen, ob sie für symmetrische bzw. nicht-symmetrische Behandlungen von Variablen und damit verknüpfte Fragestellungen geeignet sind. Im nicht-symmetrischen Fall lassen sich noch weitere Einteilungen gemäß der Anzahl der erklärenden Variablen und Zielvariablen treffen und danach, ob im Fall mehrerer Zielvariablen diese untereinander als symmetrisch betrachtet werden oder nicht.

Anhand dieser Strukturen kann eine Einteilung von Verfahren vorgenommen werden, die sich als ein baumartiges Gebilde darstellen lässt. Zur Illustration zeigt Abbildung 3 einen Ausschnitt dieser Klassifizierung.

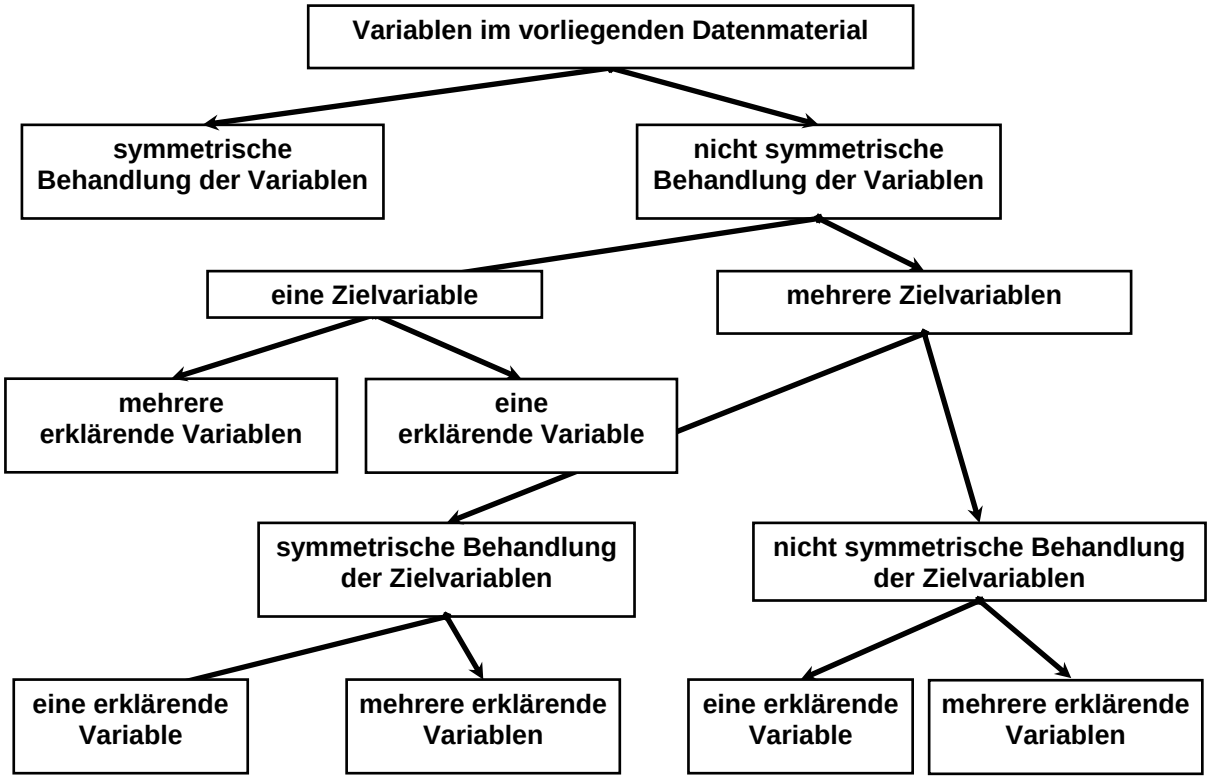


Abbildung 3: Übersicht zum Fokus Variablen-Symmetrie

2.3 Fokus Daten-Input

Ein weiterer Anhaltspunkt für die Auswahl einer statistischen Methode sowie für die Bewertung ihrer Anwendbarkeit ist die Beschaffenheit der zugrunde liegenden Daten. Daher steht im Weiteren das Datenmaterial im Mittelpunkt des Interesses. Die Bewertung des Datenmaterials erfolgt anhand von vorhandenen Kenntnissen über die erhobenen Variablen.

Zudem werden Beurteilungskriterien herangezogen, die bereits nach kurzer Inspizierung des vorliegenden Datensatzes einen ersten Eindruck über die Qualität der Daten vermitteln.

Zunächst wird der Grad des Informationsgehaltes der beobachteten Variablen betrachtet. Es erfolgt eine Unterscheidung in diskrete und stetige Variablen. Diskrete Variablen besitzen endlich viele oder höchstens abzählbar unendlich viele verschiedene Ausprägungen. Stetige Variablen hingegen können jeden beliebigen Wert in einem Messbereich annehmen. Eine weitere Differenzierung bietet das Messniveau oder auch Skalenniveau der Variablen. Entsprechend dem Informationsgehalt der Messung wird zwischen der nominal-, ordinal- und kardinal- bzw. metrischen Skala unterschieden. Werte nominal skalierter Variablen dienen lediglich der (nominalen) Unterscheidung von Untersuchungseinheiten. Weder Reihenfolge noch Abstände nominal skalierter Variablen sind interpretierbar. Bei Variablen mit ordinalem Skalenniveau werden Werte beobachtet, die sinnvoll angeordnet werden können. Die so gewonnene Reihenfolge möglicher Beobachtungen ist interpretierbar, die Abstände zwischen den Werten ordinal skalierter Variablen dagegen nicht. Das metrische Niveau ist das höchste (informativste) Skalenniveau. Die beobachteten Werte der Variablen können wie bei den ordinal skalierten Variablen in einer Reihenfolge angeordnet werden. Zusätzlich sind die Abstände zwischen den Ausprägungen interpretierbar. Nehmen nominal oder ordinal skalierte Variablen nur endlich viele verschiedene Werte an, so werden sie als kategoriale Variablen bezeichnet. Bei Vorliegen von metrischem bzw. kardinalem Skalenniveau wird zusätzlich zwischen intervall- und verhältnisskalierten Variablen unterschieden. Für die Intervallskala ist charakteristisch, dass sie keinen oder nur einen willkürlich festgelegten Nullpunkt besitzt. Damit sind zwar die Abstände (Differenzen) der Messwerte einer metrischen, intervallskalierten Variable interpretierbar, die Quotienten allerdings nicht. Die Verhältnisskala verfügt über einen natürlichen Nullpunkt, der auch die Zahl Null zugeordnet wird. Damit kann das Verhältnis zweier Werte einer metrischen, verhältnisskalierten Variablen sinnvoll durch ihren Quotienten beschrieben werden. Dabei ist zu beachten, dass der Interpretationsgehalt niedrigerer Skalenniveaus im Interpretationsgehalt höherer Skalenniveaus enthalten ist. Eine detaillierte Beschreibung der verschiedenen Skalenniveaus kann in Fahrmeir et al. (2003) nachgelesen werden.

Zur weiteren Verdeutlichung der unterschiedlichen Skalenniveaus werden in Tabelle 2 typische Beispiele sowie die entsprechenden Interpretationsmöglichkeiten angeführt.

Skalenniveau		Beispiele	Interpretation
kardinal/metrisch	**verhältnisskaliert**	Gewicht	Verhältnis
	intervallskaliert	Temperaturangabe in °C	Abstand
ordinal		Schulnoten	Größer-, Gleich-, Kleiner-Beziehungen
nominal		Haarfarbe Geschlecht	Gleichheit bzw. Unterschiedlichkeit

Tabelle 2: Beispiele und Interpretationsmöglichkeiten der verschiedenen Skalenniveaus

Weitere Bewertungskriterien des Datenmaterials sind die Dimension sowie der Umfang des zugrunde liegenden Datensatzes. Die Dimension erfasst die Anzahl der beobachteten Variablen; der Umfang beschreibt die Stichprobengröße.

Zur Überprüfung, ob das gewählte statistische Verfahren für den vorliegenden Datensatz geeignet ist, wird zusätzlich eine detaillierte Charakterisierung der Datenqualität angefertigt. Dazu wird der Datensatz bezüglich fehlender Werte, möglicher Messfehler, etc. begutachtet.

3 Beispiele

3.1 Klassifikation einiger Methoden

Im Folgenden werden verschiedene statistische Methoden kurz vorgestellt und anschließend gemäß der beschriebenen Kriterien Information, Variablen-Symmetrie und Daten-Input klassifiziert:

1) Agglomerative hierarchische Clusteranalyse mit Euklidischer Distanz und Complete-Linkage Verfahren
2) Agglomerative hierarchische Clusteranalyse mit M-Koeffizient und Complete-Linkage Verfahren
3) Multiple lineare Regression mittels KQ-Schätzung
4) Multivariate lineare Regression mittels KQ-Schätzung

Bei agglomerativen hierarchischen Clusteranalysen werden Beobachtungen innerhalb eines Datensatzes nach und nach zu Clustern zusammengefasst. Ziel ist es dabei, dass sich Beobachtungen innerhalb eines Clusters möglichst ähnlich sind und sich die Cluster voneinander deutlich unterscheiden. Zusammengefasst werden bei Complete-Linkage-Verfahren jeweils diejenigen Cluster, deren Abstand minimal ist. Dabei ist der Abstand zweier Cluster durch den maximalen Abstand zweier ihrer Mitglieder gegeben, welcher durch ein Distanzmaß gemessen wird. Bei Methode 1) ist dies die euklidische Distanz, welche für ordinale und stetige Merkmale geeignet ist und bei 2) der M-Koeffizient für nominalskalierte, binäre Merkmale.

Lineare Regressionsmodelle modellieren einen linearen funktionalen Zusammenhang zwischen Einfluss- und Zielgrößen. Die Existenz eines solchen Zusammenhanges wird vorausgesetzt und soll mit Hilfe des Verfahrens genauer spezifiziert werden. Unbekannte Parameter werden mittels der Kleinst-Quadrate (KQ-) Methode geschätzt. Die multiple lineare Regression ist durch eine Zielgröße und mehrere Einflussgrößen gekennzeichnet, die multivariate lineare Regression durch mehrere gleichberechtigte Zielgrößen und mehrere Einflussgrößen. Zielvariablen sind dabei notwendigerweise stetig, Einflussgrößen können sowohl nominal, ordinal als auch stetig sein. In Tabelle 3 sind die oben aufgeführten Klassifikationen zusammengefasst.

<table>
<tr><td rowspan="2">Klassifikations-sicht</td><td colspan="4">Methode</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td rowspan="2">Information</td><td colspan="2">Strukturerkennung</td><td colspan="2">Strukturmodellierung</td></tr>
<tr><td colspan="2">Zusammenfassung von Objekten zu Ähnlichkeitsklassen</td><td colspan="2">Modellierung funktionaler Zusammenhänge</td></tr>
<tr><td>Variablen-Symmetrie</td><td colspan="2">symmetrische Behandlung mehrerer Variablen</td><td>eine Zielvariable, mehrere erklärende Variablen</td><td>mehrere Zielvariablen, mehrere erklärende Variablen</td></tr>
<tr><td rowspan="2">Daten-Input</td><td rowspan="2">ordinales oder stetiges Skalen-niveau</td><td rowspan="2">nominal, binäres Skalenniveau</td><td>stetige Zielvariable</td><td>stetige Zielvariablen</td></tr>
<tr><td colspan="2">nominal, ordinal oder stetige erklärende Variablen</td></tr>
</table>

Tabelle 3: Beispielhafte Klassifikation statistischer Methoden

3.2 Eine Anwendungssituation

Für ein Simulationsmodell des Frachtumschlages an einem Flughafen werden Eingangsdaten benötigt. Im Rahmen des Vorgehensmodells gemäß Abbildung 2 wurde unter anderem für alle auftretenden Flugzeuge ein Datensatz mit folgenden Variablen ermittelt:

X1: Frachtgewicht in Kilo

X2: Anzahl Passagiere

X3: Anzahl Gepäckstücke

Anhand dieser Variablen sollen die Flugzeuge im Datensatz in Klassen mit ähnlicher Frachtkapazität eingeteilt werden, um im Simulationsmodell nur das Auftreten von Flugzeugen aus den jeweiligen Klassen als Systemlast handhaben zu müssen. Da es sich um stetige (X1) bzw. ordinale (X2,X3) Merkmale handelt, eignet sich Methode 1 der in Abschnitt 3.1 vorgestellten Methoden.

4 Diskussion und Ausblick

Das entwickelte Klassifikationsschemata für statistische Methoden differenziert nach den Aspekten Information, Variablen-Symmetrie und Daten-Input. Diesen Aspekten ist eine Hierarchie inhärent in dem Sinne, dass die Aufgabe und damit die zu gewinnende Information im Vordergrund steht und in Kombination mit der Variablen-Symmetrie zu statistischen Verfahren führt, die sich dann noch in ihren Anforderungen und Möglichkeiten bezüglich des Daten-Inputs unterscheiden. Es ist denkbar, auch die gewünschte Art des Daten-Outputs in die Wahl einer geeigneten statistischen Methode mit einzubeziehen. Dies ist jedoch nachrangig anzusehen, da in der Regel die mit ein und demselben Verfahren ermittelten Informationen in der jeweils gewünschten Form als Ausgangsdaten bereitgestellt werden können.

Ziel der Entwicklung von Klassifikationskriterien ist es unter anderem ähnliche Methoden zu typischen Methodenkategorien zusammen zu fassen. Hier existieren in der Statistik bereits Methodenklassen wie Faktorenanalysemethoden, Regressionsverfahren, etc., die sich entsprechend mit Hilfe der entwickelten Kriterien klassifizieren lassen und im Methodennutzungsmodell (Abbildung 1) auf der Taxonomie-Ebene zu finden sind. Einzelne, detailliert spezifizierte statistische Analysemethoden aus der Methodenebene sind dann diesen Methodenklassen zugeordnet. Aufgrund der hohen und weiter steigenden Anzahl an statistischen Methoden ist eine vollständige Einordnung aller statistischen Verfahren nicht denkbar und würde auch zu einer unnötigen Komplexität führen. Vielmehr muss eine Konzentration auf Methodenklassen und Methoden erfolgen, die typisch für die logistische Anwendung und logistische Standardprozesse sind.

5 Literatur

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (1996): Multivariate Analysemethoden – Eine anwendungsorientierte Einführung. 8. Auflage. Springer, Berlin

Bernhard, J., Fender, T., Jodin, D., Kuhnt, S., Langenbach, M., Wenzel, S. (2005): Information Acquisition for Modelling and Simulation of Logistics Networks, zur Veröffentlichung eingereicht.

Fahrmeir, L. , Künstler, R., Pigeot, I., Tutz, G. (2003): Statistik. 4. Auflage. Springer, Berlin.

VDI 3633 Blatt 1, Entwurf (2000): Simulation von Logistik-, Materialfluß- und Produktionssystemen – Grundlagen. VDI-Handbuch Materialfluss und Fördertechnik, Band 8. Verein Deutscher Ingenieure, Düsseldorf, Beuth Verlag, Berlin.

Sonderforschungsbereich 559

Bisher erschienene Technical Reports

03020 Michael Kaczmarek, Marcus Völker: Entwicklung von Simulationsmodellen für die Analyse von Supply Chain-Strategien und -Strukturen im ProC/B-Paradigma

03021 Michael Kaczmarek: Beschreibung ausgewählter Strategien zur Steuerung der Austauschprozesse in der Supply Chain

03022 Michael Kaczmarek: Organisation der Planung und Steuerung in Supply Chains

03024 Anne Schulze im Hove, Frank Stüllenberg, Stefan Weidt: Inhaltliche Ausgestaltung der Netzwerk-Balanced-Scorecard für Beschaffungsketten

03029 Hilmar Heinrichmeyer, Andreas Reinholz: Entwicklung eines Bewertungsmodells für die Depotstandortoptimierung bei Servicenetzen

03032 Marco Motta, Iwo Riha, Stefan Weidt: Simulation eines Regionallagerkonzeptes

03034 Frank Laakmann, Iwo Riha, Niklas Stracke, Stefan Weidt: Workbenchgestützte Konstruktion von Beschaffungsketten

03035 Iwo Riha, Stefan Weidt: Entwicklung einer Bewertungssystematik für Beschaffungsketten

04001 André Alberti, Bernd Hellingrath, Stefan Weidt, Markus Witthaut: Ergebnisse und Schlussforderungen der Simulationsexperimente im Szenario Automobilindustrie

04002 Kay Hömberg, Dirk Jodin, Maren Leppin: Methoden der Informations- und Datenerhebung

04003 Carsten Tepper: Prozessablauf-Visualisierung von ProC/B-Modellen

05001 Jochen Bernhard, Miroslav Dragan, Sigrid Wenzel: Evaluation und Erweiterung der Kriterien zur Klassifizierung von Visualisierungsverfahren für GNL

05002 Entwicklung eines Anaylserahmens für die Untersuchung organisatorischer Aspekte in der Supply Chain

5003 Einsatz der Response Surface Methode zur Optimierung komplexer Simulationsmodelle

05004 Automatisierte Methoden und Systeme der Datenerhebung

05005 Kriterien für die Kategorisierung statistischer Methoden im Rahmen eines Methodennutzungsmodells zur Informationsgewinnung in GNL

Alle Technical Reports können im Internet unter
http://www.sfb559.uni-dortmund.de/
abgerufen werden. Für eine Druckversion wenden Sie
sich bitte an die SFB-Geschäftsstelle
e-mail: grosseca@iml.fhg.de