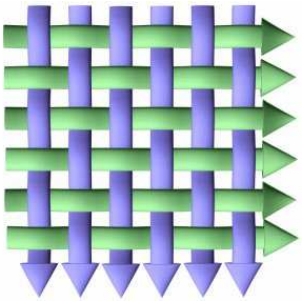
23. November 2003

Sonderforschungsbereich 559
Modellierung großer Netze in der Logistik



Technical Report 03006

ISSN 1612-1376

Statistisches Datenmanagement zur Informationsgewinnung in GNL

Sonja Kuhnt

Lehrstuhl Mathematische Statistik und
industrielle Anwendungen
Vogelpothsweg 87

44221 Dortmund

Thomas Fender

Lehrstuhl Mathematische Statistik und
industrielle Anwendungen
Vogelpothsweg 87

44221 Dortmund

Dortmund, den 23. November 2003

Inhalt

1 Einleitung

Die Informationsbeschaffung und die damit einhergehenden Beschaffung einer Datenbasis haben einen hohen Anteil am Gesamtaufwand bei der Modellbildung in GNL. Das im Teilprojekt M9 entwickelte Vorgehensmodell ist eine ganzheitliche, prozessorientierte Methodik zur kontextbezogenen, d.h. aufgaben- und zielorientierten, Informationsgewinnung unter Einbeziehung der verschiedenen Methodenbereiche Datenerhebung, Statistik und Visualisierung. Nur die problembezogene Auswahl und die koordinierte Anwendung von Verfahren aus diesen Bereichen innerhalb der einzelnen Prozessschritte erlauben eine effiziente Informationsgewinnung und die Bereitstellung von Eingangsdaten für die Modellbildung. Dabei ist sowohl die hintereinander geschaltete Verwendung, aber auch eine integrative Nutzung der einzelnen Verfahren notwendig (vgl.WBE03).

Das Informationsmanagement, das durch das Vorgehensmodell beschrieben wird, grenzt sich deutlich von explorativen Methodiken wie Data Mining ab. Denn Informationsgewinnung ist die Extraktion gesuchter, problembezogener Information aus den Daten, wogegen das Data Mining zur experimentellen Suche nach impliziten noch unbekannten Strukturen eingesetzt werden kann. Das bedeutet, dass zum Informationsmanagement die Verfahren des Data Mining zur Suche nach benötigten Informationen verwendet werden können, der Prozess damit aber nicht abgeschlossen sein kann. Nur eine Validierung und Bewertung der Resultate durch statistische Datenanalyse und eine statistische Qualitäts- und Plausibilitätsprüfung der Daten stellt sicher, dass die so extrahierten Eingangsdaten für eine effiziente und valide Modellierung von GNL nutzbar sind.

Die Verfahren aus dem statistischen Datenmanagement kommen im Prozessverlauf insbesondere innerhalb der Entscheidungsfindung zur Bestimmung des relevanten Informationsbedarfs und bei der Analyse der erhobenen Datenbasis zur Extrahierung der Eingangsdaten zum Einsatz. Die Besonderheit großer Netze der Logistik im Bezug auf die Informationsgewinnung ist die Heterogenität und Komplexität der akquirierbaren Datensätze aus verschiedenen Quellen. Diese Daten sind oft redundant, unplausibel, fehlerbehaftet und lückenhaft erhoben. Deshalb müssen für die statistischen Analysen Methoden zur Komprimierung von Daten und zur Extraktion von Zusammenhängen gewählt werden. Gleichzeitig müssen die angewendeten Algorithmen dieser Verfahren effizient sein, um auf die großen, hochdimensionalen und komplex strukturierten Datensätze angewendet werden zu können.

2 Informationsgewinnung in GNL

Im Zuge der neuen Informationstechnologien ist es heutzutage möglich bei relativ geringen Kosten große Mengen von Daten zu sammeln, zu sichern, zu übermitteln aber auch zu kombinieren oder anderweitig weiterzuverarbeiten. Somit stehen sehr oft große Datenbanken oder sogar Datenbankensysteme zur Gewinnung von Informationen zur Verfügung. Allerdings stellt sich sehr häufig heraus, dass das Verwerten der Information, die in den Datenbanken enthalten ist, meist sehr schwierig ist. Das vollständige Ausschöpfen der Datenbanken in Hinsicht auf den darin enthaltenden Informationsgehalt ist nahezu unmöglich.

Obwohl der Anwender sehr oft ein vages Verständnis von den vorliegenden Daten hat und meist auch Prämissen und Hypothesen über die vorliegenden Informationen durch Expertenwissen formulieren kann, weiß er in der Regel nicht, ob die Informationen in den Daten diese Annahmen auch wirklich unterstützen, und ob nicht noch weitere interessante und vor allem relevante, noch unentdeckte Information in den Daten vorhanden ist.

2.1 Data Mining

Eine mögliche Vorgehensweise zur Gewinnung von Information aus Daten ist Data Mining, mit dessen Hilfe nach Strukturen gesucht wird, so dass Muster, Regularitäten und Zusammenhänge in den Daten erkannt werden können. Die Philosophie bei der Anwendung von Verfahren aus dem Bereich des Data Mining ist, dass ohne Vorgaben experimentell mit Analyseverfahren nach diesen unbekannten Strukturen gesucht wird. Dabei sind diese Verfahren häufig aus dem statistischen Methodenspektrum übernommen oder mit geringen Anpassungen an die geänderten Anforderungen entliehen. Daneben werden aber auch Algorithmen aus den Bereichen der künstlichen Intelligenz und des maschinellen Lernen verwendet. Für eine gute Übersicht über die Aufgaben und die Methoden des Data Mining sei auf [FPS+96] und [FGW01] verwiesen.

Data Mining wird über eine Reihe von Aufgabenstellungen definiert, die mit vorhandenen Methoden gelöst werden sollen, wie Segmentierung, Klassifikation, Prognose, Abhängigkeitsanalysen und Trends (zeitliche Veränderungen). Aus den experimentell erhobenen Ergebnissen werden die gezogenen Schlüsse zur Generierung von Hypothesen (Parameter, Regeln, Modelle) genutzt. Durch die experimentelle Untersuchung der Daten sind die Resultate explorativ, die Überprüfung der Signifikanz der Aussagen ist beim Data Mining auf eine Überprüfung der Hypothesen an den Originaldaten mit Modellvalidierungsverfahren wie z.B. Cross-Validation beschränkt. Theoretische, generell gültige Aussagen über die Güte von Verfahren existieren in der Regel nicht. Dies liegt in der praktischen Ausrichtung des Data Mining begründet, denn oft existieren für die angewandten Algorithmen keine Aussagen zur Konsistenz bzw. zur Anpassungsgüte, oder die Annahmen, die für die Anwendung verschiedener Verfahren gemacht worden sind, können in den Daten selbst nicht beobachtet werden.

Diese Ausrichtung auf die experimentelle Suche nach Strukturen schränkt die Anwendung der Methoden aus dem Bereich Data Mining auf sekundär erhobene Daten ein, da bei Primärdaten davon ausgegangen werden kann, dass diese Daten mit dem Vorsatz erhoben worden sind, mit dem so zur Verfügung stehenden Informationsgehalt die originäre Fragestellung zu beantworten oder den originären Informationsbedarf abzudecken.

2.2 Informationsmanagement

Im Gegensatz zum Data Mining ist Informationsmanagement eine Methodik zur Extraktion der notwendigen Informationen und den damit assoziierten Daten, um den Informationsbedarf für eine konkrete Aufgabenstellung abzudecken (vgl. Tabelle 1). Anhand einer vorgegebenen Aufgabenstellung wird die relevante Information durch eine zielorientierte Informationserhebung gewonnen. Dabei sind neben der Analyse des Informationsbedarfs auch die Auswahl und die Bewertung der Informationsquellen notwendig. Bei dem Abgleich der vorhandenen mit der gemäß der Aufgabenstellung notwendigen Information kann so festgelegt werden, welche zusätzlichen Informationsquellen einerseits notwendig sind und andererseits nutzbar gemacht werden können.

Informationsmanagement	Data Mining
Ziel: Extraktion gesuchter Informationen aus Daten	**Ziel:** Suche nach impliziten, noch unbekannten Strukturen in Daten
primär und sekundär erhobene Daten	sekundär erhobene Daten
aufgabenorientierte und zielführende Extraktion von Strukturen in den Daten	experimentelle Suche in den Daten unter geringen Annahmen
Datenverdichtung (Kondensation, Komplexitätsreduktion)	Aufdeckung von Kausalitäten und Strukturen
anwender- und problemorientierte Aufbereitung und Visualisierung der Daten	Hypothesengenerierung
Validierung der Daten (z. B. durch Ausreißererkennung)	
Behandlung fehlender Werte	

Tabelle 1: Vergleich der Ziele und der Vorgehensweise bei Informationsmanagement und Data Mining

Neben den bereits vorhandenen oder einfach zu beschaffenden Sekundärdaten aus internen und externen Quellen ist auch die kontextbezogene, zielgerichtete Erhebung von Primärdaten sinnvoll und oft sogar notwendig. Mit Hilfe digitaler und auch manueller Erhebungsmethoden können an den unterschiedlichen Informationsquellen Daten extrahiert und in Datenbanken erfasst werden. Die dann digitalisierten und strukturierten Daten liegen als Rohdaten in einer Matrix mit Variablen und Beobachtungen vor.

Für die weitere Nutzung der eigentlich benötigten Informationen, zum Beispiel als Eingangsdaten zur Modellierung großer Netze in der Logistik (GNL), ist die Aufbereitung dieser Daten mit statistischen Methoden notwendig (vgl. Kapitel 3). Ziel ist eine Verdichtung der Daten, so dass ein möglichst einfacher Informationsraum aufgespannt werden kann, der aber die notwendige Information enthält. Dazu müssen im Rahmen eines umfangreichen Datenmanagement statistische Verfahren nebeneinander, aber auch integrativ genutzt werden. Zusätzlich kommen auch andere Verfahren beispielsweise aus den Bereichen des Data Minings oder der Visualisierung (vgl. [WSO02]) zur Anwendung.

Im Weiteren ist eine Qualitäts- und Plausibilitätsprüfung unabdingbar, um vor allem die Konsistenz der Daten untereinander zu gewährleisten, aber auch um Ausreißer zu erkennen, oder um eventuell fehlende Werte von Variablen zu bewerten und zu behandeln. Erst danach kann die eigentliche statistische Datenauswertung unter Beachtung der komplexen Strukturen und der Hochdimensionalität der Datensätze beginnen. Dazu gehört insbesondere die Kondensation und die Komplexitätsreduktion der Daten. Weitere wichtige Aufgaben des Datenmanagements sind die variablenspezifische Klassifikation der Daten und die Mustererkennung in den Daten. Für diese Problemstellungen ist die Erstellung eines

Modells, das den Daten entspricht, notwendig. Dazu gehören etwaige Annahmen an eine Verteilung in den Daten und statistische Inferenz mit Parameterschätzung, Hypothesentests, Konfidenzverfahren sowie Prognosen.

Zusätzlich kann die Entdeckung impliziter, unbekannter Muster in den Daten mit explorativen, statistischen Verfahren eine Aufgabe des Informationsmanagement sein. In diesem Fall können auch Verfahren aus dem Bereich des Data Mining sinnvoll sein. Die Abgrenzung des Informationsmanagements zum Data Mining ergibt sich aber vor allem aus der Aufgabenstellung der Informationsgewinnung. Die ziel- und aufgabenorientierte Extraktion benötigter Information und die transparente Aufbereitung dieser Information und der zugehörigen Daten in ausreichender Qualität und Granularität, damit ein Anwender die Information aufwandsreduziert weiterverarbeiten kann.

2.3 Statistisches Datenmanagement

Die Informationsgewinnung nimmt einen hohen Stellenwert bei der Modellierung von Logistiksystemen ein, denn die Bereitstellung der relevanten Information in einer guten Qualität und in der richtigen Granularität ist Garant dafür, dass die Resultate der Modellierung überhaupt nutzbar sind. Vor diesem Hintergrund ist eine validierbare statistische Datenanalyse im Zuge der Beschaffung von Information von großer Wichtigkeit, damit die Güte der Information, die in die Modellierung eingeht, nachprüfbar ist. Von besonderer Bedeutung ist hierbei einerseits die Beachtung der Voraussetzungen an ein statistisches Verfahren, damit dieses überhaupt validierbare Ergebnisse liefert. Andererseits müssen die relevanten Zusammenhänge in den Daten durch die Ergebnisse der statistischen Analyse aufgedeckt werden, d.h. es darf kein systembedingter, sondern nur ein zufälliger Fehler die Störgrößen innerhalb der Modellierung der Zusammenhänge ausmachen.

Bei der Modellierung großer Netze der Logistik, insbesondere für Simulationsmodelle und Modellierungsverfahren zur Optimierung der Netze, wird ein Großteil der benötigten Information über die Eingangsdaten in die jeweilige Modellierung eingebracht. Diese Daten können Systemlasten, Organisationsdaten oder auch technische Daten sein. Die Darstellungsform muss dem jeweiligen Modellierungsverfahren angepasst werden, das bedeutet in der Regel eine Datenverdichtung auf Kennzahlen (Lage- und Streuungsmaße einer Verteilung), univariaten und multivariaten Verteilungsannahmen oder auf die Klassifikation der Daten (Clusterung). Da die relevante Information für eine sachgerechte Durchführung der logistischen Modellierungsverfahren trotzdem in diesen Eingangsdaten enthalten sein muss, ergibt sich als übergeordnete Aufgabe für das statistische Datenmanagement die Reduktion sowohl des Datenumfangs als auch der Komplexität innerhalb der Datenstrukturen.

Ein wichtiger Betrachtungsgegenstand bei den Modellierungsaufgaben in GNL ist das Verhalten der modellierten Netze bei Veränderungen in den Systemen über die Zeit hinweg oder die Anwendung des Modells auf ein neues, noch aufzubauendes, aber vergleichbares System. Dazu sind Eingangsdaten notwendig, die den so neu entstandenen Situationen in den Netzen entsprechen, das bedeutet, dass eine adäquate Anpassung der Daten notwendig ist. Das statistische Datenmanagement stellt hierfür sowohl Verfahren zur zeitlichen Prognose als auch zur systembezogenen Prognose zur Verfügung.

Ein anderer Blickwinkel auf die Anforderungen des Datenmanagements ergibt sich bei der Betrachtung der zur Verfügung stehenden Informationsquellen und den dort erhebbaren Informationen in großen Netzen. Selbst mit modernen Datenerfassungsmethoden können häufig Fehler in den Daten, unplausible Daten oder fehlende Daten nicht ausgeschlossen werden. Daraus ergibt sich die Problematik der oft nicht unmittelbar ersichtlichen Fehlerquellen in den großen und komplexen Datensätzen, die dann zu Artefakten bei der Auswertung führen. Ausreißeridentifikation und -bewertung, sowie die Anwendung robuster statistischer Verfahren sind hier geeignet trotzdem valide Resultate zu liefern.

In Logistiksystemen tritt das Problem auf, dass zahlreiche Informationsquellen nicht genutzt werden können, beziehungsweise ist die Nutzung teilweise zu kostenintensiv, oder die Information liegt überhaupt nicht vor, so dass auf Informationen und somit auch auf Daten aus sekundären Quellen zurückgegriffen werden muss. Bei so erhobenen Daten ist der Informationshintergrund nicht direkt mit dem des zu betrachtenden Systems vergleichbar. Die Aufgabe bei der Verwendung solcher Daten kann die Erzeugung neuer Daten unter Verwendung der Annahmen aus dem zu betrachtenden System sein. Mit Methoden des statistischen Datenmanagement kann die Verteilung der sekundär erhobenen Daten mit den vorgegebenen Annahmen verglichen und gegebenenfalls angepasst werden, so dass Daten generiert werden können, die den Annahmen in dem zu betrachtenden Logistiksystem entsprechen.

3 Statistische Verfahren zum Datenmanagement

Im diesem Abschnitt werden statistische Verfahren gemäß ihrer Aufgabenstellung innerhalb der Informationsgewinnung klassifiziert. Speziell bei der Verdichtung der erfassten Daten sind vor allem ziel- und anwenderorientierte statistische Verfahren notwendig, die ausreichend valide Eingangsdaten liefern, deren Informationsgehalt bzw. -qualität der Modellierungsaufgabe angemessen ist, und deren Darstellungsform einen Einsatz des Modellierungsverfahrens erlaubt. Neben gängigen multivariaten Verfahren (vgl. [FHT96]) wie etwa der multivariaten Regressionsanalyse, der Faktoranalyse und der Clusteranalyse, bieten sich auch neuere statistische Entwicklungen auf den Gebieten der Dimensionsreduktion, der Ausreißererkennung und der robusten Statistik für das Datenmanagement bei Daten für GNL an.

3.1 Reduktion

Typischer Weise liegen Datensätze vor, in denen eine hohe Anzahl an Objekten durch viele Variablen charakterisiert werden. Um Strukturen aus derartigen Datensätzen heraus-zukristallisieren bietet die Statistik eine Reihe von dimensions- und datenreduzierenden Verfahren.

3.1.1 Dimensionsreduktion

Die Reduzierung der Dimension eines Variablenvektors kann sowohl durch das Entfernen einzelner Variablen aus dem Variablenvektor als auch durch die Zusammenfassung mehrerer Variablen zu einer geringeren Anzahl an Variablen geschehen. Diese Reduzierung oder Zurückführung einer größeren Menge von Variablen auf eine möglichst kleine Menge von Variablen geschieht jeweils auf Basis des gegebenen Datensatzes mit dem Ziel möglichst viel der vorhandenen Information beizubehalten.

- Variablenselektion

  Regressionsanalytische Fragestellungen gehen davon aus, dass eine interessierende Zielvariable funktional von einer oder mehreren erklärenden Variablen abhängt. Um aus einer Menge von potentiell erklärenden Variablen die wichtigsten auszuwählen werden Methoden der Variablenselektion eingesetzt. Besser zu realisieren als die vollständige Untersuchung aller möglichen Teilmengen der potentiellen Einflussvariablen sind dabei häufig schrittweise Verfahren. Bei der Rückwärtselimination (Backward Selection) werden ausgehend von der Menge aller potentiell erklärenden Variablen schrittweise Variablen eliminiert, die nicht wesentlich zur Erhöhung der Information über die Zielvariable beitragen, wobei „wesentlich“ anhand statistischer Tests beurteilt wird. Bei der Vorwärtsselektion wird umgekehrt vorgegangen und ausgehend von keiner Einflussgröße schrittweise diejenige mit dem größten Einfluss auf die Zielvariable hinzugenommen. Beide Verfahren bedürfen eines Stoppkriteriums; Ziel ist es mit möglichst wenigen Variablen möglichst viel Information über die Zielvariable zu erhalten. ([FHT96], Kap. 4). Vorteile der schrittweisen Selektion von Variablen zur Bildung eines Regressionsmodells sind neben der guten Interpretierbarkeit der ausgewählten Einflussvariablen, diese werden selber nicht verändert, die Verfügbarkeit von Algorithmen und Programmen. Es ist jedoch nicht garantiert, dass das nach einem gegebenen Kriterium „beste“ Modell gefunden wird, da nicht alle möglichen Modelle untersucht werden.

- Hauptkomponentenanalyse (PCA) / Faktorenanalyse

  Faktoranalytische Verfahren ([FHT96], Kap. 11), zu denen auch die Hauptkomponentenanalyse zählt, zielen darauf ab, eine größere Menge beobachtbarer abhängiger Variablen auf eine möglichst kleine Menge zu Grunde liegender, unabhängiger Variablen, Faktoren genannt, zurückzuführen. Diese Variablen sind als Ergebnis der faktoranalytischen Methode in der Regel hypothetisch und nicht empirisch erfassbar. Für das Datenmanagement in GNL ist der Schwerpunkt auf einer explorativen Verwendung von faktoranalytischen Methoden zu legen, da sich eher Fragestellungen ergeben, bei denen ausgehend von empirischen Daten wenige gemeinsame Faktoren gefunden werden sollen als das bereits Strukturen oder Wissen über die hypothetischen Variablen vorliegt. Eine besondere Rolle spielt hier die so genannte Hauptkomponentenmethode, da sie als rein datenmanipulierendes Verfahren angesehen werden kann, dass auf die Formulierung eines expliziten statistischen Modells verzichtet. Diese Methoden erlauben eine geometrische Interpretation der gebildeten Faktoren als orthogonale Richtungen, die sukzessive ein Maximum der Stichprobenvarianz erklären. Nachteile dieser Methoden sind die anspruchsvolle und nicht eindeutige Wahl der Anzahl der ausgewählten Faktoren und die Interpretation der Faktoren, welche lediglich daraus abgeleitet werden kann, wie diese aus den ursprünglichen Variablen gebildet werden. Faktoranalytische Methoden sind in allen statistischen Programmpaketen implementiert.

- Sliced Inverse Regression (SIR)

  Die relative neue Methode der Sliced Inverse Regression [LI91], basiert auf der Tatsache, das der durch die Einflussgrößen in regressionsanalytischen Problemen aufgespannte hochdimensionale Raum aufgrund von Abhängigkeiten zwischen den Einflussgrößen oft auf einen niederdimensionalen Raum zurückgeführt werden kann. Innerhalb eines mehrschrittigen Verfahrens wird diese Zurückführung mit Hilfe der Hauptkomponentenanalyse durchgeführt, so dass hier eine Verknüpfung von Regression und Hauptkomponentenanalyse stattfindet. Die SIR-Methode selber und robustere Varianten [GHB01] zählen noch nicht zu bekannten Standardverfahren und sind daher auch in gängigen Software-Paketen bisher nicht implementiert.

3.1.2 Datenreduktion / Datenselektion

Die Analyse großer Datensätze und ihre Nutzung zur Bestimmung von Eingangsdaten für Modellierungen in GNL wird durch die Unübersichtlichkeit, die aus der reinen Masse der Daten resultiert, erschwert. Aus diesem Grund ist neben der Dimensionsreduktion eine Reduktion der Daten auch dahingehend sinnvoll, dass Daten, die nicht zur Information beitragen, nicht berücksichtigt werden. Dies können Ausreißer sein, welche nach ihrer Identifikation entfernt werden, oder Klassen von Beobachtungen, die aufgrund ihrer Eigenschaften nicht Gegenstand der Aufgabenstellung sind. Andererseits können Beobachtungen mit ähnlichen Merkmalsausprägungen zu Gruppen zusammengefasst analysiert werden.

- Klassifikation

  Das Ziel der Klassifikation ist die Einteilung von Objekten in Gruppen, wobei sich die Objekte einer Klasse möglichst ähnlich sein sollen und zwischen den Klassen möglichst unterschiedlich. Die Entscheidung der Gruppenzugehörigkeit ergibt sich aus der Analyse der zu den Objekten erhobenen Merkmalen, d.h. aus den Daten. Die Zuordnung der Objekte erfolgt dabei durch den Vergleich der Abstände zwischen den Objekten, wobei ein geeignetes, der Struktur der Daten angemessenes Abstandsmaß gewählt werden muss. Außerdem muss als weitere Annahme die Anzahl der Klassen vorgegeben werden. Mit den so entwickelten Regeln für die Zugehörigkeit eines

Objektes zu einer Klasse können auch neu hinzukommende Objekte einer bereits bestehenden Klasse zugeordnet werden.

Durch eine Klassifizierung können die für die Modellierung in GNL interessanten Objekt- oder Datengruppen identifiziert und selektiv genutzt werden. Außerdem können die Daten durch die Schätzung der Verteilung oder zumindest der Lage und Streuungsmaße in den einzelnen Gruppen zusätzlich verdichtet werden. Dies führt zu klassierten Eingangsdaten, bei denen die Informationen der Objekte aus einer Gruppe reduziert zusammengefasst sind. Neben dem Vorteil dieser hohen Datenreduktion und -selektion gibt es wegen der hohen Anzahl von Klassifikationsverfahren für viele Problemstellungen und Datensituationen anwenderorientierte, standardisierte Algorithmen.

Weitere gängige Klassifikationsverfahren sind Diskriminanzanalyseverfahren (vgl. [MCL92]), bei denen auch Aussagen zur Güte gemacht werden, und die eher explorativen Verfahren Clusteranalyse (vgl. [ELL01], [KAU90]), sowie Support Vector Machine, k-nearest-Neighbor-Algorithmen und Classification-Tree-Algorithmen (vgl. [HTF01]). Dabei ist ein Classification-Tree hierarchisch aufgebaut, für die Clusteranalyse gibt es sowohl partitionierende als auch hierarchische Verfahren, wogegen die restlichen Verfahren partitionierende Verfahren sind.

- Ausreißeridentifikation

  In jedem Datensatz können einzelne Beobachtungen vorhanden sein, die stark von einem Muster, einer Struktur abweichen, die die restlichen Daten aufweisen. Derartige Beobachtungen werden in der Regel als Ausreißer bezeichnet. Eine formale, allgemein anerkannte Definition des Ausreißerbegriffes hat sich jedoch noch nicht durchgesetzt. Das α-Ausreißerkonzept (vgl. [DGA93], [GKP03]) liefert eine Ausreißerdefinition, bei der Beobachtungen als Ausreißer bezeichnet werden, wenn sie unter einem betrachteten Modell sehr unwahrscheinlich sind. Um Ausreißer in einem Datensatz zu erkennen, existieren eine Vielzahl von Verfahren, die sich je nach Datensituation und verwendetem Ausreißerbegriff unterscheiden. Generell können sie nach einschrittigen und mehrschrittigen Verfahren unterschieden werden. Bei einschrittige oder auch simultanen Verfahren werden in einem Schritt alle Beobachtungen gleichzeitig danach beurteilt, ob sie als Ausreißer anzusehen sind oder nicht. Mehrschrittige Verfahren hingegen beurteilen in jedem Schritt einzelne Beobachtungen, dabei beginnen outward Prozeduren mit einem reduzierten Datensatz, der keine verdächtigen Beobachtungen enthält. Sukzessive werden dann die am wenigsten auffälligen Beobachtungen hinzugefügt, solange sie zu dem Datensatz passen. Inward Prozeduren dagegen setzen beim vollständigen Datensatz an und entfernen schrittweise Beobachtungen, die als Ausreißer beurteilt werden. Die Entscheidung, welche Beobachtung als wie „auffällig“ zu beurteilen ist und wann die Prozeduren stoppen beruht jeweils auf vorher festgelegten formalen Kriterien. Die Erkennung von Ausreißern dient nicht unbedingt dazu, den Datensatz von diesen zu bereinigen. Vielmehr sind oft die Ausreißer selber von Interesse, zum Beispiel um in GNL auch worst-case Fälle zu bedenken. Obwohl eine Identifizierung von Ausreißern in jedem Datensatz stattfinden sollte, gibt es bisher kaum Standardverfahren. Auch sind existierende Verfahren bisher so gut wie nicht in statistischen Programmpaketen implementiert.

3.2 Prognose

Eine häufige Fragestellung bei der Modellierung von GNL ist das Verhalten der Systeme nicht nur bei Veränderungen der Systemkonfiguration, sondern auch bei einer Übertragung auf neue, noch nicht existierende Systeme oder in Hinsicht auf das zukünftige Verhalten des bekannten Systems. Für solche logistischen Prognosemodelle müssen aus den erhobenen Daten mit geeigneten statistischen Modellen und Verfahren adäquate Eingangsdaten geschätzt bzw. hochgerechnet werden.

3.2.1 Zeitliche Prognose

In großen Netzen der Logistik gerade bei der Modellierung von Systemlasten werden Daten in der Regel in einem zeitlichen Kontext erhoben. Eine statistische Datenanalyse, die auf der Unabhängigkeit der untersuchten Variablen basiert, wird solchen zeitabhängigen Datenstrukturen nicht gerecht. Es bleibt unbeachtet, dass die Daten zu unterschiedlichen Zeitpunkten erhoben worden sind und womöglich von vorhergehenden Werten abhängen. Die nicht zutreffende Annahme der Unabhängigkeit kann sich negativ auf die Qualität der statistischen Modellierung auswirken. Einen Ausweg zur datengerechten, statistischen Modellierung bieten die Verfahren der Zeitreihenanalyse. Ein zusätzlicher Vorteil der Nutzung von Zeitreihenverfahren ist die Möglichkeit durch geeignete Fortschreibung der Reihe direkt eine Prognoseschätzung zu erhalten.

Die Modellierung zeitlicher Prozesse in den Daten kann sowohl mit univariater Zeitreihenanalyse (vgl. [BDA96]) als auch mit multivariaten Analysemethoden (vgl. [REI97]) durchgeführt werden. Die klassischen Zeitreihenmodelle ergeben sich aus autoregressiven Anteilen (AR), bei denen Werte mit zeitlicher Verzögerung in das Modell eingehen, und Moving Average Anteilen (MA), die die Innovationen zu verschiedenen Zeitpunkten darstellen. Für Modelle dieser Art (z.B. MA-, AR-, ARMA-, ARIMA-Modelle) und Modelle, die quadratische Fehlerstrukturen berücksichtigen (z.B. ARCH-, GARCH-Modelle) gibt es sowohl Verfahren zur Modellidentifikation als auch für die Parameterschätzung in dem jeweils gewählten Modell. Da deterministische Strukturen, wie Trend, saisonale Effekte, Ausreißer, Strukturbrüche durch geeignete Filter-Verfahren erkannt werden können, ist die Berücksichtigung dieser Einflüsse zusätzlich zu der zeitlichen Modellierung möglich.

Die Fortschreibung der Daten in die Zukunft mit einem so entwickelten Modell ergibt eine Prognoseschätzung, bei der die Schwankungsbereiche aufgrund der geschätzten Prognosevarianz mit angegeben werden können (Konfidenzbänder, -intervalle). Dadurch ist eine Güteberechnung für die Prognose möglich. Bei einer hohen Güte der Modellierung oder einem hohen Anteil des deterministischen Einflusses in dem Modell ist die Präzision der Prognose gut.

Die Wahl eines geeigneten Modells kann je nach Komplexität der Daten aufwendig sein. Oft ist es notwendig Modelle, die nicht zu den Standardzeitreihenmodellen gehören, an die Daten anzupassen. Dies ist durch den Anwender aus dem logistischen Kontext ohne Vorkenntnisse nur schwer zu leisten, so dass statistischer Support notwendig bleibt. Ein weiteres Problem ist die qualitative Bewertung der erzeugten Modelle, da die statistischen Verfahren zur Bestimmung der Modellgüte nur bei einfachen Zeitreihenmodellen anwendbar sind. Weitere Güteberechnungen sind nur durch statistische Simulationen oder durch Cross-Validation möglich.

3.2.2 Systembezogene Prognose

Bei der Analyse logistischer Netze stellt sich häufig die Aufgabe, sehr kostenintensive oder aus anderen Gründen schwer erhebbare Informationen durch andere Informationen aus derselben oder aus anderen Quelle zu ersetzen. Jede vorhandene Information soll zur Lösung der Problemstellung genutzt werden. Die Anpassung dieser Information an die eigentlich benötigte Information kann durch Hochrechnungen erreicht werden.

Die systembezogene Prognose ist die Hochrechnung auf die gesuchten Kennzahlen oder Parameter aus dem zu betrachtenden logistischen System unter Ausnutzung bereits erhobener Daten eines ähnlich aufgebauten anderen Systems. Dabei werden die bekannten Ähnlichkeiten oder Zusammenhänge zwischen den Systemen als a-priori-Information genutzt, um Aussagen über die eigentliche Fragestellung machen zu können. Ebenso ist die Hochrechnung eines Parameters eines noch nicht erhobenen Merkmals möglich, wenn bereits erhobene Merkmale aus dem gleichen System existieren. Dann reicht eine kleine, gebundene, zusätzliche Stichprobe des zu betrachtenden Merkmals und der anderen Merkmale aus, um den gesuchten Parameter zu schätzen.

Dieses Vorgehen wird als gebundene Hochrechnung bezeichnet. Einfache Verfahren sind die Differenzschätzung für additive Zusammenhänge zwischen den betrachteten Daten und die Verhältnisschätzung für relativ von einander abhängige Daten. Dabei muss für beide Verfahren ein hoher Grad von Abhängigkeit (hohe Korrelation) der Merkmale angenommen werden (vgl. [LOH99]) Außerdem kann die Regressionsschätzung Anwendung finden. Dann können die gesuchten Merkmalswerte als abhängige Variablen aus den bekannten Einflussfaktoren (bekannte Merkmale) durch ein lineares, nichtlineares oder generalisiertes lineares Modell geschätzt werden. Auch hier ist eine Abhängigkeit zwischen den betrachteten Merkmalen Voraussetzung (vgl. [VDR00]).

Die systembezogene Prognose birgt den Vorteil, dass Analysen an Daten, die nur kostenintensiv oder mit großen technischen Schwierigkeiten erhoben werden können, im Verbund mit anderen bereits bekannten oder einfach zu erhebenden Merkmalen durchführbar sind. Dies gilt für die Anwendung sowohl innerhalb eines logistischen Systems, als auch zwischen mehreren ähnlichen Logistiknetzen. Beachtet werden muss dabei aber grundsätzlich, dass die Voraussetzung der Abhängigkeit zwischen den Merkmalen und damit zwischen den Systemen geben sein muss. Dies sind häufig nur Annahmen, wenn diese Zusammenhänge nicht bereits in Vorarbeiten validiert worden sind.

3.3 Generierung

Bei der Informationsgewinnung in GNL liegen teilweise nur Daten aus sekundären Quellen vor, deren Informationshintergrund verwandt mit dem zu betrachtenden logistischen System und der eigentlichen Problemstellung ist. Die Nutzung dieser Daten ist nur dann sinnvoll, wenn die Verteilungsannahmen mit denen des eigentlichen Systems übereinstimmen.

- Verteilungsanpassung

  Die Annahmen über die theoretische Verteilung der benötigten Daten kann als mathematische Hypothese dargestellt werden, die entweder auf einen oder mehrere Parameter der Verteilung beschränkt ist, oder die die gesamte Verteilung berücksichtigt. Zum Vergleich, ob die vorliegenden Daten mit dieser theoretischen Verteilung vergleichbar sind, können die empirische Verteilung oder die geschätzten Lage- und Streuungsmaße der Daten mit Hilfe von statistischen Anpassungstests mit der Hypothese verglichen werden (vgl. z.B. [SHE00], S. 51ff). Die Daten können oft mit geeigneten Transformationen bezüglich ihrer Lage und der Streuung an die theoretische Verteilung angepasst werden. Es ist aber auch möglich, die theoretischen Verteilungsannahmen (Parameter) mit Hilfe der Parameterschätzung aus den realen Daten zu modifizieren.

  Für die weiteren statistischen Analysen zur Bestimmung der Eingangsdaten können sowohl die transformierten Daten, die empirische Verteilung der Daten, aber auch die theoretische Verteilung herangezogen werden. Außerdem können auch neue Daten in großem Umfang als Zufallszahlen aus den Verteilungen generiert werden.

  Der Vorteil dieser Vorgehensweise ist, dass die Datenbeschaffung über sekundäre Quellen mit vertretbarem Aufwand durchgeführt wird, oder überhaupt erst möglich ist. Allerdings sollte beachtet werden, dass die Bestimmung der Verteilungshypothese nur mit a-priori-Informationen aus dem Logistiksystem sinnvoll ist. Außerdem kann durch die Testentscheidung nur festgestellt werden, dass die Hypothese nicht verworfen werden kann, die Fehlerwahrscheinlichkeit dafür, dass die Hypothese nicht verworfen wird, obwohl sie nicht korrekt ist, kann dagegen nicht kontrolliert werden.

4 Ausblick

Dieser Bericht nennt zahlreiche statistische Verfahren, die sich zur Datenanalyse und -validierung innerhalb der Informationsgewinnung eignen. Diese sind nach Verfahrensgruppen gemäß ihrer Aufgabenstellung und ihrer Funktionalität innerhalb der Informationsgewinnung klassifiziert und bewertet. Dabei stehen die Anwendbarkeit und die Zielorientierung der Verfahren im Vordergrund, um mit möglichst standardisiertem Vorgehen innerhalb des Informationsgewinnungsprozesses eine präzisere und anwenderfreundliche Informationsbeschaffung bei der Modellierung in GNL zu ermöglichen.

Innerhalb der Kooperationen des Teilprojekt M9 mit den Anwendungsprojekten im SFB 559 konnten einige dieser Verfahren bereits experimentell an realen Datensituationen bewertet werden (vgl. [BFE03]). Die Anwendung und Valdierung weiterer Verfahren ist durch die Fülle der unterschiedlichen Aufgabenstellungen und die unterschiedlichen Informationsbestände innerhalb des SFB 559 möglich. Gleichzeitig können durch diese Untersuchungen zusätzlich weitere geeignete statistische Verfahren oder Verfahrensklassen ermittelt werden, die der jeweiligen Modellierungsaufgabe in GNL angemessen sind.

Für die Anwendung auf die in GNL vorliegenden großen und hochdimensionalen Datensätze sind zur Reduktion und Selektion Verfahren zur Dimensionsreduktion und zur Klassifikation, aber auch zur Ausreißeridentifikation notwendig. Die Modellierung der Zukunft eines Systems oder neuer Systeme bedingen die Anwendung von zeitlichen und systembezogenen Prognoseverfahren oder sogar die Generierung neuer Daten oder Verteilungen aus Daten eines anderen Systems, die jeweils eine Fortschreibung bzw. Hochrechnung der Daten erlauben. Die Anwendung von explorativen Verfahren, wie beispielsweise Data Mining Methoden, innerhalb des Informationsmanagements ist möglich und gerade in Bezug auf die Entscheidungsfindung im Prozess der Zieldefinition sehr sinnvoll, doch reichen solche Verfahren in der Regel nicht aus, da Data Mining nur zur experimentellen Suche nach Strukturen in den Daten geeignet ist, und keine statistisch validierten Resultate liefert.

Generell ergibt sich im Kontext der Informationsgewinnung in GNL aufgrund der Größe der Datensätze die Notwendigkeit, dass für die Verfahren zum statistischen Datenmanagement effiziente und schnelle Algorithmen existieren. Außerdem sind sehr häufig robuste statistische Verfahren unabdingbar, da die Daten häufig fehlerbehaftet sind und aus sekundären Informationsquellen stammen.

Literatur

[BDA96] Brockwell P.J.; Davis, R.A.: Introduction to Time Series and Forecasting, Springer Verlag, New York, 1996.

[BFE03] Bernhard, J.; Fender, T.: Experimentelle Anwendung statistischer Verfahren und Visualisierungsmethoden zur Gewinnung ausgesuchter Eingangsdaten im Kontext von GNL. Technical Report – Sonderforschungsbereich 559 „Modellierung großer Netze in der Logistik“ 03014, 2003, ISSN 1612-1376.

[DGA93] Davis, P.L.; Gather U.: The Identification of Multiple Outliers (with Discussion and Reply), Journal of the American Statistical Association, 88, 782-792, 1993.

[ELL01] Everitt, B.S.; Landau, S.; Leese, M.: Cluster Analysis, 4th ed., Arnold, London, 2001.

[FGW01] Fayyad, U.M.; Grinstein G.; Wierse, A.: Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, Burlington, 2001.

[FHT96] Fahrmeir, L.; Hamerle, A.; Tutz, G.: Multivariate statistische Verfahren, de Gruyter Verlag, Berlin, 1996.

[FPS+96] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, Menlo Park, 1996.

[GHB01] Gather, U.; Hilker, T.; Becker, C.: A Robustified Version of Sliced Inverse Regression. In: Fernholz, L.T.; Morgenthaler, S.; Stahel, W. (Hrsg.): Statistics in Genetics and in the Environmental Sciences, Birkhäuser, Basel, 2001, 147-157.

[GKP03] Gather, U.; Kuhnt, S.; Pawlitschko, J.: Concepts of Outlyingness for Various Data Structures. In: Misra, J. C. (Hrsg.): Industrial Mathematics and Statistics. Narosa Publishing House, New Delhi, 2003, S. 545-585.

[HTF01] Hastie, T.; Tibshirani, R.; Friedman, J.: The Elements of Statistical Learning – Data Mining, Inference, and Prediction. Springer, New York, 2001.

[KAU90]: Kaufman, L.; Rousseeuw; P.J.: Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.

[LI91] Li, K.C.: Sliced Inverse Regression for Dimension Reduction, Journal of the American Statistical Association, 86, 316-342, 1991.

[LOH99] Lohr, S.L.: Design and Analysis. Duxbury Press, Pacific Grove, 1999.

[MCL92] McLachlan, G.J.: Discriminant Analysis and Pattern Recognition, Wiley, New York, 1992.

[REI97] Reinsel, G.C.: Elements of Multivariate Time Series Analysis, Springer Verlag, New York, 1997.

[SHE00] Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, 2nd ed., Chapman & Hall, Boca Raton, 2000.

[VDR00] Valliant, R.; Dorman, A.H.: Finite Population Sampling and Inference – A Prediction Approach, Wiley, New York, 2000.

[WBE03] Wenzel, S.; Bernhard, J.: Vorgehensmodell zur Informationsgewinnung für die Modellierung von Logistiksystemen. In: Hohmann, R. (Hrsg.): Simulationstechnik. Tagungsband zum 17. Symposium in Magdeburg, Reihe Frontiers in Simulation, FS 13, SCS-Europe BVBA, Ghent, 2003, S. 379-384.

[WSO02] Wegman, E. J.; Solka, J. L.: On Some Mathematics For Visualizing High Dimensional Data. In: Sankhyā Series A 64, S. 429-452, 2002.