Visuelles Data Mining und Visualisierungsdesign für die Klimaforschung

Dissertation

zur

Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

der Fakultät für Informatik und Elektrotechnik

der Universität Rostock

vorgelegt von

Thomas Nocke,
geb. am 27.09.1975 in Dresden
aus Berlin

Rostock, 6. Juli 2007

Betreuerin/Gutachterin
Prof. Dr. Heidrun Schumann, Universität Rostock, Fakultät für Informatik und Elektrotechnik

Externe Gutachter
Prof. Dr. Helwig Hauser, University of Bergen, Department of Informatics
Prof. Dr. Holger Theisel, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik

Termin der Verteidigung
30.11.2007

Zusammenfassung

Die Kombination von Verfahren der Visualisierung und der automatischen Datenanalyse – das sogenannte visuelle Data Mining – ist ein wichtiges Hilfsmittel bei der Erforschung großer Datenmengen. Dennoch besteht eine Lücke zwischen den in modernen Softwaresystemen verfügbaren Methoden und deren Einsatz in speziellen Anwendungen. Am Beispiel der Analyse von Klimadaten zeigt diese Arbeit auf, wie diese Lücke durch eine enge Verknüpfung von interaktiven Visualisierungstechniken und automatischen Analyseverfahren sowie durch einen hohen Grad an Nutzerunterstützung geschlossen werden kann. Schwerpunkt der Arbeit ist die Anpassung und Weiterentwicklung von Visualisierungstechniken für die heterogenen, multivariaten Datensätze aus der Klimaforschung. Systematisch wird die Spannbreite existierender Techniken auf ihre Einsatzmöglichkeiten in diesem Kontext untersucht und neue Techniken – insbesondere zur Darstellung von räumlichen und zeitlichen Daten sowie zur vergleichenden Visualisierung – entworfen. Weiterhin werden Methoden zur Darstellung von Ergebnissen aus Cluster- und Hauptkomponentenanalysen (weiter-)entwickelt. Des Weiteren wird ein allgemeines Vorgehensmodell zum Einsatz des visuellen Data Mining im gesamten Prozess der Modellbildung, -simulation und -evaluation entworfen und veranschaulicht. Abschließend wird, um den Einsatz der hier vorgestellten Techniken zu erleichtern, ein Mechanismus zur Auswahl und Parametrisierung von Visualisierungstechniken entworfen. Als Basis für diesen Mechanismus werden wichtige Einflussfaktoren (Metadaten, Analyseziele) spezifiziert und Methoden zu deren nutzergestützter Erhebung und Verwaltung vorgestellt.

Abstract

The combination of visualization and automated mining methods – so-called visual data mining – is an important means of analyzing large data sets. However, there is a gap between the available methods in sophisticated software systems and their use in special domains. Based on the example of climate data this thesis shows how to close this gap by tightly coupling interactive visualization and automatic analysis methods as well as by providing a high degree of user interaction. The main focus of this work is on how visualization techniques can be enhanced and adapted to the heterogeneous, multi-variate data sets used in climate research. To this end, the possible fields of application of visualization techniques are systematically investigated and new techniques are proposed, especially for spatial and temporal data analysis and for comparative visualization. Furthermore, innovative techniques displaying results of cluster and principal component analysis are provided. Moreover, a general procedure to apply the proposed visual data mining techniques to the whole process of modeling, simulation and model evaluation is designed and illustrated. Finally, to simplify the application of the proposed techniques, a selection and parameterization mechanism for visualization techniques is designed. The important factors influencing this mechanism (metadata, goals of analysis) are specified, and methods for their collection and management offering a high degree of user support are proposed.

Keywords

Visual Data Mining, Visualization, Climate Research, Visualization Design, Metadata, Goals of Analysis, Cluster Analysis, Principal Component Analysis, Modeling, Simulation

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

DGLS	Differentialgleichungssystem
DoD	Details-on-Demand (Nachladen von Details nach Bedarf)
DVR	Direct Volume Rendering (Gruppe von Verfahren zum Rendering von 3D-Raumdaten)
gDGLS	gewöhnliches Differentialgleichungssystem
GIS	Geographische Informationssysteme
GUI	Graphical User Interface (deutsch: grafische Nutzerschnittstelle)
LIC	Line Integral Convolution (Abbildung eines statischen 2D-Strömungsfeldes durch Faltung der Stromlinien mit einem verrrauschten Bild)
PCA	Principal Component Analysis (Hauptkomponentenanalyse)
PIK	Potsdam Institut für Klimafolgenforschung
PK	Parallele Koordinaten
SOM	Self Organizing Map (selbstorganisierende Karte, vgl. Kohonen 1997)
SP	Scatterplot
SPM	Scatterplot-Matrix
VDM	Visuelles Data Mining

Danksagung

An dieser Stelle möchte ich allen danken, die mir bei meiner Dissertation hilfreich zur Seite standen. Insbesondere danke ich Heidi Schumann, die mir die Chance eröffnete, als Assistent an ihrem Lehrstuhl zu arbeiten. Mit ihrer zielführenden Unterstützung und ihrer konstruktiven Kritik hat sie maßgeblich zum Gelingen dieser Arbeit beigetragen. Besonderer Dank gilt auch meinen beiden Gutachtern Holger Theisel und Helwig Hauser für das Lesen und Bewerten dieser umfangreichen Arbeit. Rupert Klein möchte für die aktive Unterstützung der Kooperation zwischen dem Potsdam Institut für Klimafolgenforschung (PIK) und der Universität Rostock danken, welche die Grundlage für die hier vorgelegten Ergebnisse bildet. Ausserdem möchte ich Rupert Klein danken, dass er es mir ermöglichte, meine Arbeit am PIK abzuschließen.

Auch gilt mein Dank meinen Kollegen und Wegbegleitern für ihre Unterstützung und ihre Anregungen. Dies sind Friedrich, Matthias, Georg, Hermann, Mathias, Christian, René, Bernd, Peter, Hans-Jörg, Dietmar, Petra, Uwe und Rosi aus Rostock und Michael, Markus, Uwe, Claus, Martin und Stefan aus Potsdam. Dank gilt auch allen Studenten, welche in unterstützenden Arbeiten zum Erfolg dieser Arbeit beigetragen haben: Ralf, Mario, Daniela, Stephan, Karsten, Ronny, Martin, Daniel, Conrad, Andreas, Torsten, Hagen, Matthias, Anja, Ulrike, Martin und Henning.

Es ist mir besonders wichtig, auch meinen Eltern und meiner Petra für ihr Verständnis, ihre Geduld, und ihre Unterstützung in allen Belangen zu danken.

Kapitel 1

Einleitung und Motivation

Die Durchdringung aller Lebensbereiche der heutigen Gesellschaft mit moderner Informationstechnologie hat eine für den Einzelnen schwer zu überschauende Informationslandschaft geschaffen. Immense Datenmengen werden in digitaler Form gespeichert und verwaltet. Um das darin enthaltene Wissen aufzuschließen, sind spezielle, auf den Kontext der Daten zugeschnittene Konzepte und Methoden zu deren Analyse erforderlich. Ziel hierbei ist es, die außerordentlichen Fähigkeiten der menschlichen Kognition mit den Fähigkeiten heutiger Computersysteme zur Verarbeitung extrem großer Datenmengen zu koppeln. Eine solche Kopplung ermöglicht insbesondere für den wissenschaftlichen Erkenntnisprozess neue Einsichten in die Daten und die den Daten zugrunde liegenden Phänomene zu erlangen, Schlussfolgerungen über damit verbundene Theorien und Modelle zu ziehen und diese Ergebnisse zu kommunizieren.

Eine essentielle Methode, die Lücke zwischen abstrakten, im Computer gespeicherten Daten auf der einen Seite und der menschlichen Kognition auf der anderen Seite zu schließen, ist die Visualisierung. Bei der Visualisierung werden für die Daten geeignete graphische Darstellungen erzeugt, welche es dem Betrachter erlauben, einen intuitiven Zugang zu den den Daten zugrunde liegenden Werten und deren Beziehungen zu erhalten. Vorteil beim Einsatz von Visualisierungen ist, dass sie es basierend auf den menschlichen Assoziations- und Mustererkennungsfähigkeiten erlauben, große, abstrakte Datenmengen zu analysieren und die Ergebnisse solcher Analysen leicht verständlich zu kommunizieren. Darüber hinaus bieten moderne Visualisierungssysteme einen hohen Grad an Interaktivität, welche die Möglichkeiten, in den Daten vorliegende, versteckte Muster aufzudecken, wesentlich verbessern.

Weiterhin sind automatische Analysemethoden, welche insbesondere Verfahren der Statistik und des maschinellen Lernens beinhalten, bei den heute vorliegenden immensen Datenmengen nicht mehr wegzudenken. Dabei wird die besondere Fähigkeit moderner Computersysteme, numerische Berechnungen und Suchvorgänge schnell und genau durchzuführen, vor allem zur Strukturierung der Daten ausgenutzt. Solche Techniken werden auch unter dem Oberbegriff *Data Mining* - Methoden zusammengefasst.

Ein aktueller Forschungsschwerpunkt im Umfeld des *Data Mining* und der Visualisierung ist, die Potenz solcher automatischer Analyseverfahren mit der Potenz menschlicher Wahrnehmungsfähigkeiten bei der Visualisierung zu koppeln. Entsprechende Techniken und Systeme werden unter dem Begriff *Visuelles Data Mining (VDM)* zusammengefasst.

Ansatz dieser Arbeit ist es, die Methoden des visuellen Data Mining speziell für das Anwendungsgebiet der Klimaforschung einzusetzen und weiterzuentwickeln. Dies stellt die Basis für ein neues methodisches Vorgehen in diesem Umfeld dar und hat das Ziel, es Klimaforschern zu ermöglichen, bei der Untersuchung von Klimadaten und -modellen neue Erkenntnisse zu gewinnen.

1.1 Herausforderungen

Problemstellung beim visuellen Data Mining ist es, die automatischen mit den visuellen Methoden eng zu verzahnen, um dabei auftretende Synergien bestmöglich zu nutzen. Dies geht über eine reine Darstellung der Ergebnisse automatischer Berechnungen deutlich hinaus, wobei die Ergebnisse und Daten aus einem Verfahren direkt durch andere Verfahren verarbeitet oder zu deren Auswahl oder Parametrisierung genutzt werden. Inbesondere hat hierbei gerade die Methodenkopplung ein hohes Potential, die Ergebnisse automatischer Berechnungsverfahren transparent zu gestalten und dadurch das Verständnis in diese Verfahren, deren Parameter und deren Resultate zu steigern. Eine spezielle Problemstellung hierbei ist die Verzahnung von statistischen Verfahren mit Visualisierungstechniken, wenn die Daten im räumlichen oder zeitlichen Kontext vorliegen.

Eine wichtige Anwendung des visuellen Data Mining, die zunehmende Bedeutung erlangt, ist die Unterstützung des gesamten Modellierungs- und Simulationsprozesses. Die Herausforderung hierbei besteht darin, deutlich über eine reine Analyse bzw. Darstellung von Simulationsdaten im Sinne eines „Postprocessing“ hinauszugehen. Dies betrifft die Datenanalyse im Vorfeld der Modellbildung im Sinne einer Hypothesenbildung und -validierung genauso wie die Untersuchung der Modellstruktur von zum Teil komplexen Modellen bis hin zur Auswertung komplexer Simulationsexperimente und der Modellvalidierung. Das Potential moderner VDM-Verfahren in diesem Prozess wird bisher erst in Ansätzen ausgeschöpft. So muss beispielsweise untersucht werden, inwieweit sich zum Teil sehr resourcenaufwendige Simulationen in einen interaktiven Analyseprozess einbinden und verarbeiten lassen.

Eine weitere Herausforderung für diese Arbeit ergibt sich, wenn die für allgemeine Problemstellungen konzipierten VDM-Methoden in einem speziellen Anwendungskontext – wie in diesem Fall der Klimaforschung – eingesetzt werden sollen. Dabei besteht eine Lücke zwischen zum Teil schwer bedienbaren Analysewerkzeugen mit einer Vielzahl in der Anwendung unbekannter (zumeist multivariater) Verfahren auf der einen Seite, und dem Wissen der Anwender und den Konventionen in deren Umfeld auf der anderen Seite. Auch werden die Chancen, die sich durch eine hochgradig interaktive Analyse verknüpfter Verfahren ergeben, in Anwendungen wie der Klimaforschung erst in Ansätzen ausgenutzt. Eine spezielle Problemstellung dieser Arbeit stellen die großen, heterogenen Datensätze der Klimaforschung dar:

- von teilweise hochauflösenden Daten über lange Zeitperioden,
- die hohe Zahl von abhängigen Variablen (teilweise > 100),
- von Klimadaten auf abweichenden Gittern,
- deren Kombination mit statistisch abgeleiteten Größen.

1.2 Ziele der Arbeit

Ansatz dieser Arbeit ist es, systematisch zu untersuchen, wie die Lücke zwischen allgemein einsetzbaren Methoden des visuellen Data Mining und speziellen Methoden und Konventionen einer Anwendung geschlossen werden kann. Dies schließt zuallererst ein zu untersuchen, inwieweit sich Methoden der Visualisierung auf das spezielle Anwendungsgebiet der Klimaforschung zuschneiden lassen. Darüber hinaus werden bei Bedarf neuartige Visualisierungstechniken - inbesondere für die multi-variate Datenanalyse und die Darstellung von Ergebnissen von Cluster- und Hauptkomponentenanalysen im zeitlichen und/oder räumlichen Bezug - entworfen.

Ein Schwerpunkt der Arbeit liegt auf der Unterstützung bei der Auswertung von simulierten Daten und der damit verbundenen Modellvalidierung. Dazu werden Verfahren zur vergleichenden Analyse von Daten - gerade auch vor dem Hintergrund abweichender Gitter - systematisch untersucht und

neuartige Methoden vorgestellt.

Die eingesetzten, angepassten und neu entwickelten Methoden bilden die Basis für eine systematische Unterstützung des gesamten Modellierungs- und Simulationsprozesses. In diesem Kontext entwirft diese Arbeit ein allgemeines Vorgehensmodell für die Anwendung von Verfahren und Methoden des visuellen Data Mining anhand eines Beispieles aus der Klimaforschung.

Um die Vielfalt an Visualisierungsmethoden, welche in unterschiedlichen Enwicklungsumgebungen umgesetzt wurden, handhabbar zu machen, werden diese in eine Bibliothek integriert. Diese Bibliothek bildet die Basis für eine Verwendung der Techniken in verschiedenen Szenarien, z.B. bei der Visualisierung in einer Simulationsexperimentierumgebung im Framework SimEnvVis.

Ein weiterer Schwerpunkt dieser Arbeit resultiert aus der Erkenntnis, dass die reine Bereitstellung von Analysetechniken in einem VDM-Framework nicht ausreicht, um dessen praktische Einsetzbarkeit abzusichern, da die Vielzahl von VDM-Techniken mit einer Vielzahl von Parametern häufig die Anwender überfordern. Neben dem Einsatz von „History-Mechanismen“ zur Verwaltung des Analyseprozesses (vgl. z.B. Kreuseler u. a. (2004)) schließt dies insbesondere ein, Visualisierungstechniken so auszuwählen und zu parametrisieren, dass expressive und effektive Darstellungen entstehen, die genau für die aktuelle Problemstellung geeignet sind. Deswegen untersucht diese Arbeit, wie Anwendern mit speziellem Hintergrund eine geeignete Schnittstelle an die Hand gegeben werden kann, mit welcher die Charakteristik der Daten und ihre Analyseziele geeignet spezifiziert, erhoben und in den Analyseprozess einbezogen werden können.

Zielstellung hierbei ist es, ein möglichst breites Wissen über die Datencharakteristik in den VDM-Prozess einzubringen. In diesem Zusammenhang spricht man auch von Metadaten, also, Daten, welche die Daten beschreiben. Im allgemeinen ist es nicht trivial, welche Metadaten für ein effektives visuelles Data Mining gebraucht werden. Diese Arbeit stellt sich dieser Herausforderung und führt eine allgemeine Konzeption von Metadaten sowie ein Konzept und dessen Umsetzung zur Erhebung von Metadaten vor und demonstriert ihren Einsatz im VDM.

Eine weitere Voraussetzung für ein effektives VDM ist die Formulierung des Analyseziels und dessen Nutzung im Analyseprozess. Herausforderung hierbei ist es, die verschiedenen Begrifflichkeiten von Anwendern und Entwicklern von VDM-Software aufeinander abbilden zu können und für die Auswahl und Parametrisierung der Techniken nutzbar zu machen. Hierzu schlägt die vorliegende Arbeit eine neue Methodik zur Spezifikation von Zielen- und Aufgaben vor und präsentiert deren Einsatz.

Basierend auf diesen Einflussfaktoren lassen sich (halb-) automatisch geeignete Visualisierungsmethoden auswählen, diese geeignet miteinander verknüpfen und passend parametrisieren. Ein solches Vorgehen wird im Umfeld der Visualisierung auch als *Visualisierungsdesign* bezeichnet, und ermöglicht es, die Lücke zwischen vorhandenen, zum Teil in der Anwendung unbekannten Techniken und dem Wissen der potentiellen Nutzer solcher Techniken zu verringern, und dadurch das Potential moderner VDM-Methoden zur Findung neuer Erkenntnisse für eine spezielle Anwendung besser auszunutzen.

1.3 Ergebnisse der Arbeit

Im Laufe dieser Arbeit wird ein allgemeines Konzept für eine Framework-Architektur für das visuelle Data Mining und dessen prototypische Umsetzung am Beispiel der Modellbildung und Simulation in der Klimaforschung vorgestellt. Hierfür werden eine Vielzahl von bekannten Verfahren an dieses Problemumfeld angepasst und neue Vorgehensweisen entwickelt und umgesetzt. Dabei beschreitet die vorliegende Arbeit insbesondere in den folgenden Punkten Neuland:

- Entwicklung von **Visualisierungstechniken für geclusterte Daten in Raum und Zeit**

(Kreuseler u. a. 2003; Nocke u. a. 2004, 2005),

- systematische Kopplung von **Visualisierung und Hauptkomponentenanalyse** in allen Schritten des Visualisierungsprozesses (Müller u. a. 2006),
- systematische Untersuchung der Einsatzmöglichkeiten von **Visualisierungsfunktionalität im Gebiet der Klimaforschung** (Nocke u. a. 2003, 2004; Böhm u. a. 2004; Nocke u. a. 2005, 2007)
- Entwurf einer allgemeinen **Methodik zur vergleichenden Visualisierung** unter Einbeziehung von Daten auf abweichenden Gittern (Nocke u. a. 2007),
- Beschreibung einer **Methodik zum Einsatz von visuellen Data Mining-Verfahren für den Prozess der Modellierung und Simulation** (Nocke u. a. 2003; Schulz u. a. 2006b; Flechsig u. a. 2007; Nocke u. a. 2007),
- allgemeine und anwendungsspezifische **Spezifikation von Metadaten für das visuelle Data Mining** als wichtige Schnittstelle innerhalb des VDM-Prozesses sowie Konzeption und Umsetzung von **Strategien zur Erhebung und Auswertung solcher Metadaten** (Nocke 2000; Kreuseler u. a. 2003; Nocke u. Schumann 2002; Lange u. a. 2006),
- verallgemeinerte **Spezifikation von Zielen und Aufgaben** innerhalb des VDM-Prozesses, welche die Nutzerintension in der aktuellen Exploration beschreiben, und auf das Vokabular des Anwendungshintergrundes flexibel angepasst werden können (Nocke u. Schumann 2004),
- **Einführung eines neuartigen Ansatzes zum Visualisierungsdesign**, der es erlaubt, anwendungsspezifische und anwendungsunabhängige Regeln zur Auswahl und Parametrisierung von Visualisierungstechniken miteinander zu verknüpfen (Lange u. a. 2006; Nocke u. a. 2007).

1.4 Struktur der Arbeit

Die Arbeit gliedert sich wie folgt: zuerst werden in Kapitel 2 grundlegende Begriffe geklärt sowie ein kurzer Einblick in die speziellen, für diese Arbeit relevanten Problemstellungen der Klimaforschung gegeben. Danach wird ein Überblick über den Stand der Forschung der in dieser Arbeit behandelten Themenkomplexe gegeben (Kapitel 3). Daran schließt sich der Entwurf einer Architektur eines allgemeinen VDM-Frameworks und die speziellen Aspekte für die Analyse von Klimadaten in einem solchen Framework an (Kapitel 4). Im Anschluss daran werden die umgesetzen und zum Teil neu- oder weiterentwickelten Visualisierungstechniken - mit den Schwerpunkten auf der Analyse von Klimadaten im räumlichen, im zeitlichem Bezug und im Merkmalsraum sowie auf der vergleichenden visuellen Analyse - vorgestellt und systematisiert (Kapitel 5). Daran schließt sich eine systematische Untersuchung der Kombination von visuellen und automatischen Mininig-Techniken an (Kapitel 6). Exemplarisch werden hierbei die Verzahnung von Visualisierungsmethoden und Verfahren der Clusteranalyse sowie von Visualisierungsmethoden und Verfahren der Hauptkomponentenanalyse untersucht sowie die Einsatzmöglichkeiten der Verknüpfung von VDM-Verfahren für die Modellbildung und Simulation diskutiert und illustriert. Anschließend werden die Konzepte zu Metadaten, Analysezielen und -aufgaben sowie zu einem Visualisierungsdesign vorgestellt (Kapitel 7), welche Anwender aus einem speziellen Anwendungsgebiet wie der Klimaforschung dabei unterstützen sollen, den VDM-Prozess gemäß ihres Analysekontextes geeignet zu steuern. Abschließend erfolgt eine Zusammenfassung der Ergebnisse sowie ein Ausblick auf verbleibende offene Probleme (Kapitel 8).

Kapitel 2

Grundlagen

In diesem Kapitel werden für die Arbeit grundlegende Begriffe (vgl. Abs. 2.1) eingeführt und der Anwendungshintergrund der Klimaforschung vorgestellt (vgl. Abs. 2.2).

2.1 Grundlegende Begriffe

Meteorologie: Meteorologie ist „die Lehre von den Vorgängen in der Lufthülle der Erde“, und ist eng mit der Ozeanographie und Geophysik verknüpft (Brockhaus 1970). Vom Stamm „meteo“ kann abgeleitet werden, dass sie sich auf Wetterphänomene konzentriert.

Klimatologie: In großer inhaltlicher Nähe zur Meteorologie steht die Klimatologie oder Klimakunde. Sie ist nach Brockhaus (1996) die „Wissenschaft vom Klima, der Klimaänderung und deren Auswirkungen“. Sie ist „primär Teilgebiet der Meteorologie, in den erdkundlichen Bezügen und Auswirkungen auch der Geographie [...], bzgl. der paläoklimatologischen Rekonstruktionen auch der Geologie, Glaziologie und Biologie, in den Grundlagen der Physik und Chemie u.a., somit ausgeprägt interdisziplinär. Inhaltlich gliedert sie sich in die klimatologische Informationserfassung [...], die Klimadiagnostik [...], die Klimamodellierung sowie die Betrachtung [...] der Klimawirkungsforschung.“

Multi-Run-Simulationsexperiment: Bei einem Multi-Run-Simulationexperiment wird die Änderung der Zustände eines Modells in mehreren Simulationsläufen bezüglich veränderter Modelleingaben (z.B. Modellparameter oder -startwerte) getestet. So lässt sich z.B. untersuchen, wie „sensitiv“ eine Ausgabegröße auf Variationen bestimmter Eingabegrößen reagiert (vgl. Cooke u. van Noortwijk 2000).

Visualisierungspipeline: Die Visualisierungspipeline fasst die Hauptschritte bei der Erzeugung einer visuellen Darstellung zusammen. Dies sind (in dieser Reihenfolge) die Schritte Vorverarbeitung, Mapping und Rendering.

Datenmenge: Der Begriff Datenmenge umfasst die Gesamtheit aller in den Analyseprozess einfließenden Daten.

Beobachtungsraum: Von Schumann u. Müller (2000) wird der „Raum, in dem die Daten erhoben werden, als Beobachtungsraum“ bezeichnet (vgl. Schumann u. Müller 2000, S. 29). Dabei wird bewusst davon abstrahiert, ob es sich um einen konkreten physikalischen oder einen abstrakten Beobachtungsraum handelt.

Dimension: Die Dimensionen des Beobachtungsraumes werden auch als unabhängige Variable bezeichnet. Diese können je nach Art des Raumes Ortskoordinaten, Zeitachsen und/oder abstrakte Dimensionen sein.

Beobachtungspunkt: Ein Beobachtungspunkt ist ein Punkt des Beobachtungsraumes, für den Daten vorliegen. Er lässt sich als Vektor der Koordinatenwerte der Dimensionen des Beobachtungsraumes beschreiben.

Datensatz: Ein Datensatz ist die Menge aller Daten eines Beobachtungspunktes.

Merkmal: Merkmale sind Größen, die im Beobachtungsraum erhoben werden. Sie werden auch als abhängige Variable bezeichnet.

Multivariat: Liegen an einem (oder auch mehreren) Beobachtungspunkt(en) einer Datenmenge mehrere Markmalswerte vor, so spricht man von einer multivariaten Datenmenge.

Gitter: Ein Datengitter DG sei definiert als ein 5-Tupel $DG = (V, L, C, M, I)$, mit

$$\begin{array}{ll} V & \text{Menge von Beobachtungspunkten im Gitter (Gitterpunkte)} \\ L & \text{Menge der Gitterlinien (Verbund)} \\ C & \text{Menge von Gitterzellen} \\ M & \text{Menge von Merkmalen pro Beobachtungspunkt} \\ I & \text{Menge der zulässigen Interpolationsverfahren.} \end{array} \tag{2.1}$$

Wirkungskreis: Der Wirkungskreis WK sei definiert als ein Tupel $WK = (V, U)$, mit U als der Menge von Umgebungen, in die die zulässigen Interpolationsverfahren I die Merkmale M um die Gitterpunkte V inter- und extrapolieren dürfen.

Gittertypen: In der Literatur werden verschiedene Gittertypen vorgestellt. Angelehnt an Schumann u. Müller (2000) und Frühauf (1997) werden die folgenden Basisgittertypen unterschieden:

- strukturierte Gitter (beliebige Gitterlinien mit äquidistanten Gitterpunkten)
  - regelmäßige Gitter (achsenparallele Gitterlinien mit konstanten Längen),
  - blockstrukturierte Gitter (achsenparallele Gitterlinien mit variierenden Längen),
  - curvilineare Gitter (regelmäßiges/blockstrukturiertes Gitter mit nichtlinearen Abbildungen der Gitterpunkte),
- unstrukturierte oder unregelmäßige Gitter (beliebige Gitterlinien mit beliebigen Abständen) und
- hybride Gitter.

Abbildung 2.1 illustriert die verschiedenen Gittertypen.

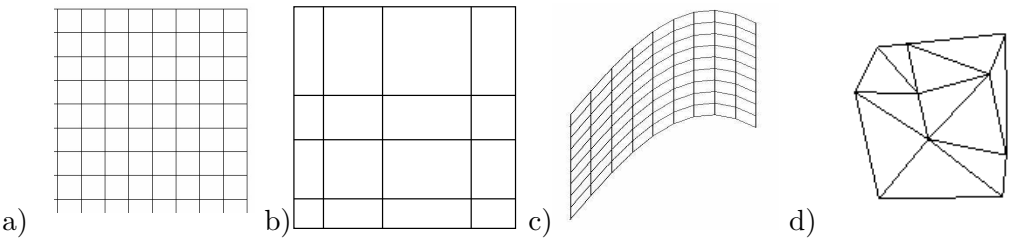


Abbildung 2.1: Illustration verschiedener Gittertypen: a) regulär b) blockstrukturiert c) curvilinear d) unregelmäßig

Metadaten: Metadaten sind „Daten über Daten“. Das bedeutet, dass sie Daten näher beschreiben, also Zusatzinformationen über sie darstellen.

Datenklasse: Der Begriff der Datenklasse stellt ein Schema für Daten dar. Unterschiedliche Datenklassen haben spezifische Eigenschaften, welche in der Visualisierung spezifisch dargestellt werden sollten. Anhand der Ausprägungen der Metadaten - insbesondere anhand der Anzahl und Art der Merkmale sowie des zugrunde liegenden Gitters - können bestimmte Datenmengen auf bestimmte

Datenklassen abgebildet werden. In der Visualisierung werden üblicherweise die Datenklassen Multiparameterdaten, Volumendaten, Strömungsdaten und gestreute Daten ($L = \emptyset$ und $C = \emptyset$) unterschieden.

Explorative Analyse: Die explorative Analyse (oder auch Datenexploration) bezeichnet die ungerichtete Suche in den Daten, um erste Muster aufzudecken. Vor allem geht es hierbei darum, ohne Vorwissen auch völlig unerwartete Erkenntnisse erzielen zu können.

Konfirmative Analyse: Die konfirmative Analyse beinhaltet eine gerichtete Suche in den Daten, wobei inbesondere Hypothesen über die den Daten zugrunde liegenden Phänomene validiert werden sollen.

Datenpräsentation: Bei der Datenpräsentation werden Erkenntnisse aus der explorativen und/ oder konfirmativen Analyse für die Kommunikation in verschiedenen Medien wie Printmedien oder Fernsehen aufbereitet. Dabei steht die leicht verständliche Darstellung der erzielten Erkenntnisse im Vordergrund.

2.2 Anwendungshintergrund

Die Klimaforschung (bzw. Klimatologie) untersucht und beschreibt das Klima und dessen langfristige Veränderungen. Neben der Identifikation und Quantifizierung wichtiger Einflussfaktoren auf das Klima beinhaltet dies, Methoden und Modelle für die Abschätzung zukünftiger Klimabedingungen zu entwickeln und zu verbessern, und dabei quantitative Aussagen z.B. über grundlegende Trends und das Auftreten von Extremereignissen abzuleiten.

Insbesondere sind langfristige Klimaszenarien Unsicherheiten bezüglich verschiedener Einflussfaktoren unterworfen. Mit der Reduktion dieser Unsicherheiten befassen sich eine Vielzahl von Forschungsinstituten, um gesellschaftlich verwertbare Aussagen über die Klimaentwicklung, und über deren ökologische, ökonomische und soziale Folgen zu gewinnen. Forschungsschwerpunkte hierbei sind ein verbessertes Verstehen von Klimaphänomenen, die Weiterentwicklung von Klimaszenarien und Analyseverfahren für zukünftige Witterungsbedingungen und die Verbesserung existierender Klimamodelle. Eines der Kernprobleme hierbei stellen die enormen Datenmengen dar, die durch Klimamessungen und Klimasimulationen entstehen, sowie deren effektive Nutzung und Verarbeitung durch den Menschen. An diesem Punkt setzt die hier vorgelegte Arbeit an, um exemplarisch am Beispiel von Klimadaten zu demonstrieren, wie mit Hilfe moderner, computergestützter Analyseverfahren die steigende Datenflut simulierter und gemessener Erdsystemdaten handhabbar gemacht werden kann.

Im Rahmen einer mehrjährigen Kooperation mit dem Potsdam Institut für Klimafolgenforschung wurden neben der Umsetzung verschiedener Visualisierungstechniken vor allem auch grundlegende Vorgehensweisen an Explorations- und Evaluationsaufgaben im Prozess der Modellierung und Simulation untersucht und teilweise auch in praktisch einsetzbare Software überführt (Nocke u. a. 2003). Dabei konnten eine Vielzahl von Herausforderungen und Aufgaben auf dem Gebiet der Klimaforschung identifiziert werden.

So ist ein vordringliches Ziel der Klimaforschung, Klimaprozesse besser zu verstehen, und im Anschluss daran dieses Wissen in die Weiterentwicklung von Klimamodellen einfließen zu lassen. Dies beinhaltet die Qualität von Modellsimulationen zu verbessern und dabei wichtige Eigenschaften des natürlichen Systems abzubilden.

Entsprechend werden im Rahmen dieser Arbeit Verfahren untersucht, welche die Analyse gemessener und simulierter klimatischer Daten verbessern und so die Klimadiagnostik und -modellierung unterstützen. Im Fokus des Interesses stehen Techniken, die für Daten auf längerfristigen Zeitskalen entworfen wurden, und damit liegt der Schwerpunkt dieser Techniken auf der Analyse längerfristiger

Klimaphänomene wie z.B. ozeanischer Zirkulationen oder langjähriger Messreihen.

In diesem Kontext lassen sich folgende wichtige Problemfelder im Grenzbereich zwischen visuellen Data Mining und Klimaforschung identifizieren:

1. Exploration von Mustern in Raum- und Zeit (z.B. von extremen Witterungsbedingungen),
2. Vergleich von Klimadaten aus Messdaten oder Simulationen,
3. Vereinfachung und Evaluation von Klimamodellen.

Ein Beispiel zur Verdeutlichung dieser Problemfelder findet sich in einer gemeinsamen Veröffentlichung mit dem Potsdam Institut für Klimafolgenforschung (vgl. Böhm u. a. 2004). Hier wird u.a. am Beispiel eines regionalen Klimamodells für das Gebiet Europa eine grundlegende Vorgehensweise zur Evaluation der Modell-Verlässlichkeit durchgeführt. Dabei werden interessente bodennahe Modellparameter identifiziert und diese auf extreme Muster – in diesem Fall z.B. auf ihre Ausprägung im Fall extrem heißer Sommer – hin untersucht (Punkt 1). Im Anschluss daran werden die simulierten Modellresultate aller Parameter mit Referenzdaten verglichen (Punkt 2). Danach werden die einzelnen Modellparameter darauf hin untersucht, wie stark sie aufgrund von Simulationsfehlern zur Ungenauigkeit der simulierten Modellresultate beitragen. In einem abschließenden Schritt wird untersucht und bewertet, ob das Modell (mit einer gewissen Genauigkeit) in der Lage ist, die realen klimatischen Prozesse adäquat wiederzugeben (Punkt 3).

Das diesem Beispiel zugrunde liegende Vorgehen (vgl. Böhm u. a. 2004) ist ein erster Ansatz zur Kopplung von Visualisierung (z.B Themenflussdarstellung) und statistischer Analyse (z.B. Clusteranalysen), lässt jedoch viel Raum für vertiefende Untersuchungen, die im Rahmen dieser Arbeit genauer disskutiert werden.

Kapitel 3

Stand der Forschung

In diesem Kapitel werden die aktuellen Arbeiten zu den vier Hauptschwerpunkten dieser Arbeit beschrieben. Dies sind *Visuelles Data Mining*, *Visualisierung und Datenanalyse in der Klimaforschung*, *Systeme* in diesem Umfeld und verschiedene *Design*-Aspekte für das Visuelle Data Mining. Dabei wird diskutiert, inwieweit aktuelle Verfahren für die Problematiken im Bereich der Analyse von Klimadaten geeignet sind und offene Punkte herausgearbeitet. Abbildung 3.1 stellt die Gliederung dieses Kapitels graphisch dar.

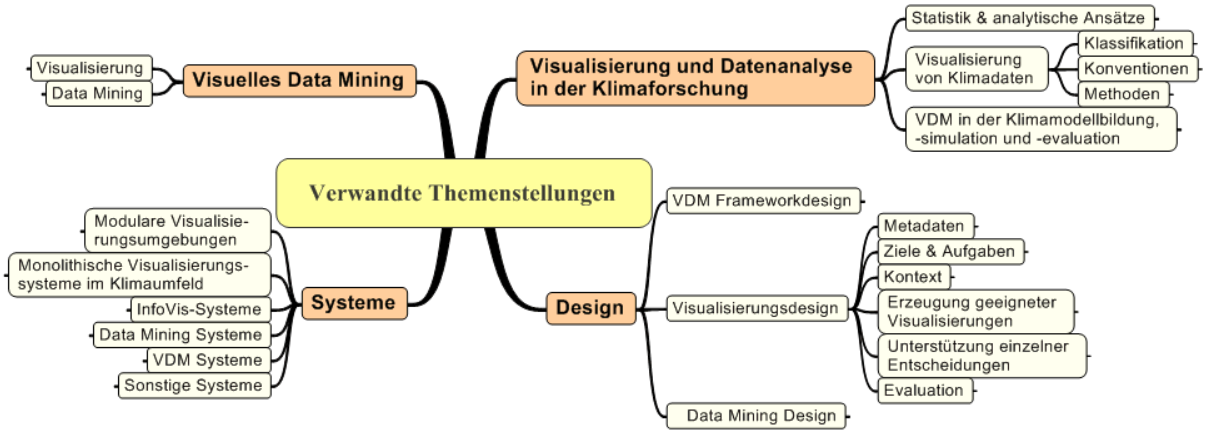


Abbildung 3.1: Überblick über verwandte Themenstellungen

Der Themenschwerpunkt *Visuelles Data Mining* (Abs. 3.1) gibt einen kurzen Überblick über aktuelle Verfahren aus den Bereichen Visualisierung und Data Mining sowie deren Verknüpfung und stellt grundlegende Prinzipien vor. Der Themenschwerpunkt *Visualisierung und Datenanalyse in der Klimaforschung* (Abs. 3.2) beschreibt statistische, visuelle und analytische Verfahren zur Analyse von Klimadaten mit dem Schwerpunkt auf der Visualisierung. Im Anschluss daran werden *existierende Systeme* im Umfeld VDM und Klimaforschung auf ihre Anwendbarkeit und Einsetzbarkeit zur Lösung der Aufgaben beim VDM von Klimadaten hin untersucht (Abs. 3.3). Der Themenschwerpunkt *Design* (Abs. 3.4) stellt zum einen Arbeiten zum Design von VDM- und Visualisierungssystemen unter dem softwaretechnischen Blickpunkt vor (Abs. 3.4.1). Zum anderen werden Arbeiten, welche (halb-) automatisch geeignete Visualisierungs- und Miningtechniken generieren, auswählen und/oder parametrisieren, vorgestellt (Abs. 3.4.2 und 3.4.3). Das Hauptaugenmerk hierbei liegt auf der Kontextbeschreibung durch Metadaten, Aufgaben und Zielen, sowie dem Treffen von Visualisierungsentscheidungen. Abschließend wird Standes der Forschung zusammengefasst (Abs. 3.5).

3.1 Visuelles Data Mining

Visuelles Data Mining wird von Keim u. a. (2002) als eine Kombination von traditionellen Data Mining-Techniken mit Informationsvisualisierungstechniken eingeführt. Es kombiniert die Stärke automatischer Analysetechniken mit den menschlichen Wahrnehmungs- und kognitiven Fähigkeiten. VDM-Techniken sind im besonderen dann geeignet,

- wenn es darum geht, große Datenmengen zu explorieren,
- wenn noch wenig über die Daten bekannt ist, die Ziele vage sind oder
- wenn die Daten sehr inhomogen oder verrauscht sind

(vgl. Keim u. a. 2002). Dabei erfolgt zumeist eine interaktive Exploration, bei der der Anwender direkt in den Analyseprozess eingebunden ist, und so ein Verständnis der unterliegenden automatischen Algorithmen und Parameter nicht (unbedingt) notwendig ist. So kann die Exploration beschleunigt, häufig bessere Resultate als bei reiner Ausführung automatischer Verfahren erzielt und das Vertrauen in deren Ergebnisse erhöht werden (a.a.O.). Ziel hierbei ist nicht, automatische Verfahren durch visuelle zu ersetzen, sondern die Vielfalt verschiedener Methoden für den Explorationsprozess miteinander zu koppeln (Westphal u. Blaxton 1998). Shneiderman (2002) motiviert die Notwendigkeit dieser Kopplung, indem er auf die Gefahren bei schwindendem Verständnis und abnehmender Kontrolle von automatischen Methoden – häufig als Blackbox – hinweist. In diesem Sinne kann der Nutzer in einem interaktiven, stark visuell geprägten Prozess einen besseren Einblick in Eignung und Eigenschaften verschiedener Verfahren erlangen. Auf der anderen Seite kann so auch der Gefahr der vielen, wunderschönen Bilder in der Informationsvisualisierung begegnet werden, die vom eigentlichen Kern der Daten ablenken können (Spence 2001).

Entsprechend dem Herkunfts-Forschungsgebiet können zwei wesentliche Zugangsweisen zum VDM ausgemacht werden. Zum einen werden Visualisierungsmethoden von Statistikern entworfen und eingesetzt, insbesondere um die Ergebnisse statischer Verfahren evaluieren zu können (vgl. z.B. Wills 1999; Unwin u. Theus 2004) und Statistiksoftware zu erweitern (z.B. SPSS 2007). Zum anderen werden statistische Methoden in der Visualisierung häufig als Vorverarbeitungsschritt eingesetzt, z.B. zur Clusterung oder zur Dimensionsreduktion (vgl. z.B. van Wijk u. van Selow 1999; Davidson u. a. 2001; Ross u. Chalmers 2003; dos Santos u. Brodlie 2004). Das Zusammenwachsen von Visualisierung auf deren einen Seite, und Statistik und Maschinellem Lernen auf der anderen Seite, ist jedoch nicht immer unproblematisch, da die verschiedenen Disziplinen entweder eher der Nutzerkontrolle oder der Stärke statistischer und automatischer Verfahren vertrauen (Shneiderman 2002).

Neuere Arbeiten haben die Kopplung von automatischen und visuellen Mining-Methoden weiterentwickelt. Ankerst (2001) unterscheidet in diesem Zusammenhang die Kopplung von VDM-Verfahren auf drei Ebenen:

1. die unabhängige Anwendung von Visualisierungstechniken und Data Mining - Algorithmen,
2. der Einsatz von Visualisierung um Muster und Resultate aus Data Mining - Algorithmen graphisch zu repräsentieren und
3. die enge Kopplung von Visualisierungs- und Data Mining - Methoden, die u.a. die Visualisierung auch von Zwischenergebnissen mit einschließt.

Insbesondere geht es bei aktuellen Forschungsanstrengungen darum, eine enge Verzahnung der Methoden auf der dritten Ebene nach Ankerst (2001) zu erreichen, welche einschließt, dass bereits Zwischenergebnisse des Mining-Prozesses visualisiert werden, aufgrund derer der Anwender direkt auf die automatischen Mining-Verfahren einwirken kann. So können Teilprozesse graphisch verdeutlicht, und der Nutzer mit seinem Hintergrundwissen direkt einbezogen werden. Ein Beispiel für nützliches visuelles Feedback ist die Vermeidung langwährender Berechnungen, falls die Start-

parameter nicht geeignet gewählt wurden. Darstellungen von Zwischenergebnissen für ein visuelles Feedback im Sinne des VDM auf der dritten Stufe finden sich unter anderem in visuell gestützten Datenbankanfragen (vgl. z.B. Avnur u. a. 1998; Hinneburg u. a. 1999), bei der Benutzung von Clusteralgorithmen (vgl. z.B. Ribarsky u. a. 1999) und bei der Klassifikation basierend auf Entscheidungsbäumen (vgl. z.B. Ankerst u. a. 2000).

Grundsätzlich werden beim (visuellen) Data Mining zwei Zugänge unterschieden: Extraktion bisher unbekannter Muster und Hypothesenbildung in den Daten (explorative Analyse) sowie die gerichtete Suche zur Verifikation von Hypothesen (konfirmative Analyse). Zusätzlich lassen sich zwei grundlegende Zielstellungen unterscheiden: Beschreibung von in den Daten enthaltenen Eigenschaften sowie Prognose zukünftiger Trends oder zu erwartender Eigenschaften in den Daten.

Die aufgeführten Arbeiten zum visuellen Data Mining konzentrieren sich auf den Bereich der Exploration (explorative Analyse), und wenden dazu vor allem statistische und Informationsvisualisierungsmethoden an. Weiterhin gibt es erst kaum Arbeiten, welche die Methoden des VDM auch auf die konfirmative Analyse anwenden. Shneiderman (2002) plädiert dafür, die Hypothesenüberprüfung stärker in den VDM-Prozess einzubeziehen, und schlägt vor, Exploration und Konfirmation gleichberechtigt zu behandeln und beide im VDM miteinander zu koppeln. Diese Kopplung ist jedoch im Rahmen des VDM ein noch weitgehend offenes Forschungsfeld. Im Rahmen dieser Arbeit werden hierzu neue Vorgehensweisen vorgestellt (vgl. Kap. 4 und 6).

Als Basismodell für die Datenbeschreibung im VDM führen Kreuseler u. Schumann (2002a) ein Informationsmodell ein, welches 3 wesentliche Aspekte beinhaltet:

1. die Menge der Informationsobjekte,
2. ihre Attribute (die einen Informationsraum aufspannen) und
3. die Struktur zwischen den Informationsobjekten (Relationen zwischen den Informationsobjekten).

Basierend auf diesem Informationsmodell können dann die verschiedenen Visualisierungsmethoden (siehe Abs. 3.1.1), aber auch automatische Mining-Methoden klassifiziert werden.

Dieses Modell wurde insbesondere für die Beschreibung abstrakter Daten[1] entworfen, wobei der Raumbezug nicht im Vordergrund steht. Diese Sicht spiegelt sich auch in der Ausrichtung aktueller Arbeiten im Bereich des VDM wieder. Hier setzt die vorgelegte Arbeit an und zeigt Wege auf, wie beliebige Datenklassen in den VDM-Prozess einbezogen werden können.

Allgemeine Übersichten zum visuellen Data Mining finden sich in Wong (1999); Keim (2002); Keim u. Ward (2002); de Oliveira u. Levkowitz (2003). Darüber hinaus sind für diese Arbeit auch spezielle Aspekte zur Visualisierung, zu automatischen Data Mining Verfahren und zu existierenden Systemen von Interesse. Im folgenden werden Arbeiten in diesen Gebieten genauer untersucht.

3.1.1 Visualisierung

Im Umfeld der Visualisierung ist es üblich, zwei Forschungsbereiche zu unterscheiden: wissenschaftliche Visualisierung (Scientific Visualization) und Informationsvisualisierung (Information Visualization). Die wissenschaftliche Visualisierung beschäftigt sich dabei überwiegend mit Daten, die in einem räumlichen Bezug gegeben sind (z.B. mit Strömungs- und Volumendaten), während sich die Informationsvisualisierung überwiegend mit abstrakten Daten befasst. Die Visualisierung zeitlicher Phänomene wird in beiden Gebieten behandelt. Die Grenzen dieser beiden Gebiete sind jedoch weitgehend fließend, da häufig auch räumliche Daten mit Techniken der Informationsvisualisierung

[1] Daten, die keinen physikalischen räumlichen Bezug haben

dargestellt werden, und auch Prinzipien der Informationsvisualisierung bei räumlichen Visualisierungstechniken eingesetzt werden können (vgl. hierzu z.B. Tory u. Möller 2004b).

Auch im Rahmen der Visualisierung von Klimadaten wird die gesamte Palette von Techniken für räumliche, zeitliche und abstrakte Daten benötigt. Deswegen sollen im folgenden kurz wichtige Arbeiten zu den drei Bereichen vorgestellt werden. Für eine allgemeine Übersicht sei hier auf zusammenfassende Bücher verwiesen, die prinzipielle Ansätze und eine breite Palette von Darstellungstechniken vorstellen (Card u. a. 1999; Ware 2000; Schumann u. Müller 2000; Spence 2001).

Visualisierungstechniken lassen sich nach den drei Elementen des Informationsmodells von Kreuseler u. Schumann (2002a) einteilen (vgl. S. 11).

Visualisierung der Elemente von Informationsbeständen. Hierbei werden Eigenschaften einzelner Informationsträger angezeigt, um ihre Identifikation und Analyse zu unterstützen oder noch nicht bekannte Eigenschaften aufzudecken (z.B. Ähnlichkeiten). Insbesondere werden zur Positionierung der Informationsträger Anordnungsalgorithmen eingesetzt, welche die Informationsobjekte aus dem n-dimensionalen Merkmalsraum in einen 2- oder 3-dimensionalen Raum abbilden. Diese Abbildung kann durch den Einsatz von

- Faktoren- und Hauptkomponentenanalyse (z.B. Müller u. a. 2006),
- Multidimensional Scaling (z.B. Kruskal u. Wish 1978; Cox u. Cox 1994),
- Federmodellen (z.B. Pfefferer 1996; Gross u. a. 1997; Theisel u. Kreuseler 1998) oder
- neuronalen Netzen (z.B. Kohonen 1997; Wright 1995; Kreuseler u. a. 2000)

erfolgen (vgl. hierzu auch Abs. 3.1.2). Neben solchen automatischen Positionierungsverfahren können auch die spezifischen Eigenschaften einer gegebenen Informationsmenge ausnutzt werden, um die Informationsobjekte darzustellen (z.B. Wise u. a. 1995; Stasko u. a. 1997; Hearst 1999).

Visualisierung der Strukturierung von Informationsbeständen. Ziel hierbei ist es, komplexe strukturelle Zusammenhänge zu veranschaulichen, um Beziehungen zwischen einzelnen Informationsträgern darzustellen sowie Navigation und Orientierung zu unterstützen. Bekannte Beispiele für Hierarchiedarstellungen sind Lamping u. a. (1995); Robertson u. a. (1991); Shneiderman (1992); van Wijk u. van de Wetering (2000); Stasko u. Zhang (2000); Kreuseler u. Schuman (1999); Barlow u. Neville (2001); Yang u. a. (2002); Nguyen u. Huang (2002); van Ham u. van Wijk (2002); Granitzer u. a. (2004); Balzer u. Deussen (2005); Zhao u. a. (2005). Für Netzwerke sind beispielhaft Tollis u. a. (1999); Abello u. a. (2001); Brandes u. Corman (2002); van Ham u. van Wijk (2004); Heer u. Boyd (2005); Dwyer u. Koren (2005) zu nennen.

Visualisierung von Werten- und Werteverteilungen. Hierbei werden die Datenwerte eines Informationsbestandes veranschaulicht. Der Schwerpunkt liegt dabei darauf, implizite Strukturen in den Datenwerten aufzudecken. Hierzu gehören Panel Matrizen, Linienzugdarstellungen (z.B. Parallele Koordinaten in Artero u. a. (2004); Johansson u. a. (2005)), ikonenbasierte Techniken und pixelbasierte Techniken (für eine Übersicht vgl. z.B. Schumann u. Müller 2000; Ware 2000; Spence 2001; Keim u. a. 2002; Kosara u. Hauser 2002).

Neben den hier aufgezählten Ansätzen mit dem Fokus auf abstrakten Daten gibt es auch spezielle Techniken für Daten im räumlichen und zeitlichen Kontext, welche die Spezifik der Visualisierung von Werten und Werteverteilungen in Raum und Zeit beachten.

Bei der Darstellung *zeitveränderlicher Daten* ist die Zielstellung, den Parameter „Zeit“ gesondert auszuweisen, um Veränderungen, Trends und Extremepisoden zu veranschaulichen. Dabei müssen die verschiedenen Ausprägungen des Parameters Zeit – z.B. auf Basis der Taxonomie von Frank (1998) – berücksichtigt werden. Von Müller u. Schumann (2003) wird zwischen statischen, dynamischen und event-basierten Techniken zur Repräsentation zeitlicher Zusammenhänge unterschieden.

Statische Techniken erzeugen ein Bild, wobei die Zeit durch eine ausgewiesene Achse repräsentiert wird. Hierzu gehören konventionelle Techniken wie Zeitdiagramme genauso wie spezielle Techniken für multivariate Daten (z. B. van Wijk u. van Selow 1999; Havre u. a. 2002a; Tominski u. a. 2004). Dynamische Techniken nutzen Animationen oder Slideshows und event-basierte Techniken stellen nicht die Datenwerte, sondern Ereignisse von Interesse zu bestimmten Zeitpunkten dar (Tominski u. Schumann 2004; Brandes u. a. 2005).

Weiterhin wurde eine Vielzahl von Visualisierungstechniken für die visuelle Analyse von Werten und Werteverteilungen auf *Daten mit räumlichem Bezug* entwickelt. Diese können nach ihrem Datentyp und dem zugrunde liegenden Gittern verschiedenen Datenklassen zugeordnet werden. Dies schließt vor allem die Darstellung von

- Multiparameterdaten,
- Volumendaten (z.B. Brodlie u. Wood 2001; Engel 2002; Silver u. a. 2004),
- Srömungsdaten (z.B. Post u. a. 2003; Laramee u. a. 2004),
- gestreuten Daten (engl.: scattered data, z.B. Treinish 1994)

ein, welche im Umfeld der Darstellung von Klimadaten relevant sind. Eine allgemeine Übersicht zur Visualisierung von Daten mit räumlichem Bezug geben Schumann u. Müller (2000). Darüber hinaus sind für diese Arbeit auch Darstellungen von Daten auf Karten relevant (vgl. z.B. Hake u. Grünreich 1994; McEachren 1994, 1995; Jung 1996; Dent 1999; Tominski u. a. 2003), wie sie z.B. in GIS-Systemen[2] erfolgen (vgl. z.B. Bill 1991; Frank 1998).

Im Umfeld der Klimadatenanalyse sind insbesondere Ansätze zur *Visualisierung von räumlichen Daten auf variierenden Gittern* von Interesse, da häufig Klimamodelle und Klimadaten in variierendem räumlichen und zeitlichen Bezug berechnet bzw. erhoben werden. Zur Validierung von Klimamodellen ist der Vergleich von Klimadaten in Raum und Zeit eine wichtige Aufgabe. Die vergleichende Visualisierung von Daten auf variierenden Gittern ist bisher noch weitgehend offener Forschungsgegenstand (vgl. hierzu auch Kapitel 5.4).

Eine Herausforderung bei der Visualisierung großer Informationsmengen – wie sie auch in der Klimaforschung auftreten – besteht in der begrenzten **Darstellungskapazität** der Ausgabegeräte. Deshalb wurden spezielle **Präsentationstechniken und Interaktionstechniken** entwickelt, um durch Nutzerinteraktion – im Sinne von Shneiderman's Mantra „Overview, Zoom and Filter, Details-On-Demand“ (Shneiderman 1996) – die Datenflut beherrschen zu können. Diese Techniken werden jedoch im Bereich der Visualisierung von Klimadaten bisher kaum eingesetzt. Weiterhin für die visuelle Analyse von Klimadaten und -modellen sind Methoden zur **Darstellung von Unsicherheiten und Features** relevant, da die resultierenden Daten aus Berechnungs- und Simulationsprozessen mit starken Unsicherheiten behaftet sein können sowie die Identifikation von Mustern (wie z.B. Fronten oder Wirbeln) wichtige Erkenntnisse über die zugrunde liegenden Phänomene geben können. Relevante Methoden werden im folgenden kurz vorgestellt.

Information-Hiding kann als grundlegendes Vorgehen zur Reduktion der Datenmenge auf eine darstellbare Größe angesehen werden. Bei Filtering, Mapping und Rendering können Informationen oder deren Repräsentationen „versteckt“ werden.

Fokus & Kontext. Ein wichtiges Beispiel in diesem Zusammenhang sind Fokus & Kontext - Techniken, die bestimmte Bereiche von Interesse (Fokus) in einem hohen Detaillierungsgrad und weniger relevante Bereiche (Kontext) in einem geringeren Detaillierungsgrad wiedergeben (vgl. Keahey 2000; Böttger u. a. 2006). Bekannte Beispiele für Fokus & Kontext-Techniken sind Fisheye Views (Furnas 1986), Perpective Wall (Mackinlay u. a. 1991) und der Hyperbolic Viewer (Lamping u. a. 1996).

[2] Geographische Informationssysteme

Overview & Detail kann als eine spezielle Form des Fokus & Kontext - Konzepts aufgefasst werden. Dabei wird eine Überblicksdarstellung der gesamten Datenmenge mit einer Detailansicht gekoppelt. Der Überblick kann in einem separaten oder auch im gleichen Fenster (vgl. z.B. Kreuseler u. Schumann 2002b) dargestellt werden. Im Sinne einer verbesserten Orientierung werden dabei häufig in der Überblicksdarstellung die Lage der Daten aus der Detailansicht hervorgehoben.

Interactive Filtering bedeutet, dass der Nutzer direkt in der Visualisierungstechnik Daten von Interesse herausfiltert bzw. selektiert. Beispiele hierfür sind Linsentechniken (Bier u. a. 1993; Rao u. Card 1994; Kreuseler u. Schumann 2002b) oder dynamische Anfragen (engl. dynamic queries, siehe z.B. Ahlberg u. Shneiderman 1994; Eik 1994).

Brushing & Linking ist eine oft angewandte Interaktionstechnik (vgl. z.B. Unwin u. a. 1990). Beim Brushing kann der Nutzer Daten selektieren, die visuell hervorgehoben werden. Dies wird oft mit dem Linking verknüpft, indem die selektierten Daten auch in anderen Darstellungen hervorgehoben oder anderweitig verändert dargestellt werden. Brushing & Linking wird u.a. in Scatterplot-Matritzen und Parallelen Koordinaten-Darstellungen eingesetzt und ist integraler Bestandteil heutiger Informationsvisualisierungssysteme.

Navigationstechniken werden überwiegend dazu eingesetzt, um die Projektion der Daten auf den Bildschirm zu modifizieren, und dienen so der Navigation im Informationsraum. Zu den Navigationstechniken zählt u.a. das *Zoomen*, welches neben der Veränderung der Auflösung auch ein verändertes Mapping – z.B. mehr Details auf einer höheren Zoom-Stufe – beinhalten kann.

Die hier aufgelisteten Techniken spielen neben der Visualisierung auch bei der Interaktion mit automatischen Mining-Methoden und deren Ergebnissen eine Rolle. So kann bsw. die Selektion eines Clusters, welcher als Überblicksdarstellung aufgefasst werden kann, zur Darstellung aller Objekte des Clusters oder zur Berechnung und Darstellung von Clustereigenschaften führen (vgl. z.B. Nocke u. a. 2004).

Weiterhin von Interesse im Umfeld dieser Arbeit sind die Ansätze zur Feature-Visualisierung und zur Visualisierung von Fehlern und Unsicherheiten. Diese werden im folgenden kurz umrissen.

Feature-Visualisierung. Hierbei geht es darum, je nach Anwendungsfall unterschiedliche Eigenschaften der Datenmenge (Features) abzuleiten und diese in Kombination mit oder an Stelle von den Ausgangsdaten zu visualisieren (z.B. Silver 1997; Kao u. Shen 1999; Reinders u. a. 2001; Reinders 2001; Doleisch u. a. 2003a; Post u. a. 2003). Insbesondere steht bei der Feature-Visualisierung die interaktive Definition und Verfolgung der Features im Vordergrund. Hierbei stellt die Berechnung abgeleiteter Informationen eine Reduzierung der ursprünglichen Datenmenge dar, die nicht nur für eine platzsparende Visualisierung Vorteile bringt, sondern auch für eine effektive Aufbewahrung und Weiterverwendung der Daten.

Visualisierung von Fehlern und Unsicherheiten. Die Visualisierung von Unsicherheiten und Fehlern fokussiert auf die gleichzeitige Präsentation der Daten und der Datenqualität in deren räumlichen und zeitlichen Bezug, was gerade auch die visuelle Veranschaulichung von Simulations- und Messungenauigkeiten einschließt. Ansätze zur Darstellung von Unsicherheiten sind (z.B. in Lodha u. a. 1996; Pang u. a. 1997; Cedilnik u. Rheingans 2000; Djurcilov u. a. 2001; Doleisch u. Hauser 2002) beschrieben. So kann bsw. das „Two-Level Volume Rendering“ (Hauser u. a. 2001) in diesem Kontext verwendet werden um verschiedene Visualisierungsmethoden für unterschiedliche Objekte in der Volumen- und Strömungsvisualisierung darzustellen. Auf diese Art kann die Aufmerksamkeit des Betrachters gezielt gelenkt werden, und es lassen sich auch Unsicherheiten durch unterschiedliche Renderingstile kodieren (vgl. auch Strothotte u. a. 1999). Eine Übersicht zur Visualisierung von Unsicherheiten und Fehlern findet sich in Griethe u. Schumann (2005).

3.1.2 Data Mining

Witten u. Frank (2000) definieren Data Mining als

„the extraction of implicit, previously unknown, and potentially useful information from data."

Alternativ zum Begriff Data Mining wird auch der Begriff „Knowledge Discovery in Data Bases" (KDD) verwendet (vgl. Fayyad u. a. 2001), der jedoch neben dem eigentlichen Analyseschritt die ganze Kette von Datenaufbereitung bis zur visuellen Auswertung einschließt.

Data Mining Verfahren lassen sich nach den Gebieten, in denen sie entwickelt wurden, einteilen. Han u. Kamber (2000) zählen hierzu Datenbanktechnologien, Statistik, maschinelles Lernen, Informationswissenschaften, Visualisierung und weitere Ansätze. Im Umfeld dieser Arbeit sind vor allem Methoden der Statistik von Interesse, da diese in der Klimaforschung häufig eingesetzt werden. Im speziellen sind dies

- Clusteranalyse: Ziel der Clusteranalyse ist es, ähnliche Datenobjekte zu Clustern zusammenzufassen, während unähnliche Datenobjekte in verschiedene Cluster einsortiert werden. Die sich ergebenden Gruppen von Objekten sollen dabei möglichst homogene Eigenschaften haben. Übersichten über Clusterverfahren finden sich z.B. in Bock (1974); Han u. Kamber (2000).
- Hauptkomponenten- und Faktorenanalyse: Hierbei werden mittels Neuausrichtung der Koordinatenachsen die Trends bzw. Faktoren in den Daten erfasst, gewichtet und nach deren Relevanz sortiert. Basierend hierauf können die Variablen diesen Trends zugeordnet werden, wodurch sich grundlegende Aussagen über Variablenabhängigkeiten ergeben (vgl. z.B. Joliffe 1986).

Weitere in Data Mining-Systemen eingesetzte Methoden sind

- Assoziationsanalyse: Hierbei werden Regeln basierend auf den Kardinalitäten der Wertebereiche der beteiligten Variablen aufgestellt (vgl. z.B. Bollinger 1996). Typische Anwendung hierfür ist die Warenkorbanalyse, bei der Regeln aufgestellt werden wie z.B. „Wird Mehl gekauft, so wird in 30% der Fälle auch Milch gekauft". Neben der Häufigkeit der Übereinstimmung einer Regel (hier 45%), die auch als Konfidenz bezeichnet wird, ist dabei weiterhin auch noch der Support wichtig, der angibt, wie häufig überhaupt Milch und Mehl gekauft werden, was der Relevanz der Regel entspricht.
- Klassifikation: Bei der Klassifikation liegt das konzeptuelle Modell zur Einteilung der Objekte (im Gegensatz zur Clusteranalyse) bereits vor. Das Klassenmodell wird gewöhnlich aufgrund einer vorklassifizierten Trainingsmenge vorgegeben. Aufgabe der Klassifikation ist es, neue Datenobjekte den vorgegebenen Klassen zuzuordnen. Hierzu werden z.B. Entscheidungsbäume oder neuronale Netze eingesetzt.
- Regressionsanalyse: Hierbei werden Modelle zur Vorhersage zu erwartender Variablenwerte aufgrund funktionaler Abhängigkeiten entworfen. Zum Beispiel werden bei der linearen Regression lineare Abhängigkeiten einer Menge von Variablen ausgedrückt.
- u.a.

In diesen Verfahren als auch separat von ihnen werden im Rahmen des Data Mining eine Vielzahl statistischer Maße und Kenngrößen verwendet, welche wichtige Charakteristika der Daten quantitativ beschreibbar machen. Hierzu zählen unter anderem Ähnlichkeits-, Distanz- und Korrelationsmaße sowie Standardmomente wie Mittelwerte und Varianz. Mit diesen Verfahren und Maßen wird eine neue, aggregierte Sicht auf die Daten erzeugt, welche die zum Teil sehr großen Datenmengen erst handhabbar machen.

Zusammenfassende Abhandlungen zum Thema Data Mining finden sich z.B. in Westphal u. Blaxton (1998); Han u. Kamber (2000); Witten u. Frank (2000).

3.2 Visualisierung und Datenanalyse in der Klimaforschung

In diesem Abschnitt soll nach einer kurzen Einleitung zum Einsatz der Statistik und analytischer Ansätze in der Klimaforschung (vgl. Abs. 3.2.1) im Hauptteil der Stand der Forschung zur Visualisierung von Klimadaten ausgeführt werden (vgl. Abs. 3.2.2). Abschließend werden wichtige Aufgaben zum VDM von Klimadaten für diese Arbeit zusammengefasst (vgl. Abs. 3.2.4).

3.2.1 Statistik und analytische Ansätze

Häufig übersteigen heutige Wetter- und Klimadaten die in einem Rechner verarbeit- und direkt visualisierbare Größe. Um diese riesigen Datenmengen im Bereich der Klimaforschung verarbeiten zu können, müssen relevante Daten ausgewählt und Informationen zur Beantwortung bestimmter Fragestellungen unter Einbeziehung von Kontextwissen extrahiert und/ oder aggregiert werden. Hierzu werden in diesem Umfeld vor allem statistische Methoden und analytische Operatoren verwendet.

Beispiele für **analytische Operatoren** sind z.B. aus der Strömungslehre bekannte Operatoren zur Bestimmung der Rotation oder der Verwirbelung eines Strömungsfeldes sowie für skalare Daten der Gradient der Daten (vgl. z.B. Frühauf 1997). Sie dienen der Charakterisierung wichtiger Dateneigenschaften, und können neben der direkten Darstellung auch zur Parametrisierung von Visualisierungen verwendet werden (vgl. auch Abs. 7.1).

Die **Statistik** findet im Bereich der Klimaforschung ein breites Anwendungsfeld. Standardmomente ermöglichen z.B. durch Mittelwert- und Varianzberechnung die Abstraktion des räumlichen Verhaltens eines Klimamodells auf dessen mittlere, zeitliche Merkmalskurven, was einen schnellen Überblick über das grundlegende Verhalten einer Wetter- oder Klimasimulation ermöglicht. Des Weiteren werden insbesondere Clusteralgorithmen verwendet, um in Klimadaten mit vielen abhängigen Variablen grundlegende Strukturen extrahieren sowie diese darstellen zu können (vgl. z.B. Böhm 1999; Gerstengarbe u. Werner 1999; Kücken u. a. 1999). Die Verwendung von Clusteralgorithmen ermöglicht auch, mehrere Merkmale in verschiedenen Modellen, Modellläufen und/oder Messstationsdaten miteinander zu vergleichen.

Auswahl und Parametrisierung von Verfahren sind auch in diesem Umfeld ein sensibler Prozess, der durchaus sehr fehlerträchtig sein kann. Ein Beispiel für zwei stark variierende Resultate findet sich z.B. zwischen Mann u. a. (1998) und McKitrick u. McIntyre (2005). Mann u. a. (1998) untersuchen mittels statistischer Verfahren die Erderwärmung seit dem 14. Jahrhundert, und identifizieren eine bis zur zweiten Hälfte des 20. Jahrhundert leicht ansteigende, und ab dort aber rasant ansteigende Erwärmung (Hockeystick-Kurve). McKitrick u. McIntyre (2005) kommen bei der Untersuchung derselben Daten unter Einsatz anderer statistischer Filter zu einem abweichenden Ergebnis, denn sie finden neben der Erwärmung in den letzten 50 Jahren auch noch ein Temperaturmaximum im 15. Jahrhundert. Zur Vermeidung solcher Unstimmigkeiten wird im VDM der Weg einer starken Kopplung von Visualisierung und Statistik beschritten.

Ein Beispiel für die Unterstützung der Parametrisierung statistischer Verfahren findet sich bei Kücken u. a. (1999), wo hierzu eine graphische Nutzeroberfläche zur Parametrisierung von Simulationsläufen und Clusterverfahren entworfen wird.

3.2.2 Visualisierung von Klimadaten

In diesem Abschnitt soll der aktuelle Stand der Forschung bei der Visualisierung von Klimadaten wiedergegeben werden. Zuerst werden Klassifikationen im Umfeld von Wetter- und Klimavisualisierungen vorgestellt (vgl Abs. 3.2.2.1). Danach sollen allgemeine Konventionen in diesem Umfeld untersucht werden (vgl Abs. 3.2.2.2). Im anschließenden Abschnitt 3.2.2.3 werden für verschiedene

Arten von Darstellungen der Stand der Forschung zur Visualisierung von Wetter- und Klimadaten wiedergegeben.

3.2.2.1 Klassifikation von Visualisierungen

Schröder (1997) bezieht sich zur Klassifikation von Visualisierungstechniken im Umfeld meteorologischer Daten auf Earnshaw u. Wiseman (1992). Dort werden Visualisierungstechniken nach der **Dimensionalität des Datenraumes** und der **Dimensionalität der graphischen Repräsentation**[3] klassifiziert (vgl. Abb. 3.2).

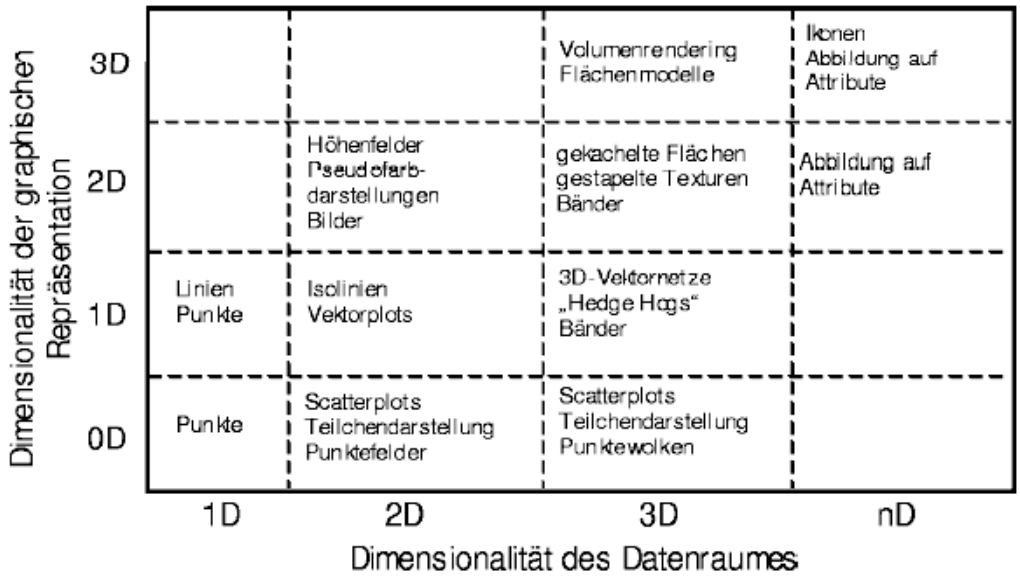


Abbildung 3.2: Klassifikation von Visualisierungen im meteorologischen Umfeld von Earnshaw u. Wiseman (1992) aus Schröder (1997)

Treinish (1999) klassifiziert abweichend davon Visualsierungstechniken für meteorologische Daten nach der **Dimensionalität des Darstellungsraumes** (2D, ...) sowie nach den **Zielen** (Analyse, Präsentation, Entscheidungshilfe) der Visualisierung in fünf Klassen:

1. **2D Visualisierungen mit geringem Interaktionsgrad**: dabei werden wenige Merkmale dargestellt, die zumeist als stetige oder konturierte Farbdarstellungen, Isolinien und einfache Ikonen für die Stromvektordarstellung wie Pfeile oder „Barbs" (Widerhaken) abgebildet werden.
2. **gekoppelte 2D und 2,5D Visualisierungen**: es handelt sich dabei um eine Oberklasse von Klasse 1 mit 3D Erweiterungen. Die Darstellungen sind sowohl für die Analyse als auch zur Präsentation geeignet, und erlauben insbesondere den Vergleich von (bis zu fünf) Merkmalen. Typische visuelle Attribute sind Farbe, Höhe und Pfeildarstellungen.
3. **Visualisierung zum 3D-Browsing**: hierbei liegt der Fokus auf Oberflächenbedingungen und Niederschlagsbedingungen zur Wettervorhersage, wobei dem Nutzer eine Vielzahl von Interaktiuonsmöglichkeiten an die Hand gegeben werden. Auch diese Darstellungen sind für Analyse und Präsentation geeignet. Typische Darstellungen sind Höhenkarten mit Oberflächenströmungsvisualisierung und 3D-Wolkendarstellungen.
4. **Visualisierung zur 3D-Analyse**: es handelt sich dabei um komplexe, interaktive 3D-Darstellungen mit Achsenbeschriftungen ohne Höhenkarte. Es können kombiniert Isoflächen,

[3] der eingesetzten Primitive

DVR[4], 3D-Ikonen (z.B. Pfeile) sowie beliebige Schnittdarstellungen eingesetzt werden.

5. **Entscheidungsunterstützende Datenpräsentation**: hierbei werden 2D-Darstellungen eingesetzt, die mit beliebigen Daten aus anderen Quellen (z.B. typische GIS-, Land- oder forstwirtschaftliche Daten) gekoppelt werden, wobei nur wesentliche Wetterinformationen dargestellt werden, um die Repräsentation nicht zu überladen.

Wesentliche Kategorie dieser Klassifikation sind die Eignung gewisser Darstellungsarten für Analyse vs. Präsentation in Abhängigkeit vom Komplexitäts- und Interaktionsgrad der Präsentationen. Treinish (1999), aber auch Gelin (2002) schlagen vor, mehrere Techniken aus verschiedenen dieser Klassen je nach Stand des Analyseprozesses zu benutzen.

Die Grenzen zwischen den Klassen 1 und 2 sowie den Klassen 3 und 4 sind dabei fließend. Klasse 1 kann als Standard für alle klimaspezifischen Visualisierungssysteme angesehen werden, und die meisten Visualisierungen in Veröffentlichungen aus Klimaforschung und Meteorologie kommen aus dieser Klasse. Nach Treinish (1999) hat sich besonders die dritte Klasse im Gegensatz zu den Klassen 1 und 2 als nutzbringend erwiesen, da eine animierte Darstellung von Bildern dieser Klasse Rückschlüsse zulässt, zu denen eine ganze Reihe von 2D-Bildern notwendig wären. Einige wenige Systeme (z.B. Vis5D) sind speziell auf die Klasse 4 zugeschnitten. Im Sinne der Kommunikation von Klimafolgen sind die Methoden der 5. Klasse von besonderem Interesse.

Nicht-räumliche Darstellungen, wie sie z.B. von Macêdo u. a. (2000) oder Doleisch u. a. (2004) in diesem Kontext zum Brushing im Merkmalsraum eingesetzt werden, fehlen in dieser Klassifikation. Ein systematischer Einsatz multivariater Techniken im Klimaumfeld steht noch aus, und erfolgt im Rahmen dieser Arbeit (vgl. Kap. 5).

Die vorgestellten Klassifikationen fokussieren auf die Darstellung von Wetterdaten, sind relativ allgemein und abstrahieren über viele Datencharakteristika, Aufgaben und Darstellungseigenschaften, beschreiben diese jedoch implizit mit. Klimadarstellungen können weiterhin auch nach dem **Datentyp** (Skalar, Vektor, Tensor), der **Anzahl und Art der dargestellten Merkmale** (uni-, bi- und multivariat sowie Temperatur, Druck, Feuchtigkeit ...), der **Art der Darstellung** (statisch, dynamisch) und **Zielstellungen** (z.B. explorieren, Hypothesen validieren, präsentieren sowie Entscheidungen ableiten) unterschieden werden. Ferner im Kontext der Klimaforschung von Interesse ist die **Art des untersuchten Phänomens** (Ozean, Atmosphäre und Phänomene der Erdoberfläche wie Wetterbedingungen, Biosphäre oder Vereisung).

3.2.2.2 Konventionen bei der Visualisierung von Klimadaten

Die Einhaltung von Konventionen beim Design von Visualisierungen erleichtert es den Nutzern, diese intuitiv und schnell zu erfassen (vgl. auch Abs. 3.4.2 und Kap. 7). Im folgenden sollen deswegen in der Literatur vorhandene Regeln im Bereich der Meteorologie und Klimaforschung kurz zusammengetragen werden. Hierbei haben insbesondere die Richtlinien von Treinish (1999) noch immer Gültigkeit.

Eine Verallgemeinerung von Richtlinien zur **Verwendung von Farbe** in meteorologischen Darstellungen wird von der Amerikanischen Wettergesellschaft (AMS) gegeben (American Meteorological Society 1993). Diese umfassen im speziellen eine Palette von Farben und deren Entsprechung im RGB-Raum für punkthafte (z.B. Stürme in Rot oder Regen in grün), linienhafte (z.B. Kaltfronten in blau und Warmfronten in Rot) oder flächenhafte Objekte (z.B. Nebel in gelb oder Lufttemperatur in einer Regenbogenskala) sowie für Kartenhintergründe (z.B. Terrain in braun). Diese Farbwahl hat sich in einem langen Zeitraum der Wetterkartographie herausgebildet und wird als Empfehlung oder Standardeinstellung bei der Darstellung von Wetterphänomenen betrachtet. Weiterhin werden

[4]Direct volume rendering: direkte Darstellung des Datengitters aufgrund der Projektion der Gitterzellen in das Bild oder durch Konstruktion von Strahlen vom Bild aus.

von der American Meteorological Society (1993) allgemeine Regeln zur Farbwahl in den Kontext der meteorologischen Visualisierung gestellt. So wird u.a. empfohlen, bekannte Farbassoziationen auszunutzen, Farben konsistent zu verwenden, die Anzahl verschiedener Farben zu begrenzen und Kontrast und wahrgenommene Helligkeit bei der Farbwahl zu berücksichtigen.

Weitere Abbildungsregeln für Wetterdaten finden sich bei Treinish (1999) und Baker u. Bushell (1995). Baker u. Bushell (1995) empfehlen, im meteorologischen Umfeld **„naturnahe" Farben** einzusetzen, um die Verständlichkeit der Darstellung zu erhöhen sowie Farbskalen entsprechend den Analysezielen einzusetzen. *Verrauschte Daten* wie z.B. Windgeschwindigkeiten werden nach Treinish (1999) üblicherweise auf die Helligkeit abgebildet, während nur *leicht variierende Merkmale* wie Temperatur auf die Farbsättigung abgebildet werden sollten. Für *feuchtigkeitsbezogene Merkmale* (z.B. Luftfeuchtigkeit oder Niederschlag) empfiehlt Treinish (1999) eine zweigeteilte Farbskala, von braun für trockene Regionen über gelb und grün für geringen Niederschlag bis zu einem hellen Blau für hohe Niederschlagswerte (vgl. hierzu auch Treinish 1994). Beim Einsatz von Farbbändern (conturing) wird eine segmentierte Farbskala mit wahrnehmbarer Ordnung auf den Farben empfohlen. Für 3D-Flächen wie Isoflächen (z.B. Wolkenoberflächen) sind uniforme Komplementärfarben zu den sonst genutzten Farben üblich, um die Überlagerung von Farbtönen zu minimieren. Baker u. Bushell (1995) empfehlen, bei Niederschlags- und Feuchtigkeitsdaten realitätsnahe, wolkenähnliche Farbskalen zu verwenden. Für DVR empfiehlt Treinish (1999) weiterhin, einen konstanten Farbton unter Variation der Helligkeit und der Transparenz einzusetzen.

Für **Oberflächen-Windvektor-Darstellungen** sind nach Treinish (1999) verschiedene Varianten üblich: für den Betrag der Vektoren werden Pseudo-Farbdarstellungen und für deren Richtung Pfeile fester Größe verwendet, um den Eindruck von Bewegungen innerhalb des Feldes bei Animationen, und um die Probleme bei der Pfeilskalierung zu vermeiden. Weiterhin werden Stromlinien mit Richtungspfeilen versehen, um Fronten, Konvergenzzonen, Wirbel u.a. darzustellen. Für Nichtmeteorologen sind nach Treinish auch wehende Fahnen üblich.

Ein weiterer, meist integraler Bestandteil von Wetter- und Klimadatendarstellungen sind **Achsen und Achsenbeschriftungen** (vgl. Baker u. Bushell 1995). Sie können z.B. eingesetzt werden, um variierende Skalierungen der Achsen zu identifizieren. Typisch in klimatischen Darstellungen ist ferner die Darstellung von Gewässern, von Gebiets- oder Ländergrenzen sowie die Darstellung der Orographie (vgl. z.B. Schröder 1997; Treinish 1999), um dem Anwender neben der örtlichen Orientierung auch Informationen über mögliche Einflüsse auf das Klima und die Ausdehnung von klimatischen Phänomenen zu geben. Für präsentative Darstellungen von Wetterphänomenen für ein weites Publikum empfehlen Baker u. Bushell (1995) sogar, zusätzliche visuelle Hinweise mit bekannten Objekten oder Entfernungen in die Darstellungen zu integrieren, um ein intuitives Verständnis für die Ausbreitung des Phänomens zu bekommen.

Darüber hinaus ist in diesem Umfeld auch die Art der **Projektion der Erdkugel in eine 2D-Darstellung** von Bedeutung. So stellen z.B. Kottek u. Rubel (2003) verschiedene im Klimaumfeld typische Projektionsverfahren und deren Vor- und Nachteile vor.

3.2.2.3 Methoden

Im folgenden sollen spezielle Arbeiten zur Visualisierung von Klimadaten vorgestellt werden. In Anlehnung an die Einteilungen von Schröder (1997) und Treinish (1999) sollen diese nach **Datenklasse**, **Gittertyp** und **Dimensionalität** eingeteilt werden. Diese Einteilung ermöglicht es, für spezielle Datenklassen zugeschnittene Techniken aus dem Visualisierungsumfeld (für beliebige Daten) leicht mit den vorgestellten Ansätzen im Klimaumfeld abgleichen zu können. Es werden dabei Aspekte zu den hier bereits aufgelisteten Konventionen datenklassenspezifisch ergänzt, und darüber hinaus auch alternative, für das Klimaumfeld untypische Darstellungsarten von Klimadaten ausgeführt. Dabei fließen sowohl Arbeiten von Klimaforschern als auch Arbeiten aus dem Visualisierungsumfeld ein.

Darstellungstechniken für skalare, gestreute 2D-Klimadaten. Von Treinish (1994) werden zur Analyse von starken Regenfällen in El-Niño-Jahren im Nordwestlichen Peru auf Basis von gestreuten Messstationsdaten - mit Hilfe des Visualisierungssystems OpenDX entwickelte - Visualisierungstechniken vorgestellt. Dazu vergleicht der Autor verschiedene Varianten zur Abbildung der gestreuten Daten auf ein regelmäßiges Gitter sowie die Abbildung mittels der Delauney Triangulation. Die Expressivität dieser Verfahren werden bezüglich Topologieerhaltung und Bildqualität untersucht, wobei auch auf die Behandlung von Fehlwerten eingegangen wird. Zum visuellen Vergleich zweier Merkmale werden verschiedene bivariate Darstellungen wie kombinierte Farb- und Isolinien-, Isoflächen- und Isolinien- sowie Farb- und Höhenfelddarstellungen vorgestellt.

Im Unterschied zu solchen interpolierenden Darstellungen mit Hauptaugenmerk auf der räumlichen Verteilung von gestreuten Daten werden im Klimaumfeld teilweise auch Ikonen eingesetzt. So verwenden Stier u. a. (2005) ikonifizierte Balkendarstellungen (4 Merkmale) oder Kreisikonen (2 Merkmale) zur vergleichenden Repräsentation von Aerosolen. Außerdem werden auch Ikonen zur Repräsentation bestimmter Wetterlagen sowie abgeleiteter Eigenschaften wie Wolkenhöhe oder Sichtverhälnisse verwendet (vgl. z.B. Scanlon 1994; Spirkovska u. Lodha 2002). Weiterhin verwendet Wrobel (2004) für die Präsentation des Typs verschiedener Messstationen im Klimaumfeld auch in der Kartographie übliche Kreis, Dreiecks- und Quadratikonen mit unterschiedlichen Farben. In der Arbeit von Saito u. a. (2005) werden darüber auch ikonifizierte, farbkodierte Zeitdiagramme zur Darstellung des Tagestemperaturverlaufes gestreuter Messstationen eingesetzt.

Darstellungstechniken für skalare, 2D-Klimadaten auf regelmäßigen Gittern. 2D-Darstellungen von 2D-Klimadaten unter Abbildung auf Farbe und/oder Isolinen sind Standardtechniken im Bereich der Meteorologie und Klimaforschung (Klasse 1 nach Treinish (1999)), und werden am häufigsten eingesetzt (vgl. z.B. Treinish u. a. 2003; Trafton u. a. 2002; Kottek u. Rubel 2003; Brockmann 2004; Stier u. a. 2005). Anzumerken ist hier, dass eine automatisch generierte, geeignete Beschriftung von Isolinien in solchen Darstellungen noch immer nicht vollständig gelöst ist, da es bei vielen, nah beieinander liegenden Isolinien leicht zu Überlappungen kommen kann (vgl. z.B. Schmidt u. a. 2004).

Weiterhin findet man solche Farb- und Isoliniendarstellungen häufig mit Höhenfelddarstellungen (Klasse 2) oder Pfeil, Barb- und/oder Stromliniendarstellungen für 2D-Strömungsdaten angereichert (vgl. S. 21). Darüber hinaus werden häufig auch sphärische Darstellungen der Erdkugel (meist unter Farbmapping) verwendet, insbesondere um die Klimaverhältnisse um die Pole adäquat wiederzugeben (vgl. z.B. Chen 1993; Brockmann 2004).

Darstellungstechniken für skalare, 3D-Klimadaten auf regelmäßigen Gittern. Zur Darstellung 3D-meteorologischer Daten werden häufig *Isoflächendarstellungen* eingesetzt. So werden z.B. von Frühauf (1997) und von Treinish (1994) Isoflächen zur Abbildung von Gebieten hoher Strömungsgeschwindigkeit bzw. Luftdrucks verwendet, die Anhand der lokalen Temperatur farbkodiert werden.

Baker u. Bushell (1995) analysieren die Möglichkeiten zur Visualisierung eines simulierten Sturmes mit Isoflächen als Kernelementen. Dabei vergleichen sie eine als Video verfügbare 3D-Visualisierung mit ihrer eigenen erweiterten Version in Bezug auf das Visualisierungsdesign. Wichtig ist ihnen dabei, das die Wahrnehmung auf die meteorologischen Daten konzentriert bleibt. Dazu schlagen sie z.B. vor, den Kontrast bei der farbkodierten Darstellung von Gitterlinien zu verringern, u.a. auch um dadurch auf der Grundfläche der umgebenden Bounding Box zusätzliche Informationen wie z.B. Isolinien kodieren zu können.

Neben Isoflächendarstellungen werden auch *Schnittdarstellungen* (vgl. z.B. Schröder 1997; Treinish

1999) und *DVR-Verfahren* (vgl. z.B. Riley u. a. 2003) zur Darstellung von Klimadaten eingesetzt. So bilden z.B. Ribarsky u. a. (2002a) Radar-Doppler-Daten auf einem radialen, curvilinearen Gitter auf ein reguläres Gitter ab und rendern dieses mittels ellipsoiden Splatting[5].

Einen speziellen Aspekt bei der 3D-Klimavisualisierung stellen *realitätsnahe Darstellungen* dar. Diese haben den Vorteil ihrer leichten Kommunizierbarkeit bei der Präsentation von Wettersituationen und -phänomenen. Überflüge über 3D Landschaften unter Verdeutlichung von Wettergeschehen sind im Rahmen des TV-Wetterberichts ein übliches Werkzeug. Daneben erlauben realitätsnahe Darstellungen auch im Rahmen der Datenanalyse, gewisse Phänomene (wie z.B. Wolkenbildung) und deren Bezug zu geographischen Einflussfaktoren (Terrain, Flussläufe, etc.) besser zu verstehen.

So können Wolkenbildung mit Hilfe von DVR-Verfahren (vgl. z.B. Riley u. a. 2003; Marchesin u. a. 2004), von geeignet schattierten bzw. texturierten Isoflächen (vgl. Baker u. Bushell 1995; Max u. Crawfis 1995) oder von fraktalen Verfahren (vgl. Schröder 1997) realitätsnah abgebildet werden. Neben dem Einsatz zur Exploration von Luftfeuchtigkeit und Niederschlag können solche wolkenähnlichen 3D-Darstellungen auch verwendet werden, um verkleinert als Ikone leicht verständlich in anderen Darstellungen eingebettet zu werden. So verwenden Baker u. Bushell (1995) Wolkenikonen innerhalb eines Zeitgraphen, um Überblick und Orientierung in einem Sturm-Datensatz zu steigern. Eine allgemeine Übersicht über Verfahren zur naturalistischen Modellierung, Animation und Beleuchtung von Wolken sowie zur realitätsnahen Visualisierung des Himmels findet sich in Trembilski (2003).

Neben Wolkendarstellungen verwendet Schröder (1997) für die Darstellung im TV-Bereich transparente, realitätsnahe Texturen zur Repräsentation von Niederschlägen oder Blitzen in 3D-Landschaften. Ferner setzt Schröder (1997) animierte Windsäcke zur Repräsentation von Windgeschwindigkeit und Windrichtung ein.

Ein weiteres Beispiel realitätsnaher Darstellungen - im Sinne der leichten Kommunizierbarkeit auch für Laien - findet sich bei Thurston u. a. (2003). Dort wird in einer 3D-Höhenkarte die Entwicklung des Baumbestandes im Bundesland Brandenburg realitätsnah dargestellt. Zusätzlich werden zur Veranschaulichung verschiedener forstwirtschaftlich und klimatisch relevanter Merkmale auch abstrakte Balken in die Darstellung integriert.

Darstellungstechniken für Strömungsdaten in 2D und 3D. Zur Visualisierung von klimatischen Strömungen ist es üblich, 2D-Pfeil- oder Barbdarstellungen einzusetzen (vgl. z.B. Treinish 1999; Macêdo u. a. 2000; Spirkovska u. Lodha 2002; Treinish u. a. 2003; Kottek u. Rubel 2003). Um die Nachteile von Pfeildarstellungen (z.B. bei der Überlappung von Pfeilen) abzumildern, werden von Macêdo u. a. (2000) zur Entschlackung der Darstellung über Brushing & Linking mit Scatterplots ein Filtering von Strömungsvektoren von Interesse durchgeführt. Um über verschiedene, in den Daten auftretende Windrichtungen selektieren zu können, werden diese kreisförmig in einem Scatterplot dargestellt. Neben der Selektion erlaubt dies dem Nutzer, schnell die Verteilung auftretender Windrichtungen zu erkennen.

Für eine kombinierte Darstellung von Vektor- und Skalardaten schlagen Crowfis u. Max (1992); Crowfis u. a. (2000) ein erweitertes DVR-Visualisierungsverfahren vor. Dieses texturbasierte Verfahren wurde speziell für die kombinierte Darstellung der Wolkenbildung und der Strömungsverhältnisse in einem globalen 3D-Atmosphärenmodell entworfen.

Treinish (1994) stellt für einen atmosphärischen Datensatz über dem Pazifik farbkodierte Stromlinien mit Pfeilen auf der Ober- und Unterseite einer rechteckigen umgebenden Box dar, ohne hier weitere Attribute flächenhaft darzustellen. Dadurch kann der Nutzer die Strömungseigenschaften in

[5]Bei der Projektion eines Datenwertes in die Bildebene wird ein - hier elliptischer - „Fussabdruck“ erzeugt, um Ausdehnung, Lage, Orientierung und Wirkungskreis der Gitterzellen im Bild geeignet wiederzugeben.

diesen Schichten leicht mit der Isofläche des Meeresspiegel-Luftdrucks vergleichen, welche zusätzlich bezüglich der Temperatur farbkodiert ist.

Chen (1993) stellt 3D-Strömungsdaten einer Klimasimulation durch Pfeile und Stromlinien zum einen in einer euklidischen 3D-Quaderdarstellung als auch sphärisch auf der Oberfläche der Erdkugel dar. Das Auftreten von Verdeckungen macht es allerdings schwer, einen guten Überblick über die Eigenschaften des Vektorfeldes zu bekommen.

Griebel u. a. (2004) führen eine neue Methode zur Dekomposition von 2D- und 3D-Oberflächenströmungen ein und demonstrieren sie am Beispiel eines Klimadatensatzes auf der Erdoberfläche. Eingefärbte zylindrische Strömungslinien in flexibel parametrisierbarer Oberflächenabtastung in Kombination mit einer LIC[6]-Darstellung geben einen schnellen, schematischen Eindruck über wichtige Eigenschaften der auftretenden Strömungen.

Doleisch u. a. (2004) stellen im Rahmen der Untersuchung des Hurrikan Isabel Feature-basierte Strömungsvisualisierungen bereit. Dazu werden mittels Feature-Extraktion Bereiche ähnlicher Stromrichtung gruppiert und die entstehenden Gruppen (von Gebieten ähnlichen Strömungsverhaltens) in einem 2D-Schnitt farbkodiert ausgegeben. Dies gibt bereits eine guten Eindruck über verschiedene Stromgebiete, und wird zusätzlich durch die symbolische Darstellung kritischer Punkte angereichert.

Da in Strömungsfeldern in Wettermodellen zwar viele kritische Punkte vorliegen, diese jedoch häufig sporadisch auftreten und nur von begrenztem Interesse sind, entwickeln Wong u. a. (2000) einen Filtermechanismus für kritische Punkte. Dabei bleiben diejenigen kritischen Punkte erhalten, in deren Umgebung besondere Scheerkräfte und Zirkulationen Indikatoren für potentielle Wetterinstabilitäten sein können. Die so extrahierten kritischen Punkte werden dann in einem sehr übersichtlichen 2D-Strombild typabhängig als variierende Kreisikonen, in Kombination mit farbkodiertem Geschwindigkeitsbetrag, Pfeildarstellungen und separierenden Linien dargestellt.

Zusammenfassend kann festgestellt werden, dass im Bereich der Klimaforschung Standard-2D-Strömungsdarstellungen wie Pfeil- oder Isoliniendarstellungen dominieren, und der Einsatz von globalen 2D-Methoden wie LIC, aber auch Stromflächendarstellungen und moderne topologiebasierte Verfahren kaum in diesem Umfeld angewendet werden.

Visualisierung von Klimadaten in ihrem zeitlichen Bezug. Gewöhnlich werden im Klimaumfeld Animationen oder 1D-Kurvendarstellungen[7] eingesetzt, um die Variable Zeit zu kodieren. **Animationen** erlauben insbesondere eine schnelle Erkennung quantitativer Änderungen in den Daten, sind aber weniger geeignet, um einzelne Werte zu erkennen. Zeitdiagramme sind leicht verständlich, haben aber Grenzen, wenn eine größere Anzahl an Merkmalen zum Beispiel auf Korrelationen hin untersucht werden soll. Im Gegensatz dazu haben aktuelle, explizite Darstellungen der Zeit, wie sie aus der Informationsvisualisierung bekannt sind (vgl. z.B. Müller u. Schumann 2003), in diesem Umfeld noch keinen Eingang gefunden.

Ausnahme hiervon ist die Arbeit von Saito u. a. (2005). Hier werden für die Visualisierung von Wetterstationsdaten sogenannte „2-Tone“-Farbskalen verwendet, um bei längeren Zeitskalen sowohl einen guten Überblick als auch einzelne Werte effektiv erkennen und somit mehrere Zeitverläufe leicht vergleichen zu können. Saito u. a. (2005) vergleichen tägliche Temperaturverläufe von über 800 Messstationen gleichzeitig und stellen sehr kompakt die jahrlichen Temperatur-, Niederschlags- und Windrichtungsschwanken einzelner Stationen dar.

Ein weiteres Beispiel für die Untersuchung des Parameters Zeit im Klimaumfeld findet sich bei Baker u. Bushell (1995). Die Autoren untersuchen den Einsatz von *Animation* für einen Sturm-Datensatz. Hierbei weisen sie darauf hin, dass, zum einen aufgrund des häufigen Auftauchens und wieder

[6] Line Integral convolution: Abbildung eines statischen 2D-Strömungsfeldes durch Faltung der Stromlinien mit einem verrrauschten Bild.

[7] bzw. Zeitdiagramme

Verschwindens kleinerer Phänomene in solchen Daten und zum anderen durch die starke Bewegung des Phänomens, die Aufmerksamkeit des Betrachters leicht von den Kernelementen des Sturms abgelenkt werden kann. Weiterhin stellen sie fest, dass ein im Laufe einer Animation springender Maximumbalken auf der Farblegende ungeeignet ist, und schlagen die Darstellung einer zusätzlichen *Kurve* vor, um die mentale Überlastung des Betrachters zu reduzieren. In dieser zeitlichen Kurve werden wichtige Kenngrößen dargestellt, und durch realitätsnahe Wolkenikonen angereichert.

Neben typischen (farbkodierten) Kurvendarstellungen mit mehreren Merkmalen (Temperatur und Luftfeuchtigkeit) fügen Treinish u. a. (2003) an die Darstellung der Windgeschwindigkeitskurve zusätzlich kleine Pfeile ein, um - für einen bestimmten Messpunkt - Windgeschwindigkeiten mit der Windrichtung in ihrem zeitlichen Verlauf vergleichen zu können.

Weiterhin, wie bereits oben erwähnt, schlägt Treinish (1999) bei animierten Pfeildarstellungen vor, zur Vermeidung des visuellen Eindrucks von Bewegungen Pfeile fester Größe und Länge zu verwenden, und Farbdarstellungen zur Kodierung der Strömungsgeschwindigkeit einzusetzen. Um ferner Muster in Animationen von Niederschlagsdaten effektiv verfolgen zu können, untersucht Treinish (1994) speziell hierfür angepasste Farbskalen.

Ross u. a. (1997) verwenden zur Darstellung der Monatsmittelwerte verschiedener Merkmale an der Zeitachse „aufgefädelte“ Rechtecke, wie sie auch in geographischen Abbildungen zur Darstellung der Niederschlagswerte z.B. in Geographielehrbüchern üblich sind.

Auch die Kombination der Zeitachse mit einer räumlichen Achse zur Verfolgung zeitlicher Muster an einem diskreten räumlichen Schnitt ist im Klimaumfeld gebräuchlich. So tragen z.B. Dameris u. a. (2005) Höhe bzw. Breite gegen die Zeit auf, um Muster bei Luftdrücken, Ozonkonzentrationen und Windgeschwindigkeiten untersuchen zu können. Auch Schröder (1997) untersucht den Austausch einer räumlichen Achse durch die Zeitachse und fokussiert dabei auf eine geeignete Nutzerkontrolle bei der Steuerung des Parameters Zeit.

Mulivariate Darstellungen. Typisch sind im Klimaumfeld einfache, meist unverbundene Scatterplots. So verwenden Stier u. a. (2005) eine Reihe solcher Plots, um gemessene und vorhergesagte Konzentrationen bestimmter Gase gegeneinander aufzutragen, wobei je nach Messeinrichtung unterschiedliche Farben verwendet werden. Die Darstellungen haben das typische Problem der Identifikation von Häufigkeiten in Scatterplots, wenn viele Punkte dicht nebeneinander liegen.

Macêdo u. a. (2000) gehen über solche einfachen, meist nicht-interaktiven Darstellungen hinaus. Dazu setzen sie insbesondere Focussing & Linking im Klimaumfeld ein. Um monatliche, gemessene atmosphärische und ozeanische Daten zu untersuchen, werden die Systeme XGobi und Arcview (vgl. Abs. 3.3.3) kombiniert, um Darstellungstechniken für multivariate Daten (Scatterplots, Perspective Wall) mit räumlichen Darstellungstechniken (Erdkugeldarstellung, Pfeildarstellungen) in verschiedenen parallelen Sichten über Brushing & Linking zu koppeln. Die Autoren weisen darauf hin, dass gerade im Bereich der atmosphärischen Wissenschaften die Definition von „bedingten Verteilungen“ auf den Wertebereichen mehrerer Variable, wie sie interaktiv z.B. in Scatterplots definiert werden können, sehr nützlich ist (vgl. hierzu auch Wilks 95). Weiterhin zeigen sie für atmosphärische Daten ausgewählte, zylindrische 3D-Projektionen im Sinne der Grand-Tour.

Bei der Darstellung des Hurrikan Isabel gehen Doleisch u. a. (2004) über ein einfaches Fokussing & Linking weit hinaus. Sie definieren in mehreren, miteinander gekoppelten Scatterplots bestimmte Features, die dann in einer 3D-Sturmvisualisierung wiedergegeben werden. Dazu werden multiple-, auch unscharfe Brushes miteinander kombiniert, was dem Anwender ein hohes Maß an Interaktivität bei der Festlegung von interessanten Kombinationen von Datenwerten gibt, und so sein Verständnis für die visualisierten Daten wesentlich steigert. Über die Arbeit von Doleisch u. a. (2004) hinaus wurden jedoch multiple, auch unscharfe Brushes bisher nicht zur visuellen Analyse von Klimadaten eingesetzt.

Visualisierungsmethoden für geclusterte Klimadaten. Clusterung ist in der Klimaforschung ein beliebtes Mittel zur Datenaggregation und Modellvalidierung. So verwenden z.B. Kücken u. a. (1999) zur Darstellung von geclusterten Klimadaten auf regelmäßigen 2D-Gittern nach der Clusterzugehörigkeit farbkodierte Kreise. Neben dem Kontrastproblem dieser (ausgefüllten) Kreise zu darunter liegenden geographischen Daten wie Land- und Meerdarstellungen tritt hier auch das - im Umfeld der Farbabbildung nominaler Daten bekannte - Problem auf, dass ähnliche Farben beim Anwender ähnliche Clustereigenschaften suggerieren, die im allgemeinen so nicht gegeben sind. Um diesem Problem zu begegnen, das sich bei großen Clusterzahlen noch verschärft, entwerfen Böhm u. a. (2005) eine semantikbasierte Clustereinfärbung aufgrund eines Clustereigenschaftsmaßes basierend auf den Clustermittelwerten. Durch diese Art der Einfärbung sind die Cluster zwar nicht immer visuell unterscheidbar, können aber gemäß ihrer Eigenschaften geordnet und ausgegeben werden.

Darüber hinaus mangelt es in der Literatur an Methoden zur Darstellung von Clusterungen in Kombination mit den Clustereigenschaften im räumlichen und zeitlichen Kontext. Hier entwirft die vorliegende Arbeit systematisch verschiedene Techniken, oder wendet bekannte Techniken zur Clusterdarstellung im Klimakontext an (vgl. Kap. 5).

3.2.3 Visuelles Data Mining in der Klimamodellbildung, -simulation und -evaluation

Typischerweise fokussiert das VDM auf die Exploration von Daten. In diesem Bereich hat es in den letzten Jahren bedeutende Fortschritte gegeben, gerade was die Kopplung verschiedener Verfahren unter starker Einbeziehung der menschlichen Wahrnehmungsfähigkeiten betrifft. Im Unterschied dazu ist der Einsatz des VDM im Prozess von Modellbildung, Simulation und Modellevaluation in der Literatur noch weitgehend unerforscht. Zwar ist der kombinierte Einsatz von Simulation und Visualisierung schon seit längerem Bestandteil von Simulationssystemen, wie es z.B. die VDI-Richtlinie VDI 3633 fordert. Jedoch ist die konsequente Kopplung von Visualisierung und statistischen Verfahren zur Unterstützung von Exploration, Hypothesenbildung und -validierung, Modellbildung, Simulation und Modellevaluierung noch weitgehend offenes Forschungsgebiet. Diese Aussage gilt insbesondere auch im Umfeld der Klimamodellierung. In der Literatur finden sich vor allem separate Lösungen für einzelne dieser Prozesse, die sich zumeist entweder auf statistische oder auf visuelle Methoden konzentrieren.

Vor allem die **visuell gestützte Exploration** wurde bisher gut untersucht (vgl. voriger Abs.). Darüber hinaus ist die statistische Unterstützung von **Hypothesenbildung und -validierung** im Klimaumfeld eine häufig eingesetzte Vorgehensweise (vgl. auch Abs. 3.2.1).

Vereinzelt - jedoch nicht im Klimaumfeld - finden sich auch Visualisierungen zur Unterstützung der **Modellbildung**. So unterstützen z.B. Matkovic u. a. (2002) sowie Koolwaaij u. van Leeuwen (2003) die Darstellung von Abläufen in Prozessmodellen. Im Gegensatz dazu ist für die Modellierung im Klimaumfeld jedoch zumeist die Quelltextprogrammierung die typische Vorgehensweise.

Des Weiteren ist die Online-**Visualisierung von Modellsimulationen** eine typische Anwendung zur Überwachung von deren Verlauf. Solche Visualisierungen greifen jedoch typischerweise nicht in die Simulation ein, sondern geben diese nur (statisch) aus. Vereinzelt gibt es aber auch Visualisierungen, die direkt in die Simulation eingreifen. Ein Beispiel hierfür ist die Arbeit von Hrdlicka u. a. (2003), wo der Nutzer interaktiv die Simulation eines Verbrennungsprozesses beeinflussen kann. Im Klimaumfeld wird eine solche Art der Interaktion bisher nicht eingesetzt.

Weiterhin wurde die **Modellevaluierung** im Klimaumfeld, die z.B. über den Vergleich von Messdaten und simulierten Daten erfolgt, sowohl durch statistische Methoden als auch Visualisierungsme-

thoden untersucht (vgl. z.B. Böhm u. a. 2005). Aus dem Visualisierungsumfeld bekannte Verfahren zur „vergleichenden Visualisierung“ (vgl. z.B. Pagendarm u. Post 1995, sowie Kapitel 5.4) werden hier jedoch kaum eingesetzt.

Für den Einsatz des VDM im gesamten Modellierungsprozess, insbesondere in der Klimafolgenforschung, besteht noch Forschungsbedarf. Erste wichtige Ansätze hierzu finden sich in Nocke u. a. (2003). Die dort vorgestellte Konzeption zum VDM in der Modellbildung und Simulation wird in Abschnitt 6.3 konzeptionell ausgebaut und dessen Anwendung demonstriert.

3.2.4 Diskussion

In diesem Abschnitt wurde - neben dem Einsatz statistischer Verfahren - vor allem der Einsatz von Visualisierungstechniken im Umfeld der Analyse von Klimadaten untersucht. Dabei lassen sich drei Kategorien von Arbeiten unterscheiden:

1. teilweise interaktive Visualisierungen von Klimaforschern zur Illustrierung ihrer Forschungsergebnisse,
2. nicht-interaktive Visualisierungen von Klimaforschern und Medienexperten zur Präsentation von Wetter- und Klimavorhersagen für ein breites Publikum und
3. interaktive Visualisierungen von Klimadaten umgesetzt von Forschern aus dem Visualsierungsumfeld.

Arbeiten der ersten Kategorie haben häufig ihren Schwerpunkt auf der Präsentation für Publikationen, und konzentrieren sich auf wenige, im Klimaumfeld übliche Techniken. Hierbei werden meist einzelne Bilder oder Animationen verwendet. Durch die Einhaltung dieser Konventionen sind sie im Klimaumfeld leicht kommunizierbar, schöpfen jedoch das Potential moderner Interaktionstechniken und neuer Visualisierungsmetaphern kaum aus. Der Fokus hierbei liegt auf statischen, räumlichen Darstellungen. Eine Untersuchung, ob Informationsvisualisierungstechniken zur **Darstellung zeitlicher Phänomene**, zur Darstellung **multi-variater Daten** ohne räumlichen und zeitlichen Bezug, zur Darstellung **von Clusterungen und Clustereigenschaften in Raum- und Zeit** sowie zur Darstellung **von Multiparameterdaten auf Karten** auch in diesem Kontext geeignet sind, steht noch aus.

Darstellungen der zweiten Kategorie sind für die Kommunikation der von den Klimaforschern generierten Erkenntnisse für ein breiteres Publikum gedacht (vgl. auch Klasse 5 nach Treinish 1999). Diese konzentrieren sich auf wenige wesentliche Muster in den Daten, erlauben aber keine Interaktionen. Solche Darstellungen sind mittlerweise relativ ausgereift (vgl. z.B. Schröder 1997).

In der dritten Kategorie wurden verschiedene, im Visualisierungsumfeld bekannte Darstellungs- und Interaktionstechniken für Klimadaten angewendet. Hierzu gehören u.a. Isoflächen und DVR Methoden für 3D-Klimadaten oder LIC-Darstellungen für 2D-Strömungsdaten. Jedoch wurde auch hier der Einsatz existierender Techniken erst in Ansätzen untersucht. So sind beispielsweise moderne (topologiebasierte) Methoden der **Strömungsvisualisierung** für dieses Umfeld kaum in der Literatur präsent. Ferner werden moderne **Interaktionstechniken** wie Brushing & Linking oder Fokus & Kontext in diesem Kontext kaum eingesetzt.

Neben dem Einsatz solcher Visualisierungs- und Interaktionstechniken im Klimaumfeld ist weiterhin kaum untersucht, wie der Anwender weitgehend bei der **Wahl und Parametrisierung** geeigneter (Visualisierungs-)Techniken für seinen Kontext zu unterstützen sei. Gerade vor dem Hintergrund vielfältiger Konventionen auf der einen Seite, und einer hohen Zahl verfügbarer Techniken und einstellbarer Parameter auf der anderen Seite, sind diese für Klimaforscher oft schwer zu überschauen.

Darüber hinaus ist die konsequente **Kopplung der Visualisierung mit statistischen Verfahren** im Sinne des VDM in diesem Umfeld bisher nicht thematisisert. Deren Einsatmöglichkeiten zur Unterstützung des gesamten Modellbildungs- und Evaluationsprozesses ist noch offener Forschungsgegenstand.

Entsprechend ergeben sich für die hier vorgelegte Arbeit eine Vielzahl von **Herausforderungen**. Insbesondere sollen in dieser Arbeit die Einsatzmöglichkeiten interaktiver Informationsvisualisierungstechniken in enger Kopplung mit statistischen Verfahren in dem für die Klimaforschung typischen Raum- und Zeitbezug untersucht, sowie dem Nutzer Hilfestellungen bei der Erzeugung von für seinen Kontext geeigneten Darstellungen gegeben werden.

3.3 Systeme

Der stete Anstieg der weltweit erhobenen Datenmengen hat zu einer Vielfalt von Analysesystemen geführt. Nachdem bisher einzelne Techniken vorgestellt wurden, sollen nun in diesem Abschnitt existierende Analysetools im Umfeld des Visuellen Data Mining vorgestellt werden und mit dem Fokus auf deren Anwendbarkeit im Klimaumfeld hin untersucht werden. Um entscheiden zu können, welche der Systeme für diesen Anwendungshintergrund besonders geeignet sind, können u.a. die folgenden Kriterien herangezogen werden (vgl. Hege 1992):

1. Standardnähe, Portabilität, Modernität,
2. Unterstützung verschiedener Arbeitsweisen (z.B. Interaktivität, oder verteiltes Arbeiten),
3. Eignung für eine bestimmte Zielgruppe (z.B. Graphikprogrammierer, Laie),
4. Komplexität der Bedienung, Einarbeitungszeit,
5. Universalität vs. spezielle Anwendungen,
6. Produktumgebungen (z.B. vorhandene Hard- und Software, Kompatibilität, Datenformate),
7. Qualitätsanforderungen,
8. Geschwindigkeitsanforderungen,
9. Zukunftsträchtigkeit (bei längeren Projekten).

Diese Kriterien, die im Visualisierungsumfeld aufgestellt wurden, lassen sich auch auf andere Systeme, z.B. auch auf Data Mining Systeme, übertragen. Im folgenden sollen verschiedene Klassen von Systemen mit den ihnen typischen Charakteristika kurz vorgestellt werden, um sie auf ihre Eignung im Klimaumfeld hin zu untersuchen.

3.3.1 Modulare Visualisierungsumgebungen

Modulare Visualisierungssysteme[8] haben sich als leicht-benutzbare visuelle Programmierumgebungen im Bereich der wissenschaftlichen Visualisierung etabliert (vgl. z.B. Brodlie 1992; Cameron 1995; Schröder 1997; Fujishiro u. a. 2000). Sie stellen eine breite Palette an Visualisierungsfunktionalität und Nutzerunterstützung bereit und ermöglichen ihren Nutzern, relativ schnell erste aussagekräftige Bilder zu generieren. Eigenschaften modularer Visualisierungsumgebungen sind

- Visuelle Programmierbarkeit: Der Nutzer kann interaktiv Visualisierungen in Form datenflussorientierter Modulgraphen zusammensetzen. Er verknüpft bestimmte Module, die jeweils eine spezielle Funktion umsetzen (z.B. Datenimport oder Stromlinienverfolgung) zu einem Netzwerk (gericheter Graph), in dem die Daten von Modul zu Modul weitergereicht werden.

[8] engl.: modular visualization environments (MVEs)

- Modularität: Der Nutzer kann entweder komplexe Teilgraphen als allgemeineres Modul definieren oder auch eigene Module in einer Basisprogrammiersprache entwerfen.
- Flexible Datenbeschreibung: Mächtige Formate zur Datenbeschreibung ermöglichen ein breites Spektrum an Datenklassen zu verwalten. Außerdem wird der Nutzer beim Datenimport unterstützt.
- Integration von Standardtechniken: Die Systeme enthalten z.T. komplexe Standardalgorithmen, z.B. Marching Cubes oder Stromlinienintegration.
- Flexible Ansteuerungskonzepte und Schnittstellen: Visualisierungen können auch über externe Schnittstellen erzeugt und angesteuert werden (z.B. über Skriptsprachen).
- Erweiterbarkeit: Falls für die Erstellung einer Visualisierung ein wie oben beschriebener Modulgraph nicht ausreicht (z.B. um eine effizientere Umsetzung eines Visualisierungsmoduls zu erreichen oder komplexe Berechnungen zu integrieren, können auch eigene Module in einer Basisprogrammiersprache entworfen werden.
- Plattformunabhängigkeit: Visualisierungssysteme werden für eine Vielfalt von Betriebssystemplatformen angeboten.

Wichtige Beispiele für modulare Visualisierungsumgebungen sind Amira[9], AVS[10], IDL[11], IRIS-Explorer[12], OpenDX[13] und VTK[14]. Diese Systeme sind auf die Analyse räumlicher Daten spezialisiert, können aber auch zum Entwurf multivariater Darstellungen ohne Raumbezug verwendet werden. Sie sind universell einsetzbar und flexibel an eine Problemstellung anpassbar. Sie können benutzt werden, um schnell Prototypen zu entwerfen und effizient komplexere Visualisierungen durch Wiederverwendung existierender Algorithmen und Designelemente umzusetzen (vgl. z.B. Treinish 1999). Insbesondere OpenDX wird wegen seiner freien Verfügbarkeit im Umfeld der Klimaforschung häufig verwendet.

Trotz der Flexibilität dieser Systeme bleibt es für Visualisierungslaien schwer, komplexere Visualisierungen geeignet zu entwerfen. Probleme stellen unter anderem die schnell anwachsende Komplexität der entworfenen Netzwerke, die Entwicklung allgemeingültiger, fehlertoleranter Visualisierungen nicht nur für fest verdrahtete – sondern auch für ähnliche – Datensätze sowie die begrenzten Möglichkeiten zum Entwurf von Parameterdialogen dar. Auch sind sie z.T. nicht dafür ausgelegt, ggf. direkt die Graphikhardware anzusprechen und eine effektive, auf große Datenmengen zugeschnitte Datenhaltung zu unterstützen. Weiterhin stoßen sie zum Teil aufgrund ihrer Allgemeingültigkeit an Grenzen, wenn es um die Einbindung spezieller Interaktionstechniken wie z.B. die Verwendung graphischer Lupen oder beliebiger visueller Anfragen an die Darstellung geht.

3.3.2 Monolithische Visualisierungssysteme im Klimaumfeld

Die im vorigen Abschnitt vorgestellten allgemeinen, modularen Visualsisierungswerkzeuge sind zumeist nicht fokussiert genug für die Erfordernisse spezieller Anwendungen, was insbesondere mit einem erhöhten Einarbeitungsaufwand für den Nutzer einhergeht (vgl. z.B. Treinish 1999). Dafür bieten sie eine breite bestehende Infrastruktur, die Datenhaltung und vorgefertigte Algorithmen einschließt. Im Gegensatz dazu reduzieren anwendungsspezifische monolithische Visualisierungssysteme[15] den Trainingsaufwand, und liefern Visualisierungen mit für die Anwendung geeigneten

[9] http://www.amiravis.com
[10] http://www.avs.com
[11] http://www.ittvis.com/idl
[12] http://www.nag.co.uk
[13] frei verfügbare Version des IBM Data Explorer, http://www.opendx.org
[14] http://www.kitware.com
[15] sogenannte Turnkey-Systeme

Parametrisierungen. Weiterhin sind sie zumeist auf bestimmte Datenklassen (z.B. 3D skalare Daten) zugeschnitten, und können so hinsichtlich Geschwindigkeit und Darstellungsqualität optimiert werden (vgl. z.B. Treinish u. a. 1992; Schröder 1997). Entsprechend können viele Aufgaben automatisiert werden. Auch diese Systeme unterstützen i. allg. Datenimport aus verschiedenen Formaten und setzen zumeist keine speziellen Kenntnisse im Bereich der Visualisierung voraus. Da die Anwender nicht mit der Struktur der internen Datenflüsse vertraut sein müssen, können sie sich ganz auf die Analyseaufgabe konzentrieren.

Solche Systeme sind häufig auf eine Menge von festverdrahteten Visualisierungen beschränkt. Dies birgt den Nachteil einer fehlenden Unterstützung bei deren Erweiterung, falls sich im Rahmen der Analyse neue, nicht durch das System abgedeckte Nutzeranforderungen oder -aufgaben herauskristallisieren. Neue Funktionen können nicht innerhalb des Systems von den Anwendern selbst spezifiziert werden, sondern müssen über ein Software-Update realisiert werden. Auch die Wiederverwendbarkeit einzelner Module ist rein auf die programmiertechnische Ebene beschränkt.

Beispiele für in Klimaforschung und Meteorologie eingesetzte monolitische Systeme sind **AWIPS**[16], **Earth System Visualizer**, **FERRET**[17], **Generic Mapping Tools**[18] (GMT), **GrADs**[19], **McIDAS**[20], **MeteoVis**[21], **Ocean Data View**[22], **RASSIN**[23], **TriVis**[24] und **Vis5D**[25]. Die meisten dieser Systeme beinhalten reine 2D-Darstellungen, wie sie unter Einfluss der Kartographie und aus den Anfängen der Computergraphik noch heute üblich sind.

So ist beispielsweise das System **GrADs** (Grid Analysis and Display System) speziell für die Anwendung im Bereich der Erdsystemanalyse zugeschnitten (Doty 2006). Bei fester Kodierung des geographischen Bezuges können 2D-Daten auf verschiedenen Gittertypen – unter anderem auch gleichzeitig – dargestellt und für das Anwendungsumfeld geeignet beschriftet werden. GrADs unterstützt unter anderem Scatterplots und Liniendiagramme für multivariate Daten, Farb- und Isoliniendarstellungen für 2D-Daten, Strömungslinien und Pfeildarstellungen für Strömungsdaten, sowie verschiedene Darstellungen von einfachen Symbolen für gestreute Stationsdaten. Über eine Fortranähnliche Skriptsprache können über die Kommandozeile oder im Skriptmodus eine Palette von vordefinierten Funktionen (z.B. Datenselektion) angesprochen sowie eine graphische Oberfläche erzeugt werden. Über selbst definierbare Funktionen ist GrADs in einem gewissen Umfang erweiterbar. Die Nutzer müssen zur Benutzung des Systems mit Skriptprogrammierung vertraut sein, so dass GrADs für Informatik-Laien eher weniger geeignet ist.

Im Gegensatz zu GrADs ist **Vis5D** speziell zur Visualisierung meteorologischer Daten in 3D ausgerichtet. Es eignet sich im besonderen zur Volumen- und Strömungsvisualisierung von zeitveränderlichen, atmosphärischen 3D Daten auf regelmäßigen Gittern (vgl. Schröder 1997; Hibbard 2001). Bereitgestellte Visualisierungstechniken sind Isolinien und -flächen, Schnittflächen und DVR für Volumen-Daten sowie Pfeildarstellungen und Trajektorien für vektorielle Daten, die insbesondere auch gleichzeitige Darstellungen mehrerer Merkmale in performanter Art erlauben. Schwächen des Systems ist die eingeschränkte Funktionalität zur Analyse dynamischer Muster über der Zeit (vgl. Schröder 1997).

Resümierend kann festgestellt werden, dass die unkomplizierte und schnelle Auswertung der Daten monolithische Visualisierungssysteme für Klimaforscher und Meteorologen attraktiv macht, obwohl

[16] http://www.weather.gov/tdl/awips
[17] http://ferret.wrc.noaa.gov/Ferret
[18] http://gmt.soest.hawaii.edu
[19] http://grads.iges.org/grads
[20] http://www.unidata.ucar.edu/software/mcidas
[21] vgl. Weihai u. Zesheng (1994)
[22] http://odv.awi-bremerhaven.de/
[23] vgl. Schröder (1997) S. 109ff
[24] vgl. Schröder (1997) S. 147ff
[25] http://vis5d.sourceforge.net

sie das Potential moderner Informationsvisualisierungssysteme zumeist nicht ausschöpfen (gerade bei der Interaktion und bei der Kombination mehrerer Sichten auf die Daten). Sie sind häufig auf bestimmte Datenklassen zugeschnitten, und erfordern einen erheblichen Aufwand bei der Erweiterung ihrer Funktionalität.

Des Weiteren können auch andere monolithische Visualisierungssysteme, die ursprünglich nicht für Anwendungen im Umfeld der Meteorologie und Klimaforschung entwickelt wurden, in diesem Umfeld eingesetzt werden. So wird beispielsweise das System **SimVis** (vgl. z.B Doleisch u. a. 2003b; Doleisch 2005), welches originär für die Darstellung von Verbrennungsvorgängen in Motoren entwickelt wurde, auch für die Visualisierung meteorologischer Phänomene eingesetzt (vgl. Doleisch u. a. 2004). Im Unterschied zu den oben genannten Systemen wird in SimVis eine breite Palette an heute üblichen Interaktionstechniken (z.B. verschiedene Brushes) angeboten.

Darüber hinaus gibt es weitere Systeme, welche Klimadaten visualisieren, jedoch keine monolithische Struktur aufweisen. Dazu gehört zum Beispiel das System **REINAS**[26], welches interaktive Visualisierungen von Wetterdaten über eine VRML-basierte Webschnittstelle einer breiten Nutzergruppe zur Verfügung stellt. Ein weiteres Beispiel für ein Webportal stellt das System **CLIMVIS**[27] dar, welches zur Web-Präsentation typischer klimatischer Bedingungen entwickelt wurde (vgl. Ross u. a. 1997), und z.B. auch für Touristen zur Urlaubsplanung verwendet werden kann.

3.3.3 Informationsvisualisierungssysteme

Im Laufe des letzten Jahrzehnts haben sich eine breite Palette an Informationvisualisierungstechniken entwickelt, die häufig in Softwaresysteme eingeflossen sind. Zur Evaluierung der Leistungsfähigkeit von Visualisierungstechniken wurden verschiedene Taxonomien bzw. Kategorisierungen entworfen. Diese Taxonomien haben auch den Vorteil, für das Visualisierungsdesign eingesetzt werden zu können (vgl. Abs. 3.4.2), da sie die Techniken u.a. nach wichtigen Datenklassen und Aufgaben einteilen. Diese Kategorisierungen wurden zwar originär für Visualisierungstechniken entworfen, lassen sich aber auch auf (Informations-)Visualisierungssysteme übertragen.

Beispiele für Taxonomien im Umfeld der Informationsvisualisierung sind die Task-by-Data-Type Taxonomie von Shneiderman (1996), die Einteilungen von Keim u. Kriegel (1996) und Card u. Mackinlay (1997), sowie die aus dem „Data-State-Reference“-Modell abgeleitete Taxonomie von Chi u. Riedl (2000). Entsprechend können (Informations-)Visualisierungstechniken in die folgenden Kategorien eingeordnet werden:

- Datentyp: Temporal, 1D, 2D, 3D, Text, Multi-Dimensional, Baum, Netzwerk (Shneiderman 1996; Chi u. Riedl 2000),
- Aufgaben: Overview, Filter, Details-On-Demand, History, ... (Shneiderman 1996; Fujishiro u. a. 2000),
- Art der Abbildung: geometrische Projektion, ikonenbasierte Techniken, pixelorientierte Techniken, hierarchische Techniken ... (Keim u. Kriegel 1996),
- Art der unterstützten Interaktionen: graphische Operationen, Mengenoperationen, Datenoperationen (Chuah u. Roth 1996; Shneiderman 1996),
- technische Aspekte: Arten eingesetzter Operatoren, ... (Chuah u. Roth 1996; Card u. Mackinlay 1997; Chi u. Riedl 2000).

Die hier aufgelisteten Zugänge sind nicht immer völlig durchschnittsfremd (z.B. Aufgaben und Art der unterstützten Interaktionen), sie liefern jedoch wichtige Einteilungsaspekte in praktikabler Weise. Ein allgemeiner Überblick zu den verschiedenen Taxonomien für Informationsvisualisierungstechniken findet sich z.B. bei Kreuseler (2004).

[26] vgl. Djurcilov u. Pang (1998)
[27] http://www.ncdc.noaa.gov

Im folgenden sollen diese Taxonomien auf Informationsvisualisierungssysteme übertragen werden. So kann ein System auf einen Datentyp bzw. auf eine Klasse von Techniken zugeschnitten oder Techniken für mehrere Datentypen mit variierenden Abbildungsparadigmen bereitgestellt werden. Bei den Aufgaben werden an moderne Informationsvisualisierungssysteme insbesondere die Unterstützung von Shneidermans' Mantra

„Overview first, Zoom and Filter, then Detail on Demand"

gefordert (Shneiderman 1996). Darüber hinaus variieren die Systeme, ob sie z.B. auch eine Historienverwaltung des Analyseprozesses durchführen. Die Interaktionen in Sytemen beinhalten neben Interaktionen auf einzelnen Techniken zusätzlich auch die Kopplung der Techniken (z.B. beim Brushing). In starkem Maße betreffen die technischen Aspekte die Einteilung von Visualisierungssystemen, denkt man z.B. an die Wiederverwendung verschiedener Operatoren von Chi u. Riedl (2000). Hierunter fällt die softwaretechnische Grundstruktur des Systems, welche bestimmt, ob es sich z.B. um ein ein Toolkit, ein ganzes System oder eine Bibliothek handelt.

Tabelle 3.1 zeigt wichtige Toolkits, Systeme und Bibliotheken im Bereich der Informationsvisualisierung. Diese sind nach dem Datentyp und der Anwendungsdomäne sortiert. Vornehmliche Charak-

Datentyp	Anwendung	System	Systemtyp	Funktionsumfang
1D & temporal	Allgemein	TimeSearcher (2007)	System	Multiple-Zeitdiagr.,Winkel- u. Boxanfrag.,Mustersuche
		VisAxes (2007)	System	bel. Streckenzugdarstellungen, versch. vis. Anfragen
2D	Allgemein	GeoVISTA (2007)	Toolkit	Kartenvis., PK, 3D-Vis., Statistik, SOMs
	Gesundheitsdaten	TeCoMed (1998)	System	Ikonen auf Karten, insb. Zeitik., geogr. Linsen
	GIS	ArcInfo (2007)	System	High-end GIS-System, hochqual. Kartendarst.
3D	Allgemein	SimVis (2007)	System	mit bel. Brushes gekoppelte 3D- u. SP-Darst.
	Medizin	VHE (2007)	System	interakt. Präsentation von 3D-Daten, DVR
		ITK (2007)	Toolkit	versch. Segmentierungs- und Registrierungsverf.
Multi-D	Allgemein	VisDB (2007)	System	Pixeltechniken, PK, „Stick Figures"
		Polaris (2007)	System	autom. VisDesign, SPs, Zeitdiag. u.v.a.
		XGobi (2006)	System	gekoppelte SPM und PK
Bäume	Allgemein	Klimt (2006)	System	Node-Link-Diag., Treemaps, Scatterplots, Anfragen
	Genomdaten	JavaTreeView (2007)	Toolkit	SPM., Dendrogramme, Farddiagramme
		HCE (2007)	System	SPM, Dendrogr., Farddiag., PK
Graphen	Allgemein	GraphViz (2007)	System	Vielfalt and Node-Link-Layouts
		Tulip (2007)	System	Vielfalt an Graph-Algos., skalierbare Node-Link-Vis.
		WilmaScope (2007)	System	Federkraftbas. Node-Link-Vis. (auch Cluster)
		Touchgraph (2007)	Bibliothek	Node-Link-Layouts
		Pajek (2007)	System	Vielfalt an Node-Link-Layouts
		JUNG (2007)	Bibliothek	Vielfalt an Graph-Algos., adaptive Node-Link-Diagr.
	Dokumentdaten	CiteWiz (2007)	System	„Growing Polygon"-Vis., Hierarchie- und Tortendiagr.
hybrid	Allgemein	VisServer (2007)	System	Hyperbol.-Hierar.-Darst., Perspective Wall, Tabellen
		Piccolo (2006)	Toolkit	2D-Grafikentw., Szenengraphen, Navigationsunterst.
		InfoVis (2007)	Toolkit	Vielzahl an Graph- und Hierarchiedarst., PK, SP u.v.a
		Spotfire (2007)	System	SP, Zeit-, Karten- u. Tortendiagr.,VisDesign, History
		Visage (2006)	System	Karten-, Ikonen-, Balken-Darst. u.a., VisDesign
		Prefuse (2007)	Toolkit	Versch. Tablellen-, Graphen-, Baum- u. Zeitdarst.
		InfoVis3D[28]	System	Hierarchiedarst., Topic-Maps, SPM, PK, History
	Wirtschaftsdaten	ILOGViews (2007)	Bibliothek	Karten-, Graphen-, PK,Balken-,Tortendiag. u.v.a.
	Dokumentdaten	InSpire (2007)	System	Zeitdiag., Galaxien, Topic-Maps, Balkendiag. u.a.

Tabelle 3.1: Ausgewählte Informationsvisualisierungssysteme und -tools, sortiert nach Datentyp und Anwendungsbezug

teristik der hier aufgelisteten Systeme ist, dass sie eine breite Palette der auf Seite 13 vorgestellten Interaktionstechniken umsetzen, und damit einen hohen Grad an Nutzerintegration in den Analyseprozess einbringen. Weiterhin stellen sie parallele, gekoppelte Darstellungen bereit, die es dem Anwender erlauben, zu tieferen Einsichten über die Zusammenhänge in den Daten zu gelangen.

Viele der hier vorgestellten Systeme haben auch einige wenige automatische Mining Verfahren integriert (z.B. Clusterungsalgorithmen oder Self-Organizing-Maps (SOMs)), wurden jedoch aufgrund ihres Fokus auf die Visualisierung hier einsortiert. Auch erreichen sie nicht die starke Interaktion von automatischen Verfahren und Visualisierungstechniken, und gehen gewöhnlich nicht über die

1. Ebene von Ankerst hinaus. Die Grenzen zu den in Abschnitt 3.3.5 vorgestellten VDM-Systemen sind jedoch fließend.

Die Palette aufgelisteter Systeme verdeutlicht das Potential moderner Informationsvisualisierungssysteme. Entsprechend ist es Ziel dieser Arbeit, die Mächtigkeit ihrer Interaktionmechanismen auch für die Untersuchung von klimatischen Phänomenen einzusetzen. Dabei sind vor allem Systeme für 1D/temporale, 2D und 3D Daten von Interesse, wie sie auch im Klimaumfeld vorliegen. So unterstützt das System **GeoVISTA** insbesondere die Visualisierung von Daten mit kartographischem Bezug, und benutzt farbige Gebietsdarstellungen gekoppelt mit einer „Parallele Koordinaten“ Darstellung und mit einer SOM-View. Zusätzlich wird in diesem System auch eine abstrakte Matrixsicht für Datensätze mit hoher Variablenanzahl integriert, in denen Variablenkorrelationen und -entropien darstellt werden. Auch die flexiblen Interaktionsmöglichkeiten zum dynamischen Filtern in **TimeSearcher** und **VisAxes** mit ihren flexiblen Filter- und Lupenfunktionen haben Potential für die Analyse von langen Klimazeitreihen über viele Merkmale. Darüber hinaus sind auch Systeme mit dem Fokus einer multi-dimensionalen Sicht für die Analyse von Klimadaten relevant, denn solche - von Raum- und Zeitbezug abstrahierende - Sichten sind in der Klimaforschung bisher eher vernachlässigt worden (vgl. Abs. 3.2.2).

3.3.4 Data Mining Systeme

Auch im Bereich des (automatischen) Data Mining finden sich eine Vielzahl von Systemen, die verschiedene der in Abschnitt 3.1.2 vorgestellten DM-Verfahren umsetzen und von Bedeutung für die in dieser Arbeit umgesetzten Frameworks sind. Han u. Kamber (2000) klassifizieren solche Verfahren nach den vier folgenden Kriterien:

- **Zugrunde liegende Datenbank und zugrunde liegende Datentypen**: Dies beinhaltet vor allem die zum Einsatz kommenden Datenbanken und deren Datenmodell. So werden u.a. relationale, transaktionale, objektorientierte und Data Warehouse Mining Systeme unterschieden. Bezüglich der zugrunde liegenden Datentypen können die Data Mining Verfahren weiterhin gemäß dem Basismodell des VDM (vgl. S. 11) in objektzentrierte (z.B. Text oder Multimedia Data Mining Systeme), attributzentrierte (z.B. raum- oder zeitgezogene Data Mining Systeme) und strukturzentrierte (z.B. Graphen-Analyse Systeme) unterteilt werden.

- **Art der gefundenen Muster**: Hierbei kann z.B. zwischen Cluster-, Hauptkomponenten-, Assoziations- oder Aussreißer-Analyse unterschieden werden. Je nach Art der extrahierten Muster können dabei verschiedene Abstraktionsgrade unterschieden werden.

- **Art der eingesetzten Techniken**: Hierbei handelt es sich um die in Abschnitt 3.1.2 dargestellten Gebiete, denen die Techniken zugeordnet werden können. Dazu gehören z.B. datenbankorientierte und Data Warehouse-orientierte Verfahren (z.B. OLAP) sowie Techniken aus den Bereichen maschinelles Lernen, neuronale Netze, Statistik und Visualisierung. Weiterhin kann hier auch nach dem Grad der Nutzerinteraktion unterschieden werden (z.B. autonome, interaktiv explorative oder Anfragesysteme).

- **Art des Anwendungsumfeldes**: Data Mining Systeme können entweder allgemein anwendbar („general purpose“) oder für ein spezielles Anwendungsgebiet („special purpose“) mit speziellen Hintergrundwissen entworfen worden sein. Die Handhabung spezieller Kontextaspekte und die Einbindung für eine Anwendung üblicher Analyseverfahren spielt auch in dieser Arbeit eine besondere Bedeutung.

[28] in Kreuseler u. Schumann (2002b)

Wichtige Data Mining Systeme sind Clementine[29], Enterprise Miner[30] und Intelligent Miner for Data[31]. Diese Systeme zeichnen sich durch eine breite Palette automatischer Mining-Techniken aus (z.B. Cluster-, Klassifikations- und Regressionsanalyse). Sie besitzen eine Vielzahl von fensterbasierten Nutzerinteraktionen wie Datenselektion und Algorithmenparametrisierung. Auch eine Vielzahl von Standard-(informations)visualisierungstechniken wie Histogramme, Liniendiagramme und Scatterplots wurde integriert, allerdings handelt es sich hierbei zumeist um nicht-interaktive Ausgabeplots, die lediglich die erste Ebene nach Ankerst (2001) unterstützen. Komplexere Visualisierungstechniken – z.B. zur Darstellung von räumlichen Daten oder von Strukturen – fehlen zumeist. Ausnahme hiervon ist das System Diamond (Parallele Koordinaten).

3.3.5 Visuelle Data Mining Systeme

Hier sollen Systeme vorgestellt werden, die bei der Kopplung von Visualisierung und automatischen Verfahren über die 1. Ebene von Ankerst hinausgehen. Sie verknüpfen eine Vielzahl von automatischen Techniken mit Visualisierungstechniken, und/oder geben dem Anwender durch die Darstellung von Zwischenergebnissen bzw. durch geeignete Unterstützungsfunktionen eine verbesserte Einsicht in die Generierung von Mining-Ergebnissen. Hier soll zwischen allgemeingültigen, auf spezielle Aufgaben zugeschnittenen und anwendungsspezifischen Systemen unterschieden werden.

Zu den allgemeingültigen Systemen gehören **OmniViz Pro**[32], SAS JMP[33] und Manet[34], **Mondrian**[35] und Mineset[36].

OmniViz Pro ist ein VDM System, welches eine Vielzahl von Visualisierungstechniken mit statistischen Verfahren koppelt. Es unterstützt verschiedene Datenklassen wie nominale und kontinuierliche Merkmale und ist skalierbar auch für sehr große Datenmengen (bis zu eine Milliarde Datenwerte oder eine Millionen Dokumente). Es ist geeignet für verschiedene Anwendungen wie Genomdaten, chemische Strukturen, Finanzdaten, Dokumente u.a. Flexibel kann der Nutzer eigene Datenformate einlesen und neue automatische Mining Verfahren anbinden.

SAS JMP verknüpft statistische Techniken mit Visualisierungen. Es benutzt einen Skript-Mechanismus, um erfolgreiche Analysesequenzen abzuspeichern und wiederherzustellen. Es ist insbesondere für Nichtexperten ausgelegt und unterstützt den Anwender bei der Auswahl geeigneter Techniken. Es ist vor allem auf wirtschaftliche und Finanzdaten zugeschnitten, kann aber auch räumliche Daten darstellen.

Beispiele für auf statistische Analysen ausgerichtete VDM Systeme sind die Syteme **Manet** und **Mondrian**. Sie stellen eine Vielzahl gekoppelter Visiualisierungstechniken (z.B. Parallele Koordinaten, Mosaik-Plots) bereit, welche auf die adäquate Darstellung statistischer Kenngrößen zugeschnitten sind. Manet hat den Fokus auf der Darstellung von fehlerhaften Daten und Fehlwerten, während Mondrian multi-variate Techniken mit geographischen Karten koppelt. Beide Systeme beinhalten zusätzlich auch eine Verwaltung verschiedener Selektionssequenzen, welche eine spätere Wiederherstellung dieser Art von Interaktion ermöglicht.

MineSet, mit starker Data Mining-Funktionalität, setzt mittlerweile auch eine enge Kopplung von Visualisierung und Mining Verfahren um, und wird deswegen als VDM-System eingeordnet. So stellt es die Baumdarstellungstechnik TreeVisualizer für Entscheidungsbäume, die Technik RuleVisualizer

[29] Integral Solutions Ltd, http://www.spss.com/clementine/
[30] SAS Institute Inc., http://www.sas.com/
[31] IBM, www.software.ibm.com/data/iminer/
[32] http://www.omniviz.com
[33] http://www.jmp.com
[34] http://stats.math.uni-augsburg.de/Manet
[35] http://stats.math.uni-augsburg.de/Mondrian
[36] Silicon Graphics Inc., http://www.purpleinsight.com/, vgl. auch Brunk u. a. (1997)

für Assoziationsregeln und verschiedene kartenbasierte Darstellungen bereit.

Zu den für spezielle Aufgaben zugeschnitten Systemen gehören die auf das VDM von Clustern spezialisierten Systeme **OPTICS Visual Clustering** (vgl. Ankerst u. a. 1999) und **HD-Eye** (vgl. Hinneburg u. a. 1999). Sie erreichen eine starke Integration von automatischem und visuellen Mining, indem dem Anwender verschiedene Sichten auf Clusterungen bereitgestellt werden, und er so ein vertieftes Verständnis für deren Arbeitsweise erhält. Das System HD-Eye erlaubt zusätzlich, Entscheidungen während des Clusterprozesses darzustellen und den Anwender in diese Entscheidungen einzubeziehen.

Darüber hinaus gibt es einige anwendungsspezifische VDM Systeme. So koppelt z.B. das System **GeneSpring**[37] eine breite Palette von statistischen Methoden mit Visualisierungstechniken für die Anwendung in der Genomanalyse. Dabei bietet GeneSpring neben auf die Anwendung zugeschnittenen Visualisierungstechniken auch eine Vielzahl von geeigneten Normierungsverfahren.

3.3.6 Sonstige Systeme

Der Vollständigkeit halber sollen hier auch gängige **Mathematiksysteme** wie Matlab[38] und Mathematika[39] sowie **Statistiksysteme** wie SPSS[40] erwähnt werden. Ihre Stärken liegen in der Verwaltung und Berechnung mathematischer Formeln und statistischer Methoden, und sie benutzen die Visualisierung im Allgemeinen als Ausgabeplot. Methoden zur Kopplung von Visualisierungstechniken oder Interaktionen, wie sie aus Informationsvisualisierungssystemen bekannt sind, werden nicht oder nur rudimentär angeboten. Aufgrund ihrer Struktur erlauben sie jedoch, ggf. Zwischenergebnisse darzustellen.

3.3.7 Diskussion

Die Diskussion zeigt, dass existierende Systeme eine breite Palette an Techniken bereitstellen. Dabei sind die Grenzen zwischen Informationsvisualisierungs-, klassischen Data Mining und Visuellen Data Mining Systemen weitgehend fließend, da Informationsvisualisierungssysteme in den letzten Jahren verstärkt automatische Verfahren einbinden, und Data Mining Systeme beginnen, interaktive, gekoppelte Visualisierungen einzusetzen. Ein ähnlicher Trend zeichnet sich auch im Bereich von Statistiksoftware ab, da sich auch in diesem Bereich die Erkenntnis durchgesetzt hat, dass die interaktive Visualisierung bei sehr großen Datenmengen ein wichtiger Faktor für das Verständnis statistischer Berechnungen ist. Auch modulare Visualisierungsumgebungen, die im Unterschied zu Informationsvisualisierungssystemen einen starken Fokus auf die Analyse räumlicher Daten legen, unterstützen neben einer Vielzahl, flexibel generierbarer Visualisierungen mittlerweile auch eine immer breitere Palette an automatischen Verfahren, um Muster in den Daten zu extrahieren und für die Visualisierung aufzubereiten. Neben diesen flexibel einsetzbaren Systemen, die jedoch die speziellen Anforderungen im Bereich der Analyse von Klimadaten jeweils nur in Teilen unterstützen können, zeigt sich eine große Kluft zu den speziell für das Klimaumfeld entwickelten Systemen. Diese sind zumeist auf eine oder wenige Datenklassen eingeschränkt und unterstützen lediglich eine eingeschränkte Menge an Interaktionen.

Zusammenfassend kann festgestellt werden, dass keines der vorgestellten Systeme in vollem Umfang dazu geeignet ist, die breit gefächerten Anforderungen im Rahmen des VDM auf Klimadaten zu unterstützen. Insbesondere die Anwendungsspezifik bei der Visualisierung, die starke Notwendigkeit zur Anwenderunterstützung bei Auswahl und Parametrisierung von visuellen und automatischen

[37] http://www.chem.agilent.com
[38] The MathWorks Inc., http://www.mathworks.com
[39] Wolfram Research Inc., http://www.wolfram.com
[40] SPSS Inc., http://www.spss.com

Methoden sowie die enge Kopplung statischer Methoden mit Visualisierungstechniken im Kontext der Simulation erfordern neue Vorgehensweisen, wie sie in dieser Arbeit bereitgestellt werden.

3.4 Design

Entwurf und Umsetzung eines VDM Systems mit einem hohen Grad an Nutzerunterstützung erfordern die Beachtung verschiedener Designkriterien. Dabei soll im folgenden auf grundlegenden Richtlinien beim Softwaredesign eines VDM Frameworks (Abs. 3.4.1), bei der Auswahl von Visualisierungstechniken (Abs. 3.4.2) und bei der Auswahl automatischer Mining-Methoden eingegangen werden (Abs. 3.4.3).

3.4.1 VDM Frameworkdesign

Eine Vielzahl von Arbeiten beschäftigt sich damit, wie gute Informationsvisualisierungssysteme entworfen werden sollten (vgl. z.B. Tang u. a. 2004; Fequete 2004; Heer u. Agrawala 2006). Neben allgemeinen Aussagen zu wichtigen Charakteristika solcher Systeme sowie Richtlinien bei deren Entwurf, die eine Anwenderakzeptanz erst ermöglichen (vgl. Wong 1999; Shneiderman 2002), finden sich auch spezielle Arbeiten, die grundlegende Modul- und Anwenderunterstützungsfunktionen diskutieren (vgl. z.B. Kreuseler u. a. 2004). Des Weiteren von Bedeutung zum Design von Visualisierungsframeworks sind geeignete Datenhaltungsstrategien, die es ermöglichen, die aktuell bearbeitete Menge an Daten zu einer verarbeitbaren Größe zu reduzieren ohne relevante Informationen zu verlieren. Beispiele für effektive Datenstrukturen auf Graphen geben z.B. Abello u. a. (2001) und Schulz u. a. (2006a), für multivariate Daten, Bäume und Grapen Fequete (2004) sowie zur Speicherung heterogener räumlich-zeitlicher Daten Ribarsky u. a. (2002a).

Darüber hinaus wurden Arbeiten mit grundlegenden Richtlinien beim Aufbau von VDM-Systemen vorgestellt. So schlagen Kreuseler u. Schumann (2002a) ein grundlegendes Informationsmodell vor und Keim u. a. (2002) identifizieren wichtige automatische Methoden, Visualisierungs-, Interaktionstechniken für das VDM.

Wie oben bereits angedeutet, gibt Shneiderman (2002) Richtlinien vor, um beim Bau von VDM-Systemen die Nutzerakzeptanz zu erhöhen. Dies schließt ein, dass

- der Nutzer spezifizieren kann, wonach er sucht und was er als interessant empfindet,
- eine flexible Anbindung externer Datenquellen gegeben ist,
- Nutzer interessante Ergebnisse einfach speichern und kommunizieren können,
- die menschliche „Verantwortung“ erhalten bleibt: übersichtliche Systeme, in denen die Nutzer befriedigend im VDM-Prozess interagieren, und sich selbst dafür verantwortlich fühlen können, ob die Analyse erfolgreich oder erfolglos ist; dies schließt insbesondere Vorhersagbarkeit und Einfachheit mit ein; Nutzer wollen Vertrauen in die Ergebnisse haben.

Auch Wong (1999) betrachtet das Design von VDM-Systemen aus einer anwenderzentrierten Sicht, betrachtet jedoch auch softwaretechnische und Datensicherheitskriterien:

- Einfachheit: VDM Systeme sollten leicht und intuitiv benutzbar sein, auch wenn die einzelnen Techniken komplexe Parameter haben,
- Nutzerautonomie: VDM Systeme sollten den Nutzer unterstützen, ihm aber weiterhin ermöglichen, die volle Kontrolle zu übernehmen,
- Verlässlichkeit: VDM Systeme sollten den Nutzer über Fehler und Genauigkeiten im VDM Prozess transparent informieren,
- Wiederverwendbarkeit: ein VDM System sollte für verschiedene Szenarien und Anwendungen adaptierbar sein,

- Verfügbarkeit: der Zugriff zu VDM Sytemen sollte ubiquitär sein. Dies schließt portable und verteilte Lösungen mit ein,
- Sicherheit: VDM Systeme sollten die Daten und das Wissen im System gegen unautorisierten Zugriff schützen.

Darüber hinaus gibt es spezielle Arbeiten die sich auf das Design von VDM-Systemen für spezielle Datenklassen konzentrieren (vgl. z.B. Schulz u. a. 2006a, VDM auf Graphen) oder auf spezielle Aspekte wie Historienkonzepte zur Verwaltung von Analyseverläufen im VDM (vgl. Kreuseler u. a. 2004) abzielen.

3.4.2 Visualisierungsdesign

Die Erzeugung geeigneter visueller Repräsentationen setzt umfangreiches Expertenwissen voraus (vgl. z.B. Fujishiro u. a. 1997). Visualisierungswissen in diesem Umfang kann man nicht von Benutzern fordern, die Experten auf anderen Gebieten sind und Visualisierung nur als Hilfsmittel einsetzen wollen. Die Folge ist, dass das Potential existierender Visualisierungssysteme nicht ausgeschöpft wird. So werden etwa fehlerhafte Darstellungen erzeugt, die die Dateninterpretation erschweren oder verfälschen (vgl. Beispiele in Jung 1998; Schumann u. Müller 2000), oder leistungsfähige Visualisierungstechniken kommen wegen mangelnder Vertrautheit nicht zur Anwendung (Kobsa 2001). Diese Beobachtungen haben dazu beigetragen, dass der menschliche Faktor in der Visualisierung zunehmend Gegenstand der aktuellen Forschung geworden ist (Tory u. Möller 2004a). Hier lassen sich sowohl die verschiedenen Arbeiten zur Bewertung von visuellen Repräsentationen einordnen als auch die unterschiedlichen Ansätze zur Unterstützung eines Anwenders bei der Erzeugung geeigneter visueller Repräsentationen. Bei deren Erzeugung sind verschiedene Einflussfaktoren zu berücksichtigen. Die Vielzahl dieser Einflussfaktoren lässt keine geschlossene Lösung des Visualisierungsproblems zu. Vielmehr fokussieren heutige Ansätze zur systematischen Generierung von visuellen Repräsentationen auf bestimmte Faktoren und vernachlässigen dabei andere Faktoren. Die meisten dieser Ansätze berücksichtigen die Eigenschaften der darzustellenden Daten sowie konkrete Ziele, die ein Anwender mit der Visualisierung verfolgt (vgl. z.B. Mackinlay 1986; Senay u. Ignatius 1994; Roth u. a. 1996; Fujishiro u. a. 2000; Zhou u. a. 2002b). Weniger auf die Visualisierung bezogen, sondern allgemein im Umfeld der Entwicklung Multiple User Interfaces (vgl. z.B. McGrenere u. a. 2002) werden zudem die zur Verfügung stehenden Ressourcen betrachtet. Auch die Wahrnehmungsfähigkeiten eines Anwenders werden eher selten berücksichtigt (Rushmeier u. a. 1997), wenn dann hauptsächlich im Hinblick auf den Einsatz geeigneter Farbskalen (Bergmann u. a. 1995). Daneben gibt es Ansätze, die ganz spezielle Fragestellungen betrachten, wie z.B. Strategien zur Visualisierung nominaler Werte (Rosario u. a. 2004) oder die geeignete Parametrisierung von Ikonen (Ward 2002). Insgesamt gilt aber, dass die Problematik eines systematischen Visualisierungsdesigns in der gängigen Literatur unterrepräsentiert ist.

Das hat zur Folge, dass Visualisierungen die Aufgaben in der Anwendung oft nicht ausreichend unterstützen („...successful decision-making and analysis are more a matter of serendipity and user experience than of intentional design and specific support for such tasks...“, Amar u. Stasko (2004)). Deshalb wurden in Keim u. a. (2005) als eine der zehn wichtigsten Herausforderungen auf dem jungen Forschungsgebiet Visual Analytics (vgl. Thomas 2005) die verstärkte Berücksichtigung der Semantik (Datencharakteristika, Anwenderziele u.a.) bei der Erzeugung von Visualisierungen genannt.

Der Begriff *Design* im Umfeld Visualisierung im Sinne einer geeigneten Abbildung der Daten in eine visuelle Repräsentation wurde von Mackinlay (1986) eingeführt. Der Begriff *Visualisierungsdesign* steht in der Literatur vorrangig für die Generierung, Auswahl und/oder Parametrisierung von geeigneten Visualisierungen. Darüber hinaus wird er allgemeiner auch im Sinne des Design eines Visualisierungssystems gebraucht (vgl. Abs. 3.4.1), was den oberen Aspekt dabei mit einschließt.

Tang u. a. (2004) sprechen in diesem Zusammenhang von Designentscheidungen bei der Definition eines internen Datenmodells, des Datenzugriffs und von Datentransformationen, sowie von semantisch angereicherten Meta-Informationen um effektive visuelle Kodierungen durchführen zu können. In diesem Abschnitt sollen kurz die wichtigsten Arbeiten in diesem Umfeld skizziert werden (vgl. dazu auch Lange u. a. 2006). Eine ausführlichere Beschreibung der Arbeiten zum Visualisierungsdesign findet sich in einer parallel laufenden Dissertation (Lange 2006).

Das Visualisierungsdesign stellt allgemein die Frage nach der Generierung von Visualisierungen mit hoher Qualität. Grundlegend für die Evaluierung von Visualisierungen sind die Kriterien

- Expressivität: eine Visualisierung stellt alle in den Daten enthaltenen Informationen und nur diese dar, und
- Effektivität: eine Visualisierung stellt die enthaltenen Informationen intuitiv dar, d.h. leicht verständlich und schnell wahrnehmbar,

wie sie in Mackinlay (1986) eingeführt wurden. Zusätzlich wurde in der Literatur das Kriterium der

- Angemessenheit: eine Visualisierung ist angemessen, wenn der Rechen- und Ressourcenaufwand in einem geeigneten Verhältnis zum Nutzen der Visualisierung steht

eingeführt (siehe z.B. Schumann u. Müller (2000)). Ziel einer „guten“ Visualisierung (und auch eines „guten“ Visualisierungssystems) ist es demnach, expressive und effektive Visualisierungen mit einem angemessenen Aufwand zu erzeugen. Dies hängt von **Gegenstand**, **Ziel** und **Kontext** der Visualisierung ab. Der Gegenstand der Visualisierung - und damit auch der wichtigste Faktor für die Visualisierung - sind die Daten, die durch die ihnen innewohnenden **Datencharakteristika** beschrieben werden können. Diese werden auch als **Metadaten** bezeichnet (siehe Abs. 3.4.2.1). Der zweite wichtige Faktor sind die **Ziele**, die der Nutzer mit der Visualisierung der Daten verfolgt, bzw. die **Aufgaben**, die er mit den Daten durchführen möchte (siehe Abs. 3.4.2.2). Der **Kontext** (siehe Abs. 3.4.2.3) beinhaltet die verfügbaren Resourcen, den Anwendungshintergrund, die Präferenzen des Nutzers und die menschlichen Wahrnehmungsfähigkeiten.

Weiterhin führt Tufte (1983) den Begriff der „graphischen Exzellenz“ (engl. graphical excellence) ein, um ein gutes Visualisierungsdesign auszumachen. Darunter fallen Klarheit, Genauigkeit, Effizienz und die Vermeidung von ungenutztem Platz innerhalb der Darstellung.

3.4.2.1 Metadaten

Metadaten sind definiert als Daten über Daten, beinhalten also jede Art von Anreicherung einer Datenmenge durch weitere Informationen. Grundlegend wird in diesem Zusammenhang zwischen beschreibenden, abgeleiteten und historischen Metadaten unterschieden (vgl. Robertson u. Hutchins 1997). Beschreibende Metadaten legen die grundlegende Struktur der Daten fest. Abgeleitete Metadaten können durch jede Art von (semi-)automatischen Verfahren berechnet werden und historische Metadaten beinhalten insbesondere Informationen zur Erhebung der Daten, z.B. zu Ort, Zeit und Qualität der erhobenen Daten. Grundlegende Notationen (z.B. Brodlie 1992; Wong u. Bergeron 1997) beschreiben kompakt wichtige Metadaten, die auch zur Unterscheidung verschiedener Datenklassen wie Multiparameter- oder Volumendaten geeignet sind. Sie schließen Anzahl und wichtige Eigenschaften unabhängiger und abhängiger Variablen ein. Allgemeine, teilweise durch den Entwickler erweiterbare Spezifikationen von Metadaten finden sich in speziellen Datenformaten (z.B. NetCDF, HDF) oder in allgemeingültigen Visualisierungssystemen (z.B. OpenDX, AVS). Auch diese Datenbeschreibungen sind auf eine bestimmte Art und/oder Menge darstellbarer Metadaten beschränkt, da sie auf eine spezielle Anwendung hin optimiert wurden. Um dieser Beschränkung zu begegnen, wurde in Nocke u. Schumann (2002) eine allgemeine, erweiterbare Metadatenspezifikation vorgestellt, die die vielfältigen Entscheidungen im Rahmen des Designs des gesamten

Visualisierungsprozesses unterstützt. Weitergehende Untersuchung zu diesem Thema findet sich in Abs. 7.1.

3.4.2.2 Ziele und Aufgaben

Ein wichtiges Instrument zur Steuerung des Nutzerinteresses sind die Ziele der Visualisierung bzw. die Aufgaben, die im Rahmen der visuellen Analyse ausgeführt werden sollen. Ein Problem bei der Spezifikation solcher Ziele sind die unterschiedlichen Begriffsbedeutungen in verschiedenen Nutzerkontexten. Entsprechend wurde eine Vielzahl von Ansätzen zur Beschreibung solcher Zielstellungen aus unterschiedlichen Perspektiven eingeführt (vgl. z.B. Roth u. Mattis 1990; Wehrend u. Lewis 1990; Robertson 1990). Der bekannteste ist die Task-by-Data-Type-Taxonomie von Shneiderman (1996), die verschiedene Datenklassen (z.B. 1D, 2D oder hierarchische Daten) mit Aufgaben (z.B. Überblick und Details on Demand) in Beziehung setzt. Eine systematische Integration dieser teilweise sehr heterogenen Aspekte ist deswegen erforderlich (vgl. Nocke u. Schumann (2004) und Abs. 7.2).

3.4.2.3 Kontext

Der Kontext einer Visualisierung beinhaltet die verfügbaren Resourcen, den Anwendungshintergrund, die Präferenzen des Nutzers und die menschlichen Wahrnehmungsfähigkeiten (vgl. Schumann u. Müller 2000). Für die hier vorgelegte Arbeit ist vor allem der **Anwendungshintergrund** von Interesse. Dieser wird vor allem durch kultur- und berufsspezifische Konventionen und Erfahrungen konstituiert (vgl. z.B. Nemcsics 1993, für die kontextabhängige Interpretation von Farben).

Im Anwendungsumfeld der Klimaforschung gibt es eine Vielzahl von Konventionen, die u.a. Abbildungsregeln für bestimmte Datenklassen und Merkmale beinhalten (vgl. Abs. 3.2.2.2). Die Literatur zu diesem Umfeld stellt eine Vielzahl von Regeln zur Farbarbbildung bereit (vgl. American Meteorological Society (1993)). Eine durchgehende Formalisierung solcher Regeln steht bisher jedoch noch aus (vgl. hierzu Abs. 7.3).

3.4.2.4 Erzeugung geeigneter Visualisierungen

Im folgenden sollen Ansätze zusammengefasst werden, die den Anwender dabei unterstützen, für seine Daten und für seine Ziele in einem gegebenen Kontext adäquate Visualisierungen zu generieren, auszuwählen und/oder entsprechend zu parametrisieren. Hierzu gibt es verschiedene Ansätze, die sich dadurch unterscheiden, ob sie auf eine Automatisierung des gesamten Visualisierungsprozesses ausgerichtet sind oder nur einzelne Schritte herausgreifen.

Die Pionierarbeit zur Automatisierung des Entwurfs graphischer Repräsentationen geht auf Mackinlay (1986) zurück. Hier werden graphische Repräsentationen als Sätze graphischer Sprachen mit Syntax und Semantik zur Beschreibung visueller Attribute definiert (vgl. Bertin 1983) und über eine Kompositionsalgebra verbunden. Das automatische Präsentationssystem APT (A Presentation Tool) geht dabei in drei Schritten vor:

1. Partitionierung: Die Datenmenge wird rekursiv unterteilt, bis jede Teilmenge durch wenigstens eine der graphischen Sprachen ausgedrückt werden kann.
2. Selektion: Für jede Datenteilmenge wird die effektivste der graphischen Sprachen bestimmt, die in der Lage ist, diese Datenteilmenge auszudrücken (anhand von datentyp- und wahrnehmungs-orientierten Rankings).
3. Komposition: Die Entwürfe für die Datenteilmengen werden mit Hilfe der Kompositionsoperatoren zu einer einheitlichen Repräsentation der gesamten Datenmenge zusammengesetzt.

Die Idee eines wissensbasierten Systems für den automatischen Visualisierungsentwurf wurde in zahlreichen nachfolgenden Arbeiten aufgegriffen und weiterentwickelt. Diese Verfahren können nach den folgenden Kriterien klassifiziert werden:

- Abdeckung
  - Unterstützte Schritte im Visualisierungsprozess
  - Unterstützte Repräsentationen
  - Unterstützte Datenklassen
- Einbindung verschiedener Einflussfaktoren (Metadaten, Ziele & Aufgaben, Visualisierungskontext)
- Abgleich (Matching) der Einflußfaktoren und der Visualisierung
  - Ausführung (automatisch vs. (semi-)automatisch)
  - Vorgehen (buttom-up vs. top-down)
  - Mechanismus (z.B. regelbasiert)
  - Grundlage (z.B. Kompositionsalgebra oder Fuzzy-Logik)

Insbesondere kann im Umfeld des Visualisierungsdesigns zwischen konstruktiven bottom-up Verfahren unter Einsatz von Kompositionsalgebren (vgl. z.B. Mackinlay 1986; Casner 1991; Roth u. Mattis 1991; Senay u. Ignatius 1994; Roth u. a. 1996; Jung 1998; Wilkinson 1999) und zwischen template-basierten top-down Verfahren (vgl. z.B. Gnanagmari 1981; Zhou u. a. 2002b; Jiawei u. a. 2004) unterschieden werden (vgl. Kamps 1999).

Hierbei basieren alle **konstruktiven Ansätze** auf dem regelbasierten Vorgehen von Mackinlay (1986). Sie unterscheiden sich in der Komplexität der graphischen Elemente, wobei das Spektrum von primitiven graphischen Objekten wie Punkten oder Linien und deren Eigenschaften über einfache Darstellungstechniken wie Scatterplots oder Liniendiagramme bis hin zu relativ komplexen grundlegenden Visualisierungstechniken wie etwa Kreissignaturenkarten reicht. Sie sind aber immer in dem Sinne „primitiv“, dass sie jeweils nur sehr wenige Datenvariablen kodieren und erst nach Komposition mit anderen primitiven Visualisierungen zur Darstellung multi-variater Daten geeignet sind. Konstruktive Ansätze haben den Vorteil, allgemein und dynamisch auch neue Visualisierungstechniken für einen Kontext zu erzeugen, an die ein Entwickler evtl. nicht gedacht hat. Sie können sehr flexibel gesteuert und erzeugt werden. Allerdings wird bei komplexeren Anforderungen des Benutzers auch eine komplexere Problembeschreibung durch den Nutzer erforderlich, die ihn überfordern kann. Weil das konstruierende Vorgehen beim Entwurf einer Visualisierung auch ein konstruierendes Vorgehen bei der Realisierung des Entwurfs erfordert, muss die Visualisierungsfunktionalität bei diesen Ansätzen neu implementiert werden. Dies kann bei existierenden hochkomplexen Visualisierungen mit langen Entwicklungszeiten ein Nachteil sein. Insbesondere besteht hier das Problem, auch die verfeinerten Interaktions- und Navigationstechniken bei solchen Ansätzen mit zu konstruieren. Ein weiteres Problem der konstruktiven Ansätze ist ihre feste Abhängigkeit vom Daten(bank)schema und von gewissen Metadaten.

Im Unterschied zu den konstruktiven Ansätzen gehen die **template-basierten Verfahren** von komplexen Visualisierungstechniken aus. Unter anderem ermöglicht dies, maßgeschneiderte Visualisierungen für bestimmte Anwendungsfelder zu entwickeln und die Anwender bei deren Auswahl und Parametrisierung zu unterstützen. Dabei können komplexe Visualisierungstechniken mit beliebigen Interaktionstechniken angesteuert und auch aus variierenden Entwicklungsumgebungen miteinander kombiniert werden. Nachteil dieser Vorgehensweise ist, dass keine beliebigen Visualisierungen erzeugbar sind und dass ein Teil der Güte des Visualisierungsdesigns beim Entwickler liegt. Ein Vorteil hierbei ist, dass die Techniken u.a. für die Verarbeitung von sehr großen Datenmengen leichter optimiert werden können. Je nach Art der Beschreibung der template-basierten Techniken kann zwischen graph-basiertem und vektor-basiertem Vorgehen unterschieden werden.

- **Graph-basiertes Vorgehen:** Ein Beispiel für das graph-basierte Vorgehen findet sich in

Zhou u. a. (2002b). Der Visualisierungsentwurf erfolgt hier anhand von Beispielen unter Verwendung von Methoden des fallbasierten maschinellen Lernens. Dazu wurde eine Datenbank mit einer großen, repräsentativen Menge guter Visualisierungsbeispiele gefüllt, wobei jedes dieser Beispiele durch einen Graphen repräsentiert wird, der die Eigenschaften der dargestellten Daten, der verwendeten visuellen Repräsentation und der Zuordnung von Datenelementen zu Visualisierungselementen (Mapping) enthält. Zu diesen Eigenschaften gehören u.a. auch sinnvolle Parametereinstellungen, Prioritäten und Präsentationsziele. Der Nutzer spezifiziert sein Visualisierungsproblem, indem er einen ebensolchen Graphen oder einen Teil davon aufstellt und damit Daten, Ziele und/oder Entwurfsvorgaben festlegt (Request). Das System sucht in einer Phase des „Skizzierens“ (Sketch Generation) aus der Datenbank durch Vergleich der entsprechenden Graphen diejenigen Beispiele heraus, die der Problemspezifikation am nächsten kommen. Nur wenn kein geeignetes Beispiel gefunden werden kann, wird eine Dekomposition in Request-Fragmente vorgenommen und das Skizzieren für diese erneut ausgeführt. Die Komposition der Fragmentskizzen wird von anderen Beispielen abgeleitet („gelernt“). Anhand der generierten Skizze wird in einer Realisierungsphase (Sketch Realization) eine analoge Visualisierung der gegebenen Datenmenge erzeugt, wobei die Feinheiten (etwa die Parameterwahl) wiederum von den Beispielen gelernt werden.

- **Vektor-basiertes Vorgehen:** Um die Eigenschaften einer (template-basierten) Visualisierungstechnik zu beschreiben, kombinieren Jiawei u. a. (2004) unterstützte Aufgaben, Metadaten und die Art des Mappings. Diese Informationen werden dann in einen „Feature“-Vektor mit binären Komponenten gespeichert. Als Charakterisierung der Visualisierungstechnik werden verschiedene visuelle „Features“ benutzt. Diese umfassen verschiedene *visuelle Anordnungen* und *strukturelle Muster*, mit der das Mapping die Beziehungen in den Daten darstellt. Hierzu gehören z.B. Anfügen (attach), Einschließen (enclosure), Zusammenfügen (composition), Verbinden (connect), Nähe (proximity), Sequenz (sequence), Überlappen (overlap), Übereinanderlegen (overlay), Nachbarschaft (adjacent), Ausrichtung (alignment) und Parallelität (parallel). Ausserdem werden auch *Informationstypen* wie Netzwerk, Hierarchie und Gitter betrachtet, sowie die *Art der Technik* wie z.B. Tabelle, Diagramm oder Karte. Weitere Eigenschaften einer Technik sind z.B. die Unterstützung von Verzerrungen und von visuellem Filtering. Die „Feature“-Vektoren werden in einer Datenbasis abgespeichert. Basierend auf dieser Visualisierungtechnik-Datenbasis kann der Nutzer Anfragen definieren, die intern in einen „Feature“-Vektor umgesetzt und dann mittels eines einfachen Ähnlichkeitsmaßes Datenbasisvektoren und Anfragevektor mit einander verglichen werden. Eine Eigenschaft dieses Maßes ist, möglichst viele angeforderte und möglichst wenige nicht benötigte Eigenschaften („Features“) zu erhalten.

3.4.2.5 Unterstützung einzelner Entscheidungen

Die vorgestellten Vorgehensweisen (top-down und bottom-up) unterstützen den gesamten Visualisierungsprozess. Dabei müssen eine Vielzahl von Einzelentscheidungen getroffen werden. Ein breites Spektrum von Arbeiten widmet sich dem Problemkreis, wie der Nutzer beim Treffen dieser Entscheidungen unterstützt werden kann. Insbesondere schließt dies die Auswahl, Kombination und Parametrisierung von Visualisierungstechniken ein.

Auswahl geeigneter Techniken. Zur Steuerung der Auswahl von Techniken werden vor allem Taxonomien eingesetzt, welche eine problemorientierte Klassifikation von Visualisierungstechniken nach den Einflussfaktoren beinhalten. Diese Taxonomien weisen aus, inwieweit Visualisierungstechniken bestimmte Datencharakteristika und Ziele unterstützen, und schließen teilweise auch den Visualisierungskontext mit ein. Eine Pionierarbeit bei der taxonomiegesteuerten Auswahl von Visualisierungstechniken ist (Wehrend u. Lewis 1990), bei der Visualisierungstechniken nach der Art der beteiligten Objekte (Datencharakteristik) und Operationen (Zielen) eingeteilt werden. Weiter-

entwicklungen dieses Ansatzes finden sich z.B. in Fujishiro u. a. (1997), Fujishiro u. a. (2000) und Jiawei u. a. (2004). Die Auswahl von Visualisierungstechniken auf der Basis von Taxonomien ist sehr intuitiv und wird daher gut akzeptiert. Auf der anderen Seite erfordert dieser Ansatz oft die Zerlegung von komplexen Visualisierungsproblemen in handhabbare Teilprobleme (z.B. wenn mehrere Zielstellungen gleichzeitig verfolgt oder verschiedene Datenobjekte gleichzeitig visualisiert werden sollen). Diese Zerlegung ist nicht immer einfach. Deshalb ist auch hier eine Nutzerunterstützung sinnvoll (vgl. z.B. North u. a. 2002).

Parametrisierung von Techniken. Nach der Auswahl einer geeigneten Visualisierungstechnik ergeben sich nun vor allem folgende Fragen:

1. Welche Datenvariablen sollen durch welche visuellen Eigenschaften der Darstellung repräsentiert werden?
2. Welche Ausprägungen einer visuellen Eigenschaft der Darstellung sollen welche Datenwerte kodieren?
3. Wie sollen weitere Parameter belegt werden?

Die erste Frage betrifft die Zuordnung von Datenvariablen zu visuellen Attributen wie Farbe, Position und Form. Legt die Visualisierungstechnik eine Reihenfolge der Visualisierungsvariablen fest (z.B. bei Parallelen Koordinaten eine Achsenreihenfolge) so muss dementsprechend auch die „Abbildungsreihenfolge“ der Datenvariablen zugeordnet werden. Arbeiten zur Parametrisierung von Visualisierungstechniken sind z.B. die automatische Zuordnung zu den Attributen einer natürlichen Szene (Robertson 1990) oder die Zuordnung von Variablen zu Attributen der hierarchischen Technik „worlds within worlds“ (Beshers u. Feiner 1993). Weiter gehören hierzu auch die Arbeiten zur Parametrisierung von Ikonen (Ward 2002). Die zweite Frage beinhaltet die Abbildung des Wertebereiches einer Datenvariablen auf den Wertebereich einer Visualisierungsvariablen. Hierbei wurde insbesondere die Farbabbildung in verschiedenen Arbeiten untersucht (z.B. Bergmann u. a. 1995; Robertson u. Hutchins 1997; Alexa u. Müller 1999; Schulze-Wollgast u. a. 2005), da sie die am häufigsten eingesetzte Abbildung und aufgrund der sensitiven menschlichen Wahrnehmungsfähigkeiten und Bedeutungszuweisungen in verschiedenen Anwendungen ein sehr „sensibler“ Prozess ist.

3.4.2.6 Evaluation

Eine wichtige Methode zur Bewertung visueller Repräsentationen ist die Durchführung von Nutzerstudien. Diese gewinnen wegen ihres Potentials und in Ermangelung gleichwertiger Alternativen zunehmend an Bedeutung (Kosara u. a. 2003; Jiawei 2003; Tory u. Möller 2004a). Hierbei lösen verschiedene Testpersonen mit verschiedenen Visualisierungen verschiedene Aufgaben. Dabei werden Bearbeitungszeit und Korrektheit der Ergebnisse gemessen. Dies führt je nach Granularität der Untersuchungsgegenstände zu Rankings von visuellen Attributen (z.B. Nowell u. a. 2002), von Visualisierungstechniken (z.B. Jung 1998) oder von Visualisierungssystemen (z.B. Kobsa 2001). Je nach Fokus sind diese Rankings spezifisch für spezielle Aspekte, z.B. ausgewählte Dateneigenschaften (z.B. nominal vs. ordinal (Nowell u. a. 2002)) oder spezielle Aufgaben (z.B. Trend erkennen vs. Werte ablesen bei Jung 1998). Das heißt, es werden vordergründig einzelne oder wenige Faktoren getestet, die die Effektivität visueller Repräsentationen beeinflussen. Dies liegt in der Natur der Sache, da stets die Mehrzahl dieser Faktoren konstant gehalten werden muss, um die Wirkung der verbleibenden Faktoren ohne störende Einflüsse untersuchen zu können. Damit können aber auch nur Aussagen über einfache Zusammenhänge getroffen werden. Für die Beurteilung komplexer Zusammenhänge muss nach alternativen Wegen gesucht werden (Kosara u. a. 2003). Hierzu gehört vor allem auch das Nutzen von Expertenwissen, das an verschiedenen Stellen und unter verschiedenen Aspekten veröffentlicht wurde, wie z.B. in Tory u. Möller (2004a) zum Thema „Human Factors in Visualization“ oder in Ware (2000) über wahrnehmungsbasierte Entwurfsrichtlinien. Unabhängig davon ist der Stand der Forschung aber noch weit davon entfernt, objektive Maße, formale Modelle

oder gesicherte Berechnungsverfahren für die Qualitätsbestimmung visueller Repräsentationen bereitzustellen. Vielmehr existiert ein umfangreiches informelles Wissen über Visualisierungen im Allgemeinen und über deren Effektivität im Besonderen, welches in geeigneter Form verfügbar gemacht werden muss, z.B. durch die Entwicklung und Integration von Tools zur Anwenderunterstützung in bestehende Visualisierungssysteme. Hierzu gibt es verschiedene Ansätze, die im nächsten Abschnitt genauer besprochen werden sollen.

Eine Nutzerstudie im Umfeld der Visualisierung von Klimadaten findet sich bei Trafton u. a. (2002). Die Autoren untersuchen mittels Eye-Tracking am Beispiel von Klimadaten den Unterschied zwischen explizit und implizit dargestellter Information, wobei sie sowohl Klimaexperten als auch Laien in diesem Umfeld einbeziehen.

3.4.2.7 Diskussion

Eine Vielzahl von Arbeiten untersucht die Unterstützung des Anwenders bei der Generierung aussagekräftiger Bilder. Unabhängig von der gewählten Vorgehensweise wirft dabei die Automatisierung des Visualisierungsprozesses allgemeine zum Teil noch ungelöste Probleme auf. Zum Beispiel kann das Bestreben, dem Nutzer möglichst viel Arbeit abzunehmen, leicht dazu führen, dass er in seinen Bedürfnissen und Zielen übergangen wird und dadurch die Akzeptanz eines Hilfesystems sinkt (vgl. Jung 1998). Besonderes Augenmerk ist also darauf zu richten, dass er zwar unterstützt wird, den Grad der Automatisierung bzw. der Interaktivität aber selbst steuern kann (vgl. hierzu auch die Kriterien von Wong (1999) und Shneiderman (2002) aus Abs. 3.4.1).

Ein weiteres Problem ergibt sich bei der adäquaten Spezifikation des Visualisierungsproblems: Je weniger der Nutzer in den Entwurfsprozess eingreifen kann, desto besser muss die Beschreibung seines Problems sein, wenn eine geeignete Lösung gefunden werden soll. Problematisch ist auch, dass ein einheitliches Vorgehen für das Visualisierungsdesign - oder sogar eine Standardisierung von Metadaten, Zielen, Technikbeschreibungen, Auswahl- und Abbildungsregeln in diesem Umfeld - noch nicht abzusehen ist. Bisher ist das Visualisierungsdesign weitgehend auf die Generierung einzelner Repräsentationen bei Einsatz einer bestimmten Menge an Einflussfaktoren ausgerichtet, was die Notwendigkeit einer geeigneten Kombination von visuellen Repräsentationen nach sich zieht (vgl. Chen 2003). Auch die Abbildung allgemeiner Beschreibungen bzw. Vorgehensweisen für ein spezielles Anwendungsfeld ist bisher nicht ausreichend untersucht worden.

3.4.3 Data Mining Design

Analog zum Visualisierungsdesign ist auch die Auswahl und Parametrisierung von automatischen Mining Verfahren ein sensibler Prozess. Betrachtet man die Variation, die sich z.B. zwischen den Resultaten verschiedener Clusterverfahren mit verschiedenen Proximitätsmaßen ergibt, führt dies zu einer breiten Interpretationsspannbreite für die Anwender solcher Verfahren. Auf der anderen Seite hat der Nutzer beim Einsatz einzelner Verfahren keine Sicherheit darüber, ob alle relevanten Muster aufgedeckt wurden. Wichtige Kriterien zur Einschätzung von Mining-Verfahren sind die (Vorhersage-)Genauigkeit aber auch deren Recourcenverbrauch (Zeit und Speicher).

Eine Möglichkeit, den Anwender bei der Durchführung automatischer Data Mining Verfahren zu unterstützen, sind so genannte Meta-Lerner (Chan u. Stolfo 1997; Laudien 2000). Diese unterstützen die Auswahl von Mining-Verfahren anhand bereits durchgeführter Mining-Läufe durch Einsatz von Methoden des maschinellen Lernens. Hier lassen sich (analog zum Visualisierungsdesign) zwei grundsätzliche Vorgehensweisen identifizieren. Zum einen werden in der als Stacking bekannten Methode verschiedene (Mining-)Modelle miteinander kombiniert, um mehrere Wissensquellen geeignet zu koppeln (Chan u. Stolfo 1997). Zum anderen gibt es verschiedene Ansätze zur Wahl und/oder Ranking von Mining-Techniken (vgl. z.B. Berrer u. a. 2000; Brazdil u. Soares

1997). Beispiele zum Einsatz von Meta-Lernern zur Steuerung von Klassifikationen finden sich bei Brazdil u. Soares (1997) und Todorovski u. Dzeroski (2003), sowie zum Ranking von maschinellen Lernverfahren in Brazdil u. a. (2003).

3.5 Zusammenfassung

Zusammenfassend ist festzustellen, dass im Gebiet des VDM noch immer großer Forschungsbedarf besteht. Gerade bei der engen Kopplung von automatischen Verfahren und Visualisierung besteht ein hohes Potenzial, welches heute erst in Ansätzen ausgeschöpft wird. Shneiderman (2002) stellt in der Visualisierungs- und in der Data Mining-Literatur eine Unterrepräsentation des anderen Gebietes fest. Es ist jedoch zu konstatieren, dass diese Unterrepräsentation seit einigen Jahren abnimmt. Auch die Zahl existierender Systeme am Markt im Visualisierungs- und Data Mining-Umfeld hat stark zugenommen. Moderne Interaktionstechniken finden hier eine immer stärker werdende Verwendung.

Um die Nutzbarkeit solcher Analysetools auch für Laien zu sichern, bestehen noch vielfältige Herausforderungen. So bieten heutige Systeme punktuelle Lösungen für spezielle Entscheidungen im Visualisierungsprozess oder bei der Verwaltung von Analysesitzungen an. Einheitliche, für verschiedene Anwendungsgebiete erweiterbare Standards für Metadaten oder Nutzerziele, wie sie z.B. im Datenbankumfeld üblich sind, liegen nicht vor.

Weiterhin ist eine stärkere Kopplung von automatischen und visuellen Verfahren im Designprozess noch offenes Forschungsgebiet. Bisherige Ansätze fokussieren entweder auf die Automatisierung von Visualisierungs- oder von Data Mining Techniken. Inwieweit verschiedene automatische und visuelle Miningtechniken kontextabhängig generiert, ausgewählt, parametrisiert und z.B. in Form von Netzwerken miteinander gekoppelt werden können, ist noch weitgehend offen, und soll im neuen Forschungsfeld des Visual Analytics (Keim u. a. 2005) genauer untersucht werden.

Darüber hinaus besteht eine große Lücke zwischen allgemeingültigen und auf eine bestimmte Anwendung zugeschnittenen Systemen. Gerade im Klimaumfeld werden modernen Visualisierungs- und Interaktionstechniken selten eingesetzt. So werden in der Visualisierungsliteratur etablierte Techniken zur Darstellung multi-variater Daten ohne Raum- und Zeitbezug, zur Darstellung von zeitveränderlichen Daten und Techniken zur Darstellung multivariater Daten auf Karten unter Einsatz von Fokus & Kontext sowie Brushing & Linking in diesem Umfeld kaum eingesetzt. Ferner besteht ein Mangel an Visualisierungsmethoden zur Darstellung der Resultate statistischer Verfahren, die vor allem bei Daten mit Zeit- und/oder Raumbezug neue Vorgehensweisen erfordern. Um die Akzeptanz der Anwender für solche Techniken zu erhöhen, müssen speziell auf das Umfeld zugeschnittene Metadaten und Zielstellungen erhoben und verarbeitet werden. Neben einer passgenauen Darstellung der Daten nach den Anwenderzielen können so auch entsprechend den Konventionen des Gebietes parametrisierte Visualisierungen erzeugt werden.

Abschließend besteht eine große Potenz, die Fähigkeiten des VDM auf den gesamten Modellierungsprozess - nicht nur im Klimaumfeld - zu übertragen, um so die Stärken insbesondere statistischer Verfahren und von Visualisierungstechniken auszunutzen. Dazu muss untersucht werden, ob und wie bei der Modellierung, Simulation, Hypothesentestung und Modellevaluation die Methodik des VDM eingesetzt werden kann.

Kapitel 4

Konzeption und Entwurf eines softwaretechnischen Rahmens für das VDM von Klimadaten

Eine breite Palette allgemeingültiger Visualisierungs-, Data Mining- und VDM-Systeme steht einer Vielzahl von speziell für die Darstellung meteorologischer und klimatologischer Daten entworfener Systeme gegenüber. Während die einen moderne Visualisierung- und Interaktionstechniken enthalten, sind die anderen in ihrer Funktionalität stark eingeschränkt, können dadurch jedoch genau auf die Anforderungen im Umfeld der Anwendung fokussieren. Ferner ist vor allem in datenflussorientierten modularen Visualisierungssystemen, aber auch in vielen speziellen Systemen ein relativ hoher Einarbeitungsaufwand für den Anwender erforderlich. Dieser hohe Aufwand ermöglicht es jedoch, auch spezielle, auf das aktuelle Problem zugeschnittene Visualisierungen zu entwerfen.

Zur Lösung der genannten Probleme verfolgt die hier vorgelegte Arbeit den Ansatz, auf der einen Seite den Aufwand des Anwenders bei der Umsetzung und Durchführung des VDM weitgehend zu reduzieren, ihm auf der anderen Seite jedoch in Abhängigkeit von seinem Wissensstand den Zugriff auf Details des Analyseprozesses zu erlauben. Systeme, welche ihn in Abhängigkeit von seinen Zielstellungen und Wissensstand unterstützen, werden bisher im Umfeld der Klimaforschung nicht eingesetzt.

In diesem Kapitel soll deswegen ein softwaretechnischer Rahmen für auf spezielle Anwendungen zuschneidbare VDM-Frameworks entworfen werden. Dabei sollen die Vorteile allgemeiner Systeme genutzt und mit dem Wissen über das spezielle Umfeld verknüpft werden. Über bisherige Arbeiten hinaus sollen dabei Visualisierung und automatische Verfahren stark gekoppelt werden, und auch ihr Potential in der konfirmativen Analyse sowie zur Unterstützung von Modellbildung, Simulation und Evaluation untersucht werden.

Zunächst werden dafür die sich aus einer Kooperation mit dem Potsdam Institut für Klimafolgenforschung ergebenden Rahmenbedingungen konkretisiert und spezielle Anforderungen abgeleitet (vgl. Abs. 4.1), um darauf aufbauend eine allgemeingültige Komponentenbibliothek entwerfen zu können (vgl. Abs. 4.2). Diese Komponentenbibliothek schließt alle im Rahmen der Arbeit entworfenen Konzepte ein und gibt grundlegende Aspekte zu deren Umsetzung und Integration vor (vgl. Abs. 4.3). Sie bildet die Basis für die im Rahmen der Kooperation mit dem Potsdam Institut für Klimafolgenforschung entworfenen Systeme SimEnvVis und VisAna.

4.1 Anforderungen und allgemeine Vorgehensweise

In Anlehnung an die von Hege (1992), Wong (1999) und Shneiderman (2002) formulierten Kriterien zur Auswahl und Entwurf von Visualisierungs- und VDM-Systemen sollen in diesem Abschnitt wichtige Anforderungen für die Konzeption einer Komponentenbibliothek im Klimaumfeld gegeben werden. Die Kriterien werden im folgenden durch *kursive Schrift* gekennzeichnet.

Allgemeine Anforderungen. In einer Bibliothek von VDM-Verfahren für Klimadaten müssen universell einsetzbare VDM-Verfahren mit speziellen Anforderungen der Anwendung verknüpft werden (*Universalität vs. spezielle Anwendungen*), sowie GUIs und Visualisierungen bezüglich Normen des Anwendungsfelds (z.B. Begrifflichkeiten oder Farbskalen) angepasst werden (*Eignung für die Zielgruppe*).

Da zur Lösung einer Analyseaufgabenstellung oft ein einzelnes VDM-Verfahren nicht ausreicht, ist es notwendig, solche Verfahren flexibel zu verknüpfen. Neben dem Linking verschiedener (gleichzeitig ausgeführter) Verfahren[1] schließt dies auch ein, Mining-Ergebnisse eines Verfahrens in einem anderen weiterzuverarbeiten. Entsprechend können sich so auch Netzwerke von Mining-Verfahren[2] bilden, die geeignet verwaltet werden müssen.

Eine Vorgabe aus der Anwendung ist, dass die umgesetzten Verfahren eine *hohe Fehlertoleranz* bezüglich verschiedener Nutzereingaben (auch bei der Verknüpfung mehrerer Techniken), aber auch bezüglich variierender Eingabedaten haben sollen (*Qualität, Verlässlichkeit*). Zur Absicherung der Systemzuverlässlichkeit sollen insbesondere (beschreibende) Metadaten eingesetzt werden, die es dem System ermöglichen zu entscheiden, ob eine spezielle VDM-Technik auf den vorliegenden (Eingabe-)Daten ausgeführt werden kann oder nicht.

Den Nutzer betreffende Anforderungen. Um die hohe Einarbeitungszeit - speziell zur Erlangung von Visualisierungs- und graphischem Wissen - für die Klimaforscher zu reduzieren, sollen Techniken auf einem allgemeinen Level mit überschaubaren Parameterdialogen angeboten werden (*Komplexität der Bedienung*). Dies bedeutet, dass dem Anwender weitgehend die Arbeit mit komplexen Netzwerken oder Skriptsprachen, wie sie in modularen Visualisierungssystemen üblich sind, abgenommen wird (*Einarbeitungszeit*). Er soll lediglich auf einem sehr allgemeinen Level mit grundsätzliche Techniken und deren Verknüpfung interagieren, jedoch keine Kenntnisse von internen Datenstrukturen und Datenflüssen besitzen müssen.

Weiterhin für die Akzeptanz eines VDM-Systems wichtig sind die *Zeitanforderungen* des Nutzers angesichts großer Klimadatensätze, was sich sowohl auf die Ausführung automatischer Algorithmen als auch auf die Generierung interaktiver, visueller Darstellungen auswirkt. Hierbei sind insbesondere Verfahren mit interaktiven Antwortzeiten von langandauernden zu trennen, um eine hohe Benutzungsfreundlichkeit zu erreichen. So können dann eine Vielzahl moderner Interaktionstechniken zur Unterstützung des Analyseprozesses (*Interaktivität*) passgenau eingesetzt werden.

Um eine leichte Bedienbarkeit auch komplexer Techniken in komplexen Analyseprozessen zu ermöglichen, soll ferner ein hoher Grad an Nutzerunterstützung bereitgestellt werden (*Einfachheit*). Dies beinhaltet neben dem oben angedeuteten Verbergen innerer Systemdetails auch eine starke Unterstützung bei Wahl und Parametrisierung geeigneter VDM-Techniken. Der damit verbundene Design-Prozess beschleunigt die Findung von Mustern in den Daten (*Zeitanforderungen*) und erhöht die *Qualität* und *Verlässlichkeit* der Resultate des VDM-Prozesses, indem z.B. auch die Analyseziele des Anwenders - dem Vokabular der Anwendung entsprechend - explizit einbezogen werden. Dabei ist besonders wichtig, den Anwender nicht zu entmündigen (*Nutzerautonomie*). Die Unterstützung soll deswegen lediglich als Leitfaden dienen, so dass der Nutzer den automatisch vorgegebenen Weg

[1] Dies sind zumeist Visualisierungstechniken.

[2] Im folgenden sollen solche Netzwerke von VDM-Techniken in Anlehnung an Chi u. Riedl (1998) auch als Operatorgraphen bezeichnet werden.

jederzeit verlassen kann.

Neben der Auswahl und Parametrisierung von Techniken benötigt der Anwender weiterhin Unterstützung, den teilweise komplexen Analyseprozess bei der Verwendung mehrerer VDM-Techniken zu überschauen. Um ihn dabei zu unterstützen, kann eine Analyseprozessverwaltung eingesetzt werden, um erfolgreiche Analysen zu verwalten und getätigte Nuterinteraktionen zu rekapitulieren.

Anforderungen an die Umsetzung. Um eine hohe *Wiederverwendbarkeit* für verschiedene Szenarien zu erreichen, soll als Design eine adaptierbare Bibliothek gewählt werden, die es erlaubt, Techniken in verschiedene Systeme - so auch in SimEnvVis und in VisAna - einzufügen, und die um beliebige Methoden erweitert werden kann.

Eine weitere Anforderung der Anwender war es, die VDM-Verfahren auf einem zentralen Visualisierungsserver auszuführen zu können, der es verschiedenen Klimaforschern erlauben soll, ihre auf lokalen Rechnern verteilten Daten auszuwerten (*verteiltes Arbeiten, Verfügbarkeit*). Bezüglich der Produktumgebung soll die Komponentenbibliothek und die daraus abgeleiteten Systeme im Umfeld der am PIK üblichen Linux- und AIX-Workstations lauffähig (*vorhandene Hard- und Software*), und für verschiedene Versionen kompatibel sein (*Kompatibilität*).

Während Klimamodelle heute zumeist noch in Fortran umgesetzt werden, soll insbesondere die Software in einer objektorientierten Programmiersprache umgesetzt werden, um die Möglichkeiten moderner GUI- und Graphikprogrammierung nutzen zu können (*Modernität*). Dabei ist es wichtig, portable, flexible Standardbibliotheken bzw. -systeme für GUI, Datenhaltung und Visualisierung zu nutzen (*Portabilität, Standardnähe*).

Als Austauschformat steht das Standardformat NetCDF im Vordergrund (*Standardnähe, Datenformate*). Neben einer flexiblen Datenbeschreibung stellt NetCDF auch wichtige Metadaten bereit, die für das VDM benötigt werden. Darüber hinaus soll die Bibliothek auch für andere, im Klimaumfeld übliche Datenformate (z.B. Klimadaten im Grib-Format oder im ASCII-Format) offen sein (*flexible Anbindung externer Datenquellen*).

4.2 Entwurf einer Komponentenbibliothek

Nachdem im vorangegangen Abschnitt Anforderungen und Vorgaben aus der Anwendung ausgeführt und darauf aufbauend die grundlegende Vorgehensweise skizziert wurde, soll nun eine flexible, aus einem Baukastensystem bestehende Komponentenbibliothek entworfen werden. Neben der Spezifikation grundlegender Schnittstellen einer solchen Bibliothek soll auch der Datenaustausch der einzelnen Komponenten auf einem allgemeinen Abstraktionsniveau diskutiert werden. Abbildung 4.1 zeigt die grundlegende Struktur der Bibliothek.

Den Kern bildet eine **VDM-Technikbibliothek** bestehend aus Visualisierungstechniken und statistischen Verfahren, die wiederum auf eine Menge von **Basisalgorithmen** zurückgreifen, welche jedoch für den Anwender weitgehend unsichtbar sind. Die VDM-Techniken können gekoppelt und zu (einfachen) Operatorgraphen zusammengesetzt werden (**Verfahrenskopplung**). Um einzelne VDM-Techniken, aber auch ganze Operatorgraphen auswählen und parametrisieren zu können, enthält das System eine Komponente zum **VDM-Design**. Insbesondere greift diese Komponente auf **Analyseziele** und **Metadaten** zurück, die durch zwei separate Komponenten verwaltet werden. Um den Nutzer ferner bei der Verwaltung des Analyseprozesses und erfolgreicher Analysen zu unterstützen, kann eine **Analyseprozessmanagementkomponente** mit den anderen Komponenten verküpft werden.

Schnittstellen des Systems stellen die graphische Nutzerschnittstelle (GUI) und eine Vielzahl von Verfahren zum Datenimport und -export dar. Die Bibliothek - und die darauf aufbauenden Systeme SimEnvVis und VisAna - wurde(n) als Serverapplikation(en) entworfen, auf welche verschiedene

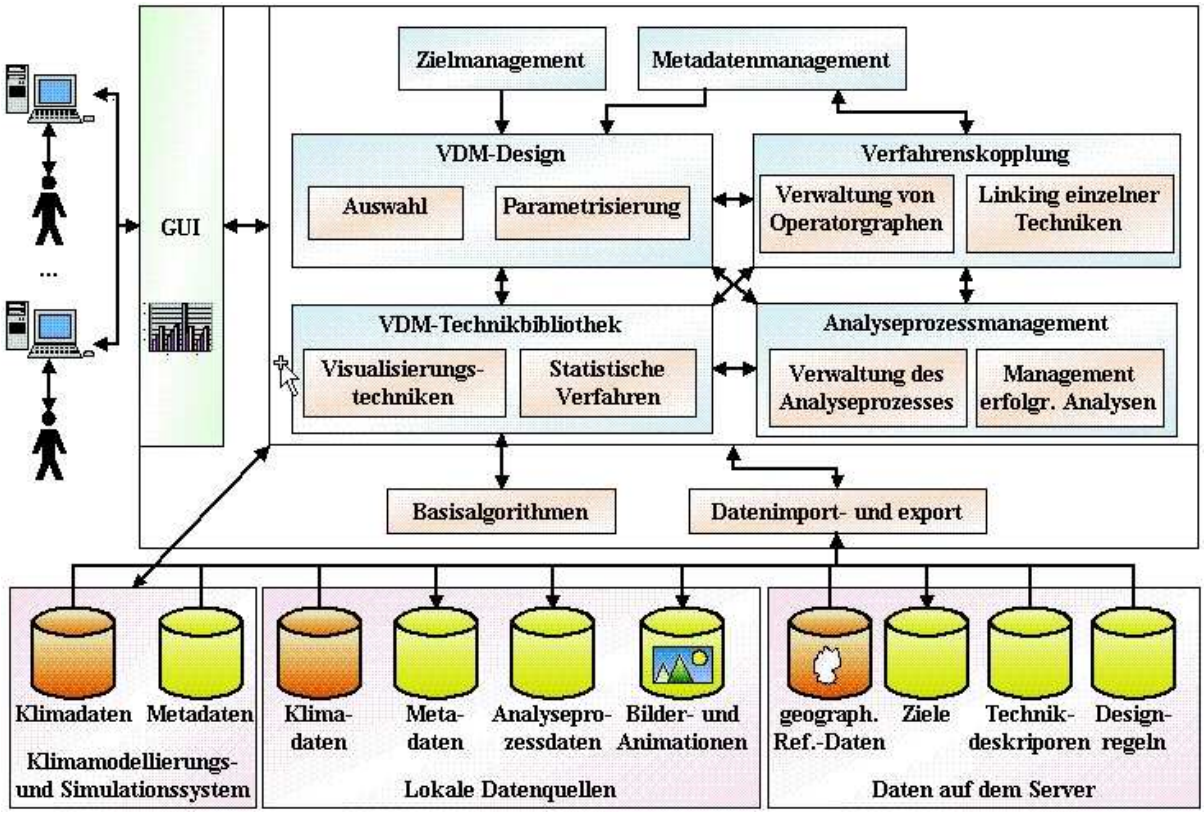


Abbildung 4.1: Grundlegende Komponenten einer VDM-Komponentenbibliothek für die Analyse von Klimadaten

Anwender von lokalen Workstations aus zugreifen können. Komplexe Prozesse können auf dem Server berechnet werden, während die Graphikausgabe durch die lokale Graphikhardware beschleunigt erfolgen kann.

Für den Datenimport und -export sind drei grundlegende Datenquellen bzw. -senken vorgesehen. Zum ersten werden **Daten**, die unabhängig von Analyseproblem und Anwender sind, **auf dem Server** gehalten. Dies sind geographische Referenzdaten (Ländergrenzen, Reliefinformationen, u.a.), allgemeine Zielstellungen sowie deren anwendungsspezifische Anpassungen, Technikdeskriptoren, welche die Eigenschaften einzelner Techniken - insbesondere zur Verarbeitung im VDM-Design - beschreiben sowie allgemeine und anwendungsspezifische Designregeln.

Zum zweiten kann jeder Anwender die für seine Untersuchungen speziellen Analysedaten lokal speichern. Bei diesen **lokalen Daten** handelt es sich um gemessene oder simulierte Klimadaten sowie um die sie beschreibenden Metadaten, um gespeicherte erfolgreiche Analyseprozesse sowie für die Kommunikation und Publikation geeignete Bilder und Animationen.

Zum dritten soll die hier entworfene Bibliothek auch erste Verfahren zum direkten Austausch von **Modellierungs- und Simulationssystemen** mit VDM-Verfahren entwerfen. Hierbei sollen Klimadaten und ihre Metadaten direkt in das VDM einbezogen werden, und das im VDM gewonnene Wissen soll direkt in die Modellierung und Simulation zurückfließen.

4.3 Eckpunkte für die Umsetzung der Komponentenbibliothek

In diesem Abschnitt sollen wichtige Entscheidungen zur Umsetzung der Komponentenbibliothek dokumentiert werden.

Visualisierungstechniken. OpenDX hat sich als flexibles Tool im Umfeld der Klimavisualisierung

etabliert (vgl. z.B. Treinish 1999; Kücken u. a. 1999) und wurde in Abstimmung mit den Kooparationspartnern vom PIK deswegen als Basistool zur Entwicklung von Visualisierungstechniken gewählt. OpenDX ist frei verfügbar und enthält insbesondere eine Vielzahl von Basisalgorithmen für die Darstellung von räumlichen Daten.

Um darüber hinaus auch Techniken einbinden zu können, die einen direkten Zugriff auf die Graphikhardware und auf interne Datenstrukturen ermöglichen, wurden weitere Darstellungstechniken in C++ mit der Graphikbibliothek OpenGL umgesetzt. Ferner wurde auch die Programmiersprache Java eingesetzt, die insbesondere für ein Rapid Prototyping von 2D-Visualisierungsanwendungen geeignet ist, um den Klimaforschern schnell erste Darstellungen präsentieren und diese auf ihr Potential in der Anwendung einschätzen zu lassen.

Statistische Techniken. Zum einen wurden im Rahmen der Arbeit durch die Kooperationspartner vom PIK evaluierte, in der Programmiersprache Fortran umgesetzte Clusteralgorithen eingebunden. Um diese integrieren zu können, wurden sie in C/C++ gewrapped. Vor allem aus Performance-Gründen wurden darüber hinausgehende Verfahren direkt in C++ und teilweise auch als Module in OpenDX umgesetzt.

Grafische Nutzerschnittstelle (GUI). Um den am PIK üblichen Betriebssystemen AIX und Linux Rechnung zu tragen, wurde die GUI mit der QT-Bibliothek[3] implementiert. Sie zeichnet sich durch die leichte Portierbarkeit auf andere Plattformen, durch ein ausgereiftes Botschaftenkonzept und eine große Klassenbibliothek aus. Insbesondere unterstützt sie auch die Graphikausgabe in 2D (QCanvas-Klasse) und 3D (Schnittstelle zu OpenGL).

Weitere Komponenten wurden als flexibel erweiterbare Module in der Programmiersprache C++ umgesetzt in der KDevelop-Projektumgebung programmiert und getestet.

4.4 Zusammenfassung

In diesem Kapitel wurde eine allgemeine Komponentenbibliothek für das VDM von Klimadaten entworfen. Die Bibliothek ist allgemeingültig konzipiert und hat das Potential, in verschiedenen Anwendungsfällen eingesetzt zu werden (vgl. die Systeme VisAna in Abs. 6.3 und SimEnvVis in Abs. 7.3). Es können je nach Bedarf einzelne Module ausgewählt und fallspezifisch verknüpft werden. In den folgenden Kapiteln sollen nun die Herausforderungen beim Entwurf der einzelnen Komponenten im Detail ausgeführt werden. Hierbei stellen sich die folgenden drei Probleme:

1. **Entwurf von Visualisierungsmethoden für Klimadaten**
   - **passgenaue Visualisierungen für den Anwendungshintergrund:** Schwierigkeit hierbei ist, im Klimaumfeld unbekannte Darstellungstechniken an diesen Anwendungshintergrund anzupassen, und dabei eine hohe Nutzerakzeptanz zu erreichen. Insbesondere ist dafür zu untersuchen, wie **passgenaue, intuitiv verständliche Metaphern** eingesetzt werden können.
   - **heterogene Klimadaten:** Klimadaten haben stark variierende Strukturen. Dies schließt z.B. lange Zeitreihen für einzelne Messstationen, Simulationen auf strukturierten räumlich-zeitlichen Gittern oder gestreute Messdaten ein. Dabei können fehlende und mit Fehlern behaftete Werte auftreten.
   - **vielfältige Interaktions- und Navigationstechniken:** Für eine flexible Untersuchung von Klimadaten sind eine Vielzahl von Interaktions- und Navigationtechniken erforderlich, welche bisher im Klimaumfeld in nur geringem Maße eingesetzt werden.
2. **Verknüpfung von Visualisierungstechniken und automatischen Verfahren:** Bei der engen Kopplung von statistischen Verfahren mit Visualisierungstechniken sind noch immer

[3] Trolltech: http://www.trolltech.com

viele Probleme ungelöst. Dies betrifft vor allem die wechselseitige Unterstützung in den verschiedenen Phasen des Analyseprozesses. Gerade bei räumlichen und zeitlichen Daten wurde diese Kopplung bisher kaum untersucht. In dieser Arbeit soll auf die drei folgenden Punkte vertiefend eingegangen werden:

- **Visualisierung und Clusteranalyse:** Die Clusteranalyse ist ein in der Klimaforschung häufig eingesetztes Verfahren, um durch die damit verbundene Datenaggregation einen erleichterten Überblick über große, multivariate Daten zu erhalten. Insbesondere die interaktive Visualisierung der Clustereigenschaften in ihrem räumlichen und zeitlichen Bezug wurde bisher kaum untersucht.
- **Visualisierung und Hauptkomponentenanalyse:** Um wichtige Trends in den Daten effektiv analysieren und leichter verstehen zu können, ist eine enge Verzahnung der Ergebnisse aus der Hauptkomponentenanalyse mit den einzelnen Schritten der Visualisierungspipeline erforderlich.
- **visuelle Unterstützung der Klimamodellbildung, -simulation und -evaluation:** Neben der reinen Auswertung simulierter oder gemessener Klimadaten können die Verfahren des Visuellen Data Minings auch eingesetzt werden, um den gesamten Modellbildungsprozess zu unterstützen. Eine Unterstützung zu diesem vielversprechenden Ansatz liegt bisher nicht vor.

3. **Unterstützung des Nutzers bei der Generierung von aussagekräftigen Bildern**
   - **Spezifikation des Analyseproblems:** Soll der Anwender dabei unterstützt werden, aus der Menge möglicher Darstellungen geeignete auszuwählen, müssen die Einflussfaktoren auf die Visualisierung explizit vorliegen. Häufig liegen diese aber nur rudimentär vor. Entsprechend muss er dabei unterstützt werden, Metadaten und Analyseziele zu spezifizieren und zu verwalten. Insbesondere muss dabei die Spezifik des Anwendungsumfeldes einbezogen werden.
   - **Generierung aussagekräftiger Darstellungen:** Entsprechend der Spezifikation des Analyseproblems muss der Anwender dann dabei unterstützt werden, geeignete Bilder für seinen Problemkontext zu erzeugen. Auch hierbei muss der Nutzer einbezogen werden, denn *ein* automatisch erzeugtes Bild ist häufig nicht ausreichend.

Entsprechend liegt der erste Schwerpunkt der Arbeit im **Entwurf intuitiver, interaktiver Visualisierungstechniken** für die Anwendung. Deshalb soll auf die Komponente *Visualisierungstechniken* im besonderen in Kapitel 5 eingegangen werden. Zweiter Schwerpunkt der Arbeit ist die Enge **Verzahnung von Visualisierung und automatischen Verfahren**. Dieser Schwerpunkt wird in Kapitel 6 ausgeführt. Hierzu werden Visualisierungstechniken und statistische Methoden auf ihre Kombinierbarkeit im Visuellen Data Mining sowie auf ihre Potenz, im Modellierungs- und Simulationsprozess als unterstützendes Werkzeug eingesetzt zu werden, untersucht. Dritter Schwerpunkt dieser Arbeit ist die **Unterstützung des Anwenders bei der Spezifikation des Analyseproblems sowie der Findung geeigneter Visualisierungstechniken**. Entsprechend sollen die zugehörigen Komponenten Metadaten- und Zielmanagement sowie Visualisierungsdesign in Kapitel 7 diskutiert werden. Zu speziellen Aspekten der Verfahrenskopplung und des Analyseprozessmanagement, der allgemeinen Interaktion der verschiedenen Komponenten sowie zu softwaretechnischen Aspekten bei der Umsetzung sei hier auch auf Nocke u. a. (2003), Kreuseler u. a. (2004) und auf Schmidt (2004) verwiesen.

Kapitel 5

Visualisierung von Klimadaten

In diesem Kapitel soll nun als erster Schwerpunkt dieser Arbeit der Entwurf von Visualisierungsmethoden für Klimadaten diskutiert werden. Dafür wurden bekannte Visualisierungstechniken an die spezifischen Anforderungen der Anwendung angepasst und in deren technischen und semantischen Hintergrund eingebettet, sowie neue Techniken entworfen und umgesetzt.

Für die Visualisierung von Klimadaten ergeben sich dabei die folgenden vier Schwerpunkte:

1. **Visualisierung im räumlichen Bezug:** Hierbei steht die Untersuchung von örtlichen Werteverteilungen in Klimadaten im Vordergrund (insb. Identifikation und Lokalisation von Mustern, z.B. extreme Witterungsbedingungen).
2. **Visualisierung im zeitlichen Bezug:** Eine wichtige Aufgabe ist die Untersuchung von Trends und besonderen Ereignissen in Klimadaten (z.B. langfristige Temperaturbedingungen).
3. **Visualisierung im Merkmalsraum:** Darstellungen, die vom räumlichen und zeitlichen Bezug der Klimadaten abstrahieren, werden bisher kaum eingesetzt. Entsprechend lassen sich verschiendene, aus der Visualisierungsliteratur bekannte Standardtechniken (z.B. Parallele Koordinaten oder Scatterplot-Matrizen) zur Untersuchung des durch die abhängigen Variablen definierten Merkmalsraum verwenden.
4. **Vergleichende Visualisierung:** Der Vergleich mehrerer Datensätze über Raum und Zeit ist ein schwieriges Problem. Vor allem gilt dies für Datensätze mit abweichenden Gittern, wie sie auch im Klimaumfeld auftreten. Dadurch können grundlegende Abhängigkeiten, Verteilungen und Gruppierungen in Klimadaten aufgedeckt werden, die gerade in räumlichen Darstellungen nur schwer identifizierbar sind.

5.1 Darstellung von Klimadaten im räumlichen Bezug

Die Darstellung räumlicher Daten hat eine lange Tradition in der Forschung zur Visualisierung. So wurden Geographische Informationssysteme (GIS), Volumenvisualisierungen und multi-variate Darstellungen intensiv untersucht. Trotzdem gibt es in diesem Umfeld noch immer vielfältige Herausforderungen, welche auch das Gebiet der Klimadatenvisualisierung betreffen.

In diesem Abschnitt werden zuerst Anspruch und Probleme bei der Darstellung von Klimadaten mit räumlichem Bezug diskutiert (vgl. Abs. 5.1.1). Zur Lösung dieser Probleme bedarf es einer breiten Palette von Standardtechniken (vgl. Abs. 5.1.2), welche auf die Bedürfnisse des Anwendungsumfeldes angepasst, in einem Visualisierungssystem bereitgestellt und geeignet parametrisiert werden müssen. Daran anschließend wird eine neue Technik zur metapherbasierten Ikonendarstellung vorgestellt (vgl. Abs. 5.1.3). Abschließend werden die Eigenschaften der vorgestellten Visualisierungstechniken zusammengefasst (vgl. Abs. 5.1.4).

5.1.1 Anspruch und Probleme

Bei der Darstellung von Daten im räumlichen Bezug ergeben sich eine Vielfalt von Zielstellungen. Neben der **Untersuchung der Werteverteilung** im Raum ist es bei Multiparameterdaten nützlich, mehrere Merkmale gleichzeitig anzeigen zu können, um deren **Abhängigkeiten** in ihrem räumlichen Bezug explorieren (vgl. auch Treinish 1999) und sich dabei herausbildende räumliche Muster untersuchen und ggf. verfolgen zu können. Je komplexer solche Abhängigkeiten werden, umso schwerer wird es für den Nutzer, die Darstellung zu erfassen. Viele komplexere Darstellungsmethoden haben deswegen bisher kaum Einzug in die Anwendung gefunden. Deswegen wird hier zum einen der Weg beschritten, den Nutzern **hochinteraktive Standardtechniken** an die Hand zu geben, und diese zum anderen mit **leicht wahrnehmbaren visuellen Metaphern** anzureichern.

Weiterhin ist bei Darstellungen im geographischen Kontext, wie sie z.B. in Klimadaten häufig vorliegen, neben einer geeigneten Achsenbeschriftung wichtig, dass solche Darstellungen auch um **geographische Informationen** erweiterbar sein müssen (Land- und Seemasken, Ländergrenzen, Relief), ohne die Wertedarstellung zu stören oder sogar zu dominieren. Entsprechend müssen für geographische Informationen und die räumlichen Daten geeignete visuelle Attribute, wie z.B. leicht unterscheidbare Farben oder Farbskalen, gewählt werden.

5.1.2 Standardtechniken

In diesem Abschnitt wird ein kurzer Überblick zu den Einsatzmöglichkeiten von Standardtechniken bei der Darstellung räumlicher Klimadaten gegeben. Dies umfasst Darstellungen für skalare Klimadaten auf regulären oder blockstrukturierten 2D-Gittern, Darstellungen für skalare, gestreute 2D-Klimadaten, Darstellungen für skalare Klimadaten auf regelmäßigen und blockstrukturierten 3D-Gittern und Darstellungen für vektorielle Klimadaten auf regelmäßigen Gittern.

Obwohl es sich hierbei um Standardtechniken handelt, sind sie nicht in einer für die Anwender geeigneten, einfach bedienbaren Form mit einem Visualisierungssystem wie OpenDX direkt verfügbar. Deshalb wurden die vorgestellten Techniken im Rahmen der Dissertation unter Einbeziehung studentischer Arbeiten im Visualisierungssystem OpenDX umgesetzt, in eine VDM-Technikbibliothek (vgl. Abs. 4.2) und/oder in das System SimEnvVis (vgl. Abs. 7.3) eingebunden.

Darstellungen für skalare Klimadaten auf 2D-Gittern. Bei der Visualisierung von 2D-Klimadaten sind Abbildungen auf Farbe, Isolinien und Höhe übliche Vorgehensweisen. Abbildung 5.1 illustriert verschiedene Darstellungen für den *Meeresspiegeldruck* im Umfeld des Golfes von Guinea. Je nach Bedarf können so

- interpolierte Farbskalen (Abb. 5.1a,b,f,e),
- Isolinien (Abb. 5.1b - 5.1e),
- eine segmentierte Farbskala (Abb. 5.1c) oder
- Höhenfelder (Abb. 5.1f,e)

verwendet werden. Zusätzlich geben - hier grau dargestellte - Ländergrenzen einen Eindruck über die örtliche Geographie (Abb. 5.1a - 5.1d). Bei gleichzeitiger Verwendung mehrerer visueller Attribute kann entweder der visuelle Eindruck eines Merkmals verstärkt (Abb. 5.1b - 5.1f) oder mehrere Merkmale gleichzeitig dargestellt werden (Abb. 5.1e).

Darüber hinaus ergeben sich spezielle Anforderungen an die Visualisierung von regulären 2D-Klimadaten. So reichen planare Abbildungen des Gitters nicht aus, wenn die Verhältnisse an den Polen untersucht werden sollen oder flächentreue Abbildungen erforderlich sind. Dazu lassen sich z.B. 3D-Kugeldarstellungen einsetzen (vgl. Abb. 5.2).

Um über die vorgestellten Techniken hinaus auch Abhängigkeiten höherer Ordnung im räumlichen Bezug analysieren zu können, werden häufig Ikonen eingesetzt. Vorteil solcher **Ikonen** ist, dass

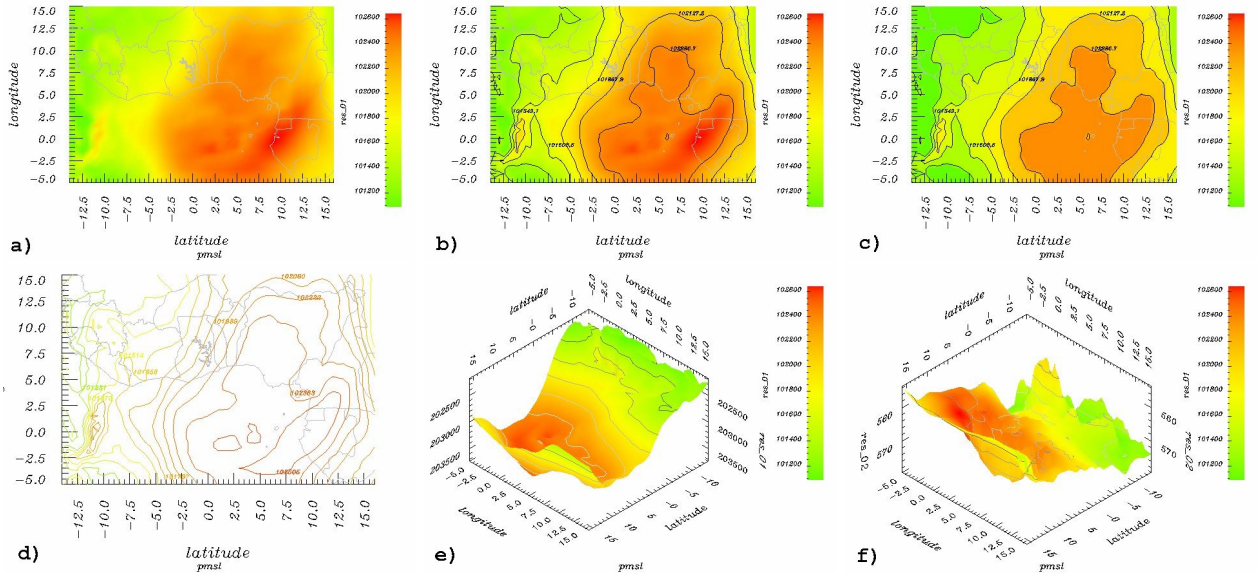


Abbildung 5.1: Standarddarstellungen für Daten auf regelmäßigen 2D-Gittern in SimEnvVis (*Meeresspiegeldruck*); a) interpolierte Farbdarstellung; b) interpolierte Farbdarstellung & Isolinien; c) Farbbänder & Isolinien d) farbkodierte Isoliniendarstellung; e) kombinierte Farb-, Isolinien- und Höhenfelddarstellung; f) analoge Darstellung zu e), jedoch unter Abbildung der *bodennahen Lufttemperatur* auf Höhe und Isolinien und des *Meeresspiegeldrucks* auf Farbe

sie mehrere Datenwerte in einem Primitiv verschlüsseln können und es dem Anwender so erlauben, auch Abhängigkeiten zwischen mehr als zwei Merkmalen zu untersuchen.

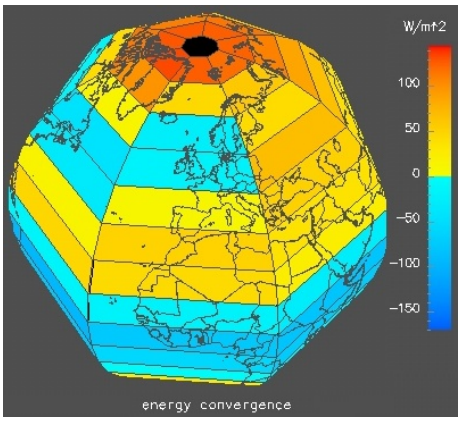


Abbildung 5.2: Kugeldarstellung mit nicht-interpolierter Farbdarstellung aus SimEnvVis

Abbildung 5.3 zeigt beispielhaft eine **vollständige** und eine **aggregierte Darstellung** eines Multi-Run-Datensatzes mit einer „Stick figure"-Ikone[1] (m-Armglyph). Durch die Aggregation können die grundlegenden Charakterista der Daten überblicksartig präsentiert und weitere Details bei Bedarf verfeinert ausgegeben werden.

Darstellungen für skalare, gestreute 2D-Klimadaten. Auch bei der Darstellung von skalaren, gestreuten 2D-Klimadaten sind Abbildungen auf Farbe, Isolinien und Höhe üblich. Hierbei handelt es sich im Allgemeinen um an verschiedenen Messstationen erhobenen Daten, die in ihrer räumlichen Verteilung untersucht werden sollen. Abbildung 5.4 zeigt beispielhaft Repräsentationstechniken für gestreute Klimadaten am Beispiel eines Maisanbau-Datensatzes im Nordosten Brasiliens im Jahr 1983 (mit einer rot-gelb-grün-Farbskala).

Dabei bieten sich die vier folgenden Techniken an:

- **punkthafte Darstellung** der Stationsdaten mit Hilfe eingefärbter Kreisikonen (Abb. 5.4a,b),
- **interpolierte Darstellung** auf Grund einer Delauney-Triangulation der Stationspositionen (Abb. 5.4c,d),
- die **Abbildung auf ein reguläres Gitter** kombiniert mit einer diskreten Farbabbildung der Werte der generierten Gitterzellen (Abb. 5.4e) sowie
- eine **flächenhafte Darstellung** basierend auf einer Voronoi-Zerlegung (Abb. 5.4f).

Darstellungen für skalare Klimadaten auf regelmäßigen und blockstrukturierten 3D-

[1] Die Umsetzung erfolgte im Rahmen eines betreuten Studentenprojektes (Ohl 2005).

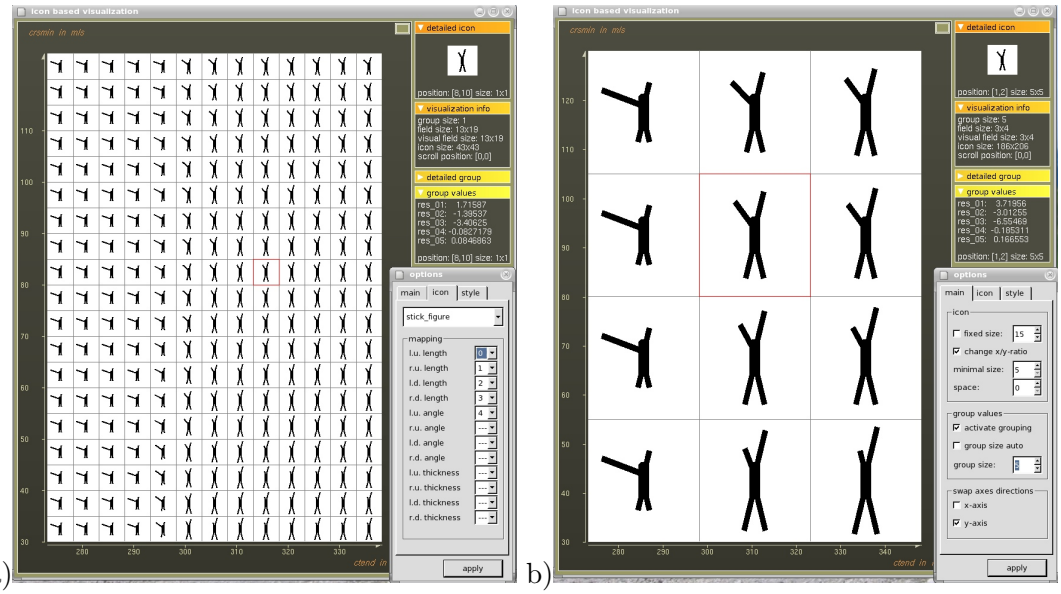


Abbildung 5.3: Ikonendarstellung für reguläre 2D-Gitter; Abbildung von fünf (aggregierten) Merkmalen im Zustandsraum eines Multi-Run-Experiments auf eine „Stick figure“-Ikone (m-Armglyph) aus SimEnvVis; a) Detailbild mit Ikonen für alle beteiligten Beobachtungspunkte; b) Überblicksbild unter Zusammenfassung von Datenpunkten im Zustandsraum (gemittelte Datenwerte)

Gittern. Die Mehrzahl aller Klimamodelle werden auf regulären oder blockstrukturierten Gittern modelliert und simuliert. Insbesondere atmosphärische Simulationen führen zu großen Datenmengen auf zeitveränderlichen, regulären 3D-Gittern. Damit lassen sich für solche Daten Visualisierungstechniken für Volumendaten anwenden (vgl. Abb. 5.5):

- **Dekompositionsmethoden** (Quadermethode (Abb. 5.5a), Schnittdarstellungen mit sich kreuzenden (Abb. 5.5b,c) und parallelen Schnitten (Abb. 5.5d)),
- **Isoflächendarstellungen** (Abb. 5.5e) und
- **DVR-Verfahren** (vgl. Abb. 5.5f)

Bei Anwendung dieser Standardtechniken ist es oft schwierig, die räumliche Verteilung extremer Werte zu untersuchen. Deswegen wurden geeignete Parametrisierungen entworfen, um durch den Einsatz von Transparenzen und (geeigneten) angepassten Transferfunktionen die Analyse von Extremwerten zu unterstützen (Abb. 5.6).

Darstellungen für vektorielle Klimadaten auf regelmäßigen Gittern. Da es sich bei klimatischen Prozessen vor allem um Strömungsprozesse handelt, können für die vorliegenden Vektorfelder Visualisierungstechniken für Strömungsdaten angewendet werden. Herausforderung hierbei ist es, verschiedene Eigenschaften des Vektorfeldes im den räumlichen Bezug mit der darunter liegenden Karte zu visualisieren.

Abbildung 5.7 zeigt wichtige Standardtechniken für die horizontale Strömung an der Erdoberfläche:

- farbkodierte **Stromliniendarstellung** (Abb. 5.7a,b)
- **Pfeildarstellung** (Pfeile variabler (Abb. 5.7c) und konstanter (Abb. 5.7d) Länge

Dabei können je nach Bedarf geographische Informationen und ein weiteres skalares Merkmal hinzugeschaltet werden (vgl. auch Abb. A.1 im Anhang).

Bei der Untersuchung der Strömungseigenschaften um die (Erd-)Pole ist eine planare Darstellung wie in Abbildung 5.7 häufig nicht ausreichend, da diese kein kontinuierliches Bild in diesen Regionen geben kann. Bei der Untersuchung solcher regionaler Klimabedingungen, insbesondere bei der Darstellung von Stromlinien, bietet sich deswegen die Abbildung des Gitters auf eine **Kugel** an (vgl.

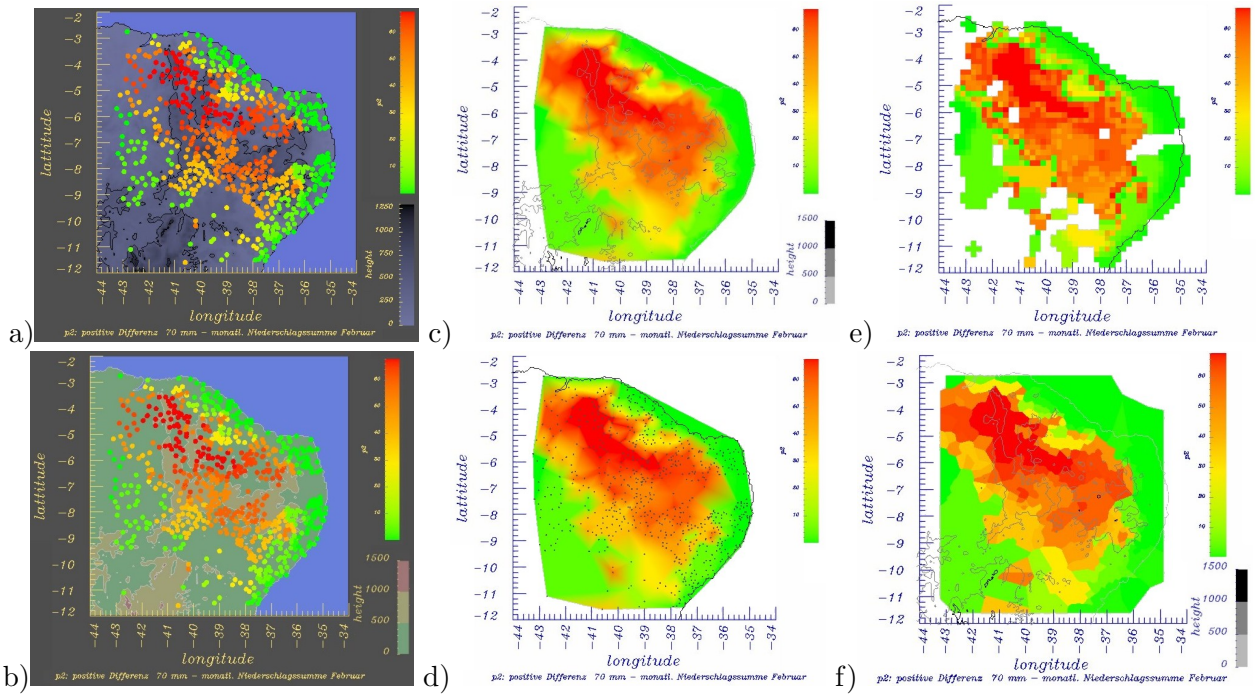


Abbildung 5.4: Standarddarstellungen für skalare, gestreute 2D-Klimadaten (normierte monatliche Niederschlagssumme für Februar(p2)); a) Darstellung farbkodierter Kreise mit kontinuierlichen Reliefinformationen; b) Darstellung farbkodierter Kreise mit diskreten Reliefabstufungen; c) Interpolierte Farbabbildung basierend auf der Delauney-Triangulation mit Relief-Isolinien; d) analog zu c), jedoch Darstellung der Messstationen; e) Abbildung auf ein regelmäßiges Gitter unter Beschränkung der regionalen Ausbreitung und Farbdarstellung; f) Farbabbildung basierend auf einer Voronoi-Zerlegung

Abb. 5.8a). Darüber hinaus unterstützt auch die **Darstellung kritischer Punkte**[2], die Topologie des Vektorfeldes besser zu verstehen und dessen Änderung über die Zeit zu verfolgen (vgl. 5.8b). Gerade bei simulierten Klimamodellen mit geringem Grad an Rauschen können diese das (visuelle) Verständnis des zugrunde liegenden Strömungsverhaltens stark verbessern. Zusätzlich können durch Kodierung der Art[3] der kritischen Punkte weitere Eigenschaften des Feldes abgebildet werden.

Zusammenfassung. In diesem Abschnitt wurden Standardvisualisierungstechniken zur räumlichen Darstellung von Daten verschiedener Charakteristik vorgestellt. Dabei stand im Vordergrund, diese wohlbekannten Techniken im allgemeinen Visualisierungssystem OpenDX zu integrieren und zu parametrisieren und sie so an die Bedürfnisse der Anwendung anzupassen. Damit wird den Anwendern aus der Klimaforschung ein Basiswerkzeug an der Hand gegeben, um die räumliche Datenverteilung zu untersuchen.

Neu aus Sicht der Anwendung ist dabei, dass hiermit die Klimadaten durch verschiedene Darstellungen mit verschiedenen Levels-of-Detail in einem interaktiven Analyseprozess untersucht werden können. Existierende Systeme unterstützen derzeit das Mantra von Shneiderman (1996) nicht oder nur in Ansätzen. Durch das in dieser Arbeit bereitgestellte Basiswerkzeug wird der Anwender mit den vorgestellten Darstellungen in die Lage versetzt, nach einen schnellen Überblick über die räumliche Verteilung der Daten Details über bestimmte Regionen des Beobachtungsraumes oder über bestimmte Werte(-bereiche) flexibel nachladen zu können.

[2]Kritische Punkte sind Punkte im Beobachtungsraum, an denen alle Komponenten des Strömungsvektors gleich 0 sind.

[3]Die Art eines kritischen Punktes kann auf Basis der Eigenwerte der Matrix der partiellen Ableitungen der Geschwindigkeitsvektoren in den kritischen Punkten berechnet werden.

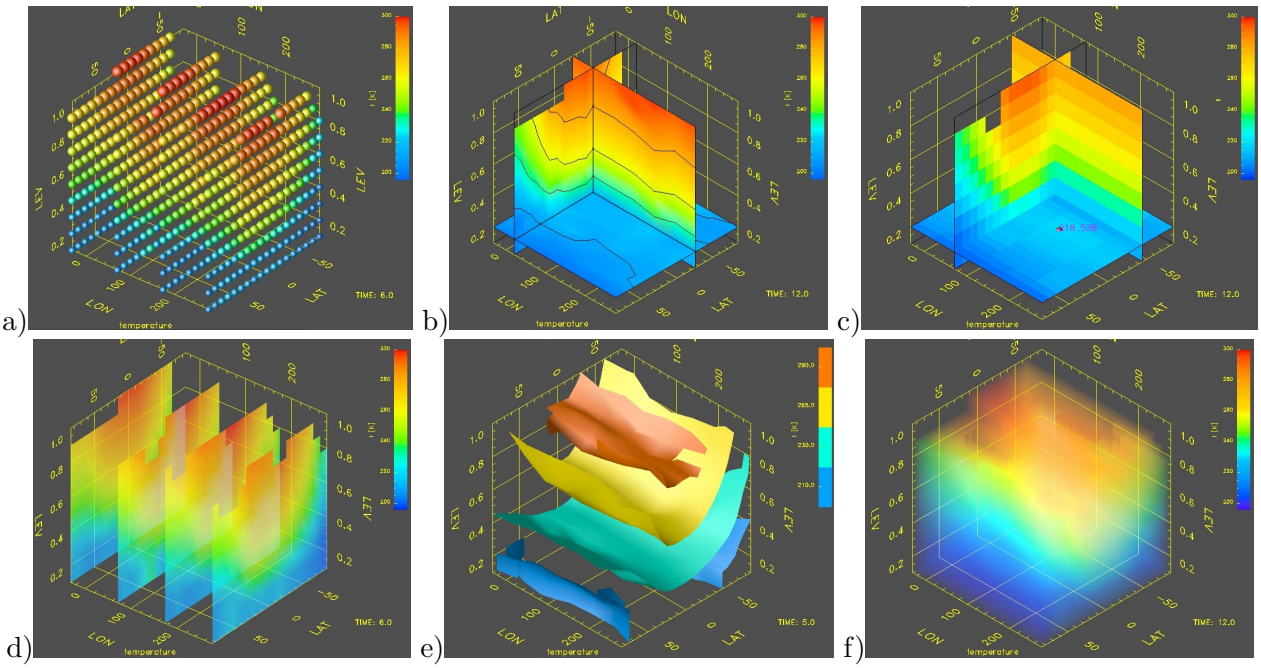


Abbildung 5.5: Standarddarstellungen für skalare Klimadaten auf regulären 3D-Gittern aus SimEnvVis (Verteilung der Lufttemperatur t mit einer angepassten Regenbogenskala); a) Quadermethode mit Kugeldarstellungen; b) Schnittdarstellung mit sich kreuzenden Schnitten unter Abbildung auf Farbe (interpoliert) und Isolinien; c) Schnittdarstellung mit sich kreuzenden Schnitten unter Abbildung auf Farbe (nicht interpoliert); d) Darstellung paralleler Schnitte; e) Isoflächendarstellung; f) DVR-Darstellung mit konstanter Transparenz für alle Datenwerte

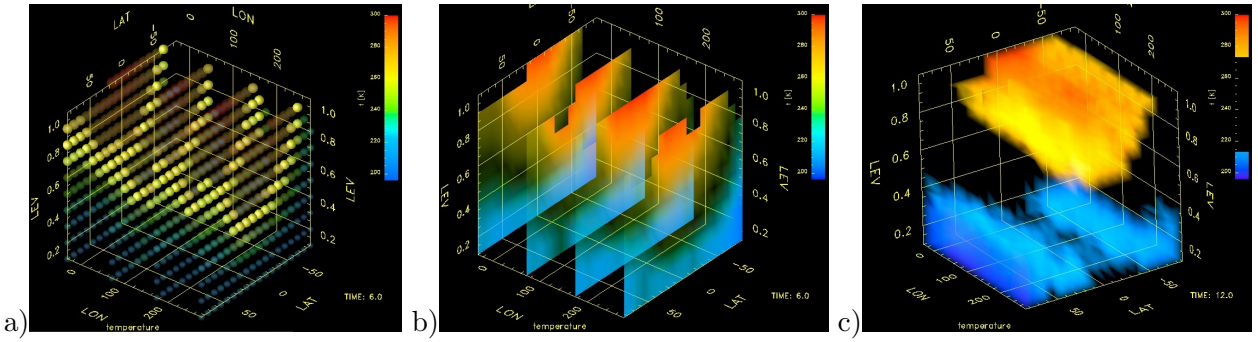


Abbildung 5.6: Fokussierte Darstellungen für skalare Klimadaten auf regulären 3D-Gittern aus SimEnvVis (Verteilung der Lufttemperatur t); a) Hervorhebung mit der Quadermethode durch Darstellung von Kugeln mit variierender Transparenz (Fokus auf Datenwerten um 260K); b) Einsatz von Transparenz bei der Darstellung paralleler Schnitte zur Hervorhebung der Extrema; c) Fokussierung auf die Extreme beim DVR

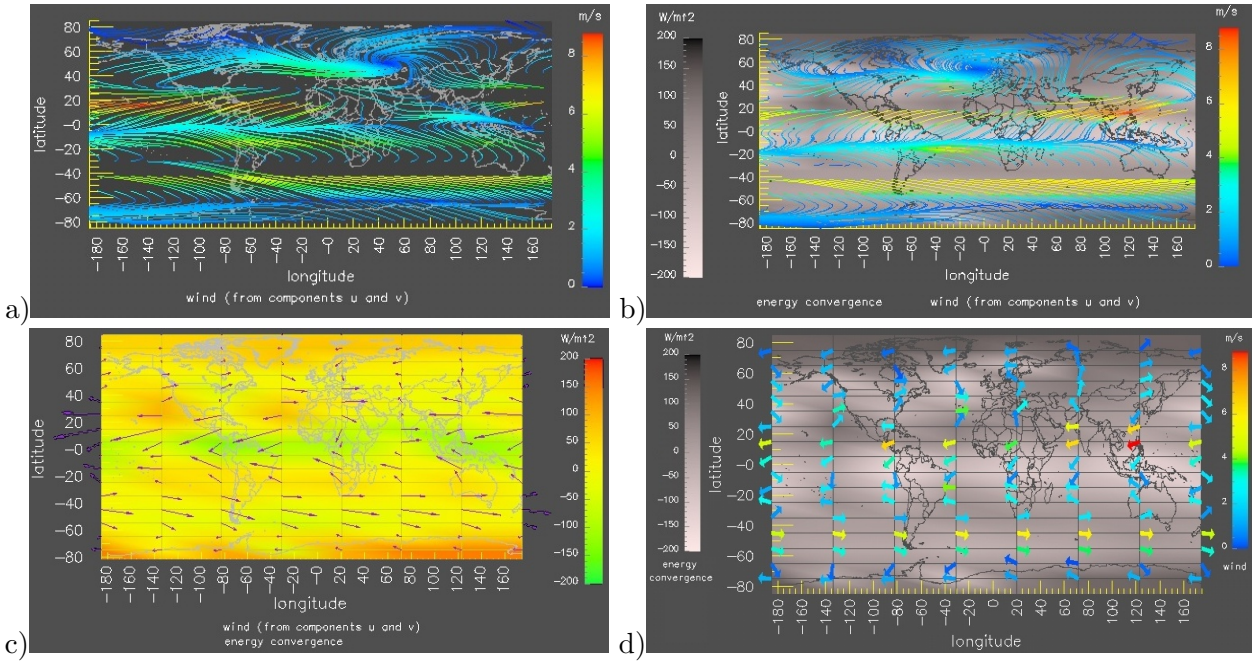


Abbildung 5.7: Darstellungen von 2D-Strömungsdaten (atmosphärische Horizontalströmung mit zusätzlichem, farbkodierten Merkmal aus SimEnvVis (Energiekonvergenz)); a) nach der Stromgeschwindigkeit eingefärbte Stromlinien (ohne farbkodiertes Merkmal); b) eingefärbte Stromlinien; c) Pfeildarstellung mit Pfeilen variabler Länge; d) Pfeildarstellung mit Pfeilen konstanter Länge unter Farbkodierung des Vektorbetrags

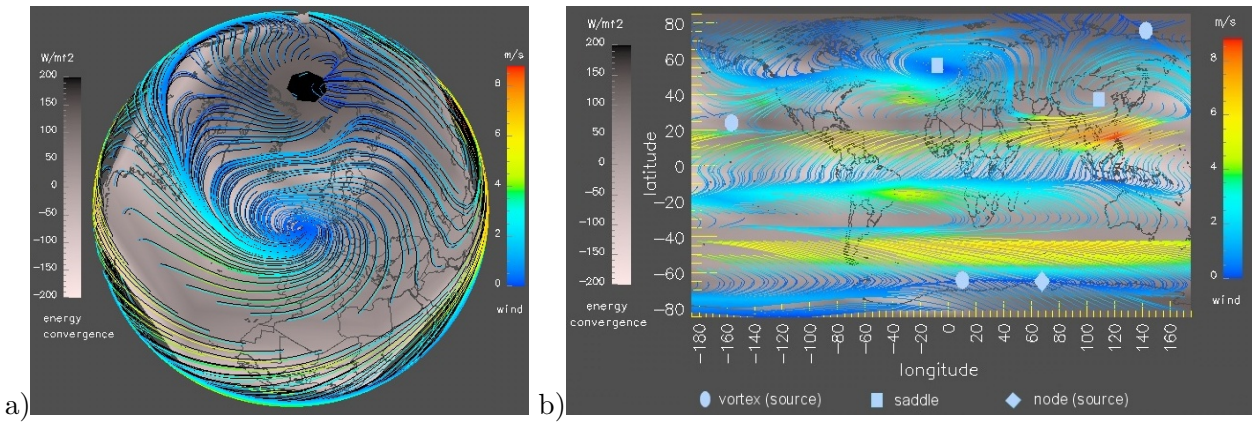


Abbildung 5.8: Weitere Darstellungen von 2D-Strömungsdaten (atmosphärische Horizontalströmung); a) Stromliniendarstellung auf der Erdkugel (Farbkodierung der Strömungsgeschwindigkeit) in Kombination mit einem farbkodierten Merkmal; b) analoge Stromliniendarstellung zu Abbildung 5.7b unter Darstellung von typabhängigen Glyphen für kritische Punkte verschiedener Art

5.1.3 Metapherbasierte Ikonendarstellungen

Die beschriebenen Standardtechniken bieten eine breite Basis, um verschiedene Klimadatensätze in ihrem räumlichen Bezug untersuchen zu können. Dabei haben sich bestimmte Techniken wir z.B. 2D-Farb- und Isoliniendarstellungen in diesem Umfeld etabliert. Dahingegen ist der Einsatz anderer Techniken wie z.B. von Ikonen zur Kodierung multi-variater Daten (vgl. Abb. 5.3) in der Anwendung eher selten. Dies liegt insbesondere daran, das solche, häufig komplexen Methoden einen höheren kognitiven Aufwand erfordern. Um die Lücke zwischen den Potentialen solcher Techniken und deren Anwendung in der Praxis zu schließen, wird im Rahmen dieser Arbeit die neue Methode der „metapherbasierten Ikonendarstellung“ eingeführt (vgl. hierzu auch Nocke u. a. (2005)).

Unter dem Begriff **metapherbasiert** ist in diesem Zusammenhang zu verstehen, dass die Daten in einer für das Umfeld der Anwendung leicht verständlichen Art und Weise dargestellt werden und es so erleichtern, diese zu interpretieren und zu kommunizieren. Solche visuellen Metaphern werden z.B. in der „ambienten Informationsdarstellung“ (vgl. z.B. Stasko u. a. 2004; Shen u. Eades 2004), zur Konzeptvisualisierung (vgl. z.B. Nesbitt 2004) und in der Kartographie (vgl. z.B. Hake u. Grünreich 1994) eingesetzt.

Ikonenendarstellungen wurden für die Untersuchung von Abhängigkeiten mehrerer Merkmale eines Datensatzes entwickelt. Dabei werden Ikonen eingesetzt, die mehrere Datenwerte in einem Primitiv verschlüsseln. Zumeist werden dabei abstrakte Ikonen verwendet, die nicht auf den dargestellten Sachverhalt verweisen[4], und so für den Anwender nur eingeschränkt nutzbar sind. Um die Effektivität solcher Ikonen in der Anwendung zu verbessern, soll hier der Ansatz der **metapherbasierten Ikonendarstellung** beschritten und dessen Nutzung für der Anwendung untersucht werden. Erste Arbeiten zur Darstellung von leicht zu interpretierenden, metapherbasierten Ikonen finden sich bei Chernoff (1973); Chuah u. Eick (1998); Mao u. a. (2000).

Dabei ergeben sich die folgenden zwei Herausforderungen: (1) Entwurf von geeigneten Ikonen und (2) Anordnung dieser Ikonen im räumlichen Bezug. Am Beispiel eines eines Datensatzes zur Abschätzung des Risikos des Ernteverlustes beim Maisanbau soll im Folgenden der hier beschrittene Lösungsweg skizziert werden.

Ikonenentwurf. Der Entwurf metapherbasierter Ikonen beinhaltet die folgenden Teilschritte:

- *Entwurf* einer kompakten, ausdrucksstarken Metapher für den Datensatz,
- *Definition von visuellen Attributen* sowie deren Änderungen, in denen sich einzelne Ikonen voneinander unterscheiden sollen,
- geeignete *Abbildung der Merkmale* auf die visuellen Attribute und
- *Bindung der Wertebereiche* der Merkmale an die Änderungen der Attribute der Ikone.

Entsprechend der Bedeutung der Merkmale des Datensatzes (Indikatoren für das Risiko des totalen Ernteverlustes der Maisernte) wurde die bildbasierte Metapher von intuitiv erkennbaren und leicht verständlichen Maiskolben ausgewählt (*Entwurf*, vgl. Abb. 5.9).

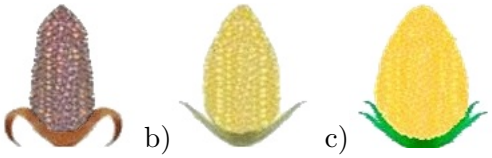


Abbildung 5.9: Basisikonen für schlechte, mittlere und gute Anbaubedingungen

Die Ikone bildet auf die visuellen Attribute *Farbe* und *Größe* ab (*Definition von visuellen Attributen*). Sie erfordert eine Separierung der Wertebereiche der - das Ernterisiko indizierenden - Merkmale in drei Intervalle (*Bindung des Wertebereiches*). Geringe Risiken bei der Maisernte werden durch einen dicken, gelben Maiskolben (Abb. 5.9a), hohe Risiken durch einen dünnen, braunen Maiskolben (Abb. 5.9c) und mittlere werden durch einen in Farbe und Dicke gemittelten Maiskolben repräsentiert (Abb. 5.9b).

[4] wie z.B. Arm- und Beinlängen sowie -winkel der „Stick-figure“-Ikonen aus Abbildung 5.3

Um die sechs Merkmale des Datensatzes darzustellen, reicht diese einfache Metapher jedoch nicht aus (*Abbildung der Merkmale*). Eine Möglichkeit wäre nun, an jeder Beobachtungsstation sechs einzelne Maiskolben für jedes Merkmal darzustellen. Da jedoch einzelne Ikonen erkennbar bleiben sollen, müssten diese relativ groß dargestellt werden, was die Ikonen weniger kompakt macht und schnell zu Überlappungen in der Darstellung in Gebieten mit hoher Messstationsdichte führen kann.

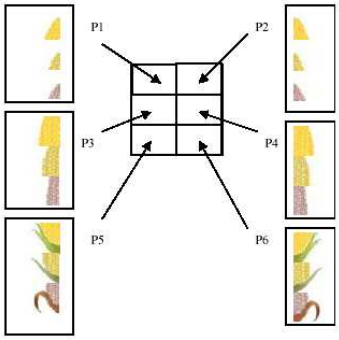


Abbildung 5.10: Segmentierung der Basisikonen und deren Zusammensetzung

Deswegen wurde der Ansatz verfolgt, die Bilder in sechs Regionen einzuteilen (*Entwurf*), und das resultierende Bild als Kombination aus sechs Bildteilen zu generieren (*Definition von visuellen Attributen*), die in Abhängigkeit der Werte der zugehörigen Merkmale ausgewählt werden (*Abbildung der Merkmale* und *Bindung des Wertebereiches*, vgl. Abb. 5.10). So führen beispielsweise hohe Werte des Merkmals p_1 zu einem dicken, gelben Kopf auf der linken Seite des Maiskolbens. Auf diese Weise können kompakte Bilder generiert werden, die es dem Anwender erlauben, alle Merkmale und deren Abhängigkeiten mit einem Blick zu erfassen.

Anordnung von Ikonen auf Karten. Nach dem Entwurf der Ikone ist die Frage zu entscheiden, wie die Ikonen auf der Karte angeordnet werden. Ein erster Ansatz ist es, diese genau an den Beobachtungspunkten zu platzieren. Jedoch treten dabei in Bereichen hoher Punktdichte Überlappungen auf, so dass wichtige Informationen verloren gehen können. Die Idee dieser Arbeit ist es, Anordnungsalgorithmen für die Erzeugung von *Mosaikbildern* einzusetzen, um Maisikonen anzuordnen und dabei Überlappungen zu vermeiden. Nach einer kurzen Einführung in die Grundlagen der Generierung von Mosaikbildern sollen im folgenden verschiedene, teilweise neue Ikonenanordnungen vorgestellt und diskutiert werden.

Ein Mosaikbild wird aus einer Menge kleinerer Bilder erzeugt (so genannte „Tiles“), die zusammen ein größeres Objekt bilden. Dazu werden Bildbereiche des Originalbildes durch kleine Bilder ersetzt, welche den Farbton-Eigenschaften des Originalbildes nahe kommen (vgl. Finkelstein u. Range 1998; McKenna u. Arce 2000). Dazu werden die folgenden vier Schritte durchlaufen: (1) Auswahl von Bildern, die als Tiles genutzt werden sollen, (2) Auswahl eines Gitters zur Anordnung der Tiles, (3) Zuweisung der Tiles auf das Gitter sowie (4) ggf. Durchführung der Farbanpassung einzelner Tiles, um sie an die Farben des Originalbildes anzupassen.

Diese Vorgehensweise kann nun leicht für die Anordnung und Darstellung von Ikonen auf Karten angepasst werden (vgl. hierzu Nocke u. a. 2005). Insbesondere die Anordnungstechniken für verschiedene Gitter, die bei Techniken zur Erzeugung von Mosaikbildern Verwendung finden, führen zu neuen Darstellungen von Ikonen. Abbildung 5.11 zeigt verschiedene Ikonendarstellungen für den Maisdatensatz.

Diese Ikonen-Darstellungen entsprechen den drei grundlegenden, von Strothotte u. Schlechtweg (2002) eingeführten Layout-Typen für Mosaikbilder: *gestreutes*, *reguläres* und *Multi-resolution-Layout*. Im folgenden sollen nun die drei Layoutvarianten in ihrer Anwendung auf Ikonen vorgestellt und ihre Vor- und Nachteile kurz diskutiert werden.

Das *gestreutes Layout* für Mosaikbilder ist ein zufälliger Prozess, der im allgemeinen künstlerischen Zwecken dient. Dieses Vorgehen ist jedoch für die Visualisierung von numerischen Daten im allgemeinen nicht geeignet, weil bei einer Positionierung der Ikonen auf zufällige Positionen die wichtige Beziehung zwischen Daten und deren Messorten verloren geht. Deswegen sollen lediglich die zwei folgenden gestreuten Ikonen-Layouts betrachtet werden: exakte Positionierung von Ikonen konstanter Größe auf die Positionen ihrer Beobachtungspunkte sowie deren Platzierung auf gegebene oder

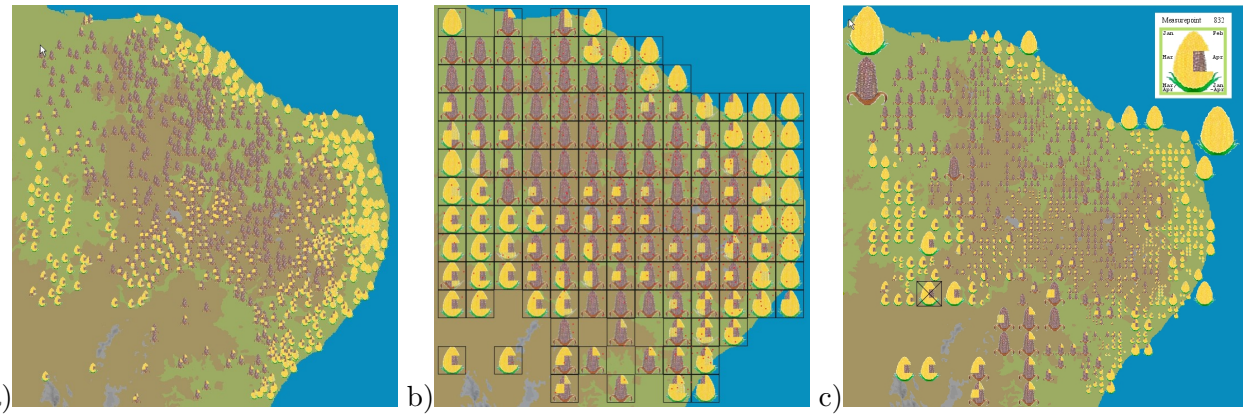


Abbildung 5.11: Metapherbasierte Ikonendarstellungen für skalare, gestreute 2D-Klimadaten (Maisanbau-Datensatzes mit sechs Parametern); a) Darstellung der Mais-Ikonen an ihren Messstationen; b) Überlagerung des Beobachtungsraumes mit regelmäßigem Gitter, Werteaggregation aller Stationen innerhalb einer Gitterzelle und Darstellung von Maisikonen für gemittelte Datenwerte; c) Darstellung der Mais-Ikonen in Bereichen eines Quadtree sowie Selektion einer Ikone und deren vergrößerter Darstellung

berechnete[5] Regionen, wobei Position und Größe der Ikonen in Abhängigkeit der Ausdehnung der Region angepasst werden können. In Abbildung 5.11a wird der erste Mechanismus verwendet, der die Ikonen auf die Positionen der Messstationen abbildet. Vorteil dieser Vorgehensweise ist, dass so Fehlinterpretationen bezüglich der Verteilung der Beobachtungspunkte vermieden werden können. Allerdings ist die Interpretierbarkeit solcher Darstellungen begrenzt, da die oben angesprochenen Überlappungen Informationsverluste erzeugen können bzw. relativ kleine Ikonen verwendet werden müssen.

Um dieses Problem zu lösen, wurden zusätzlich reguläre und Multi-resolution-Layouts umgesetzt. Beim *regulären Layout* (Abb. 5.11b) zeigt jede Gitterzelle aggregierte Informationen aller ihr zugehörigen Messstationen, in diesem Fall die Ikone der gemittelten Stationswerte. Ein solches reguläres Gitterlayout ermöglicht einen schnellen *Überblick* über die Daten in ihrer räumlichen Werteverteilung. Nachteil solcher Darstellungen ist, dass hierbei die ursprünglichen Datenwerte aggregiert werden und nicht mehr explizit erkennbar bleiben. Um dieses Problem zu lösen, kann die Gitterauflösung interaktiv geändert werden (*details-on-demand*). In einer geringen Gitterauflösung erhält der Anwender einen guten Überblick mit wenigen großen Ikonen, während eine hohe Auflösung *Details* über bestimmte Regionen bei kleinen Ikonen aufdeckt. Um auch weiterhin einen Überblick über die Verteilung der Messpunkte in einer bestimmten Region zu bekommen, können kleine Punkte an den Positionen der Messstationen eingeblendet werden (vgl. Abb. 5.11b).

Als dritte - durch die Mosaikbilder inspirierte - Anordnungstechnik wurde ein *Multi-resolution-Layout* für Ikonen umgesetzt (vgl. Abb. 5.11c). Mithilfe einer Quadtree-Unterteilung des Beobachtungsraumes werden die Ikonen - in der aktuellen Auflösungsstufe des Quadtree - gezeichnet, wenn sich genau ein Messpunkt in der aktuellen (rechteckigen) Region befindet. Ansonsten wird diese weiter unterteilt. In Regionen geringer Messpunktdichte können so große Ikonen gezeichnet werden, während in dichten Regionen kleine Ikonen dargestellt werden. Vorteil dieses Vorgehens ist, dass für jede Messstation einzelne Ikonen gezeichnet werden, die sich nicht überlappen. Dies ermöglicht die Identifizierung und Untersuchung einzelner Ikonen, wobei der Darstellungsplatz optimal ausgenutzt wird. Allerdings führt die Darstellung von Ikonen verschiedener Größe zu einer stärkeren Gewichtung der Daten in dünn besetzten Regionen. Wenn dies nicht erwünscht ist, können alternativ auch für alle Regionen Ikonen gleicher Größe gezeichnet werden.

[5] z.B. durch Voronoi-Tesselierung

Ein Quadtree führt eine sukzessive Unterteilung von Rechtecken durch und schätzt so die zugrunde liegenden Regionen ab. Da diese Regionen in der Realität jedoch unregelmäßige Begrenzungen haben, kann es passieren, dass der Quadtree auch in nicht zulässige Regionen Ikonen zeichnet (in diesem Fall in den Atlantischen Ozean, vgl. Abb. 5.11c). Um dies zu vermeiden, muss der Quadtree-Algorithmus um ein Clipping erweitert werden.

Zusammenfassend lässt sich feststellen, dass mit metapherbasierten Ikonen im Vergleich zu farbbasierten Standardtechniken (vgl. Abb. 5.4) und zu Standardikonendarstellungen (vgl. z.B. Abb. 5.3) die Daten kompakt dargestellt werden können, wobei Merkmalsabhängigkeiten leicht identifiziert werden können. Der Nutzer kann durch Wahl verschiedener Layouts flexibel einen allgemeinen Überblick bekommen (reguläres Layout) und Details über Gebiete (reguläres Layout mit feinerer Gitterauflösung oder Multi-resolution-Layout) und über einzelne Messstationen (gestreutes Layout) nachladen. Dies ermöglicht es, schnell extreme Bedingungen - wie z.B. sehr gute Bedingungen an der Küste und sehr hohe Risiken in nördlichen Regionen des Innenlandes - zu erkennen und weiter zu untersuchen (vgl. Abb. 5.11).

Im Rahmen einer studentischen Arbeit (vgl. Baalcke 2005) konnte gezeigt werden, dass die vorgestellten Techniken zur metapherbasierten Ikonenvisualisierung breite Einsatzmöglichkeiten für die Analyse von Klimadaten haben. Dabei entstand ein interaktives Tool, in dem eine Vielzahl von Parametern den Anwender dabei unterstützen, seine Fragestellungen passgenau zu beantworten. Hierzu gehören neben Layouttyp und Gittergröße auch die Einbindung verschiedener Hintergrundbilder für das Relief, Schwellwertdefinitionen und verschiedene Arten von Ikonenberechnungen. Um erfolgreiche Parametrisierungen wiederherstellen zu können, können zusätzlich erfolgreiche Parameterkombinationen abgespeichert und wiederhergestellt werden. Im Anhang A befindet sich ein Screenshot des entstandenen Tools (vgl. Abb. A.2).

Metapherbasierte Ikonendarstellungen sind im besonderen auf die Wertedarstellung zugeschnitten, erlauben aber im begrenzten Maße auch die Untersuchung von Merkmalsabhängigkeiten. Um Abhängigkeiten höherer Ordnung in den Daten zu finden, werden typischerweise statistische Verfahren wie z.B. die Clusteranalyse eingesetzt. Wie sich die hier vorgestellten Darstellungstechniken durch geeignete Anpassungen zur Darstellung von geclusterten 2D-Klimadaten einsetzen lassen, wird in Abschnitt 6.1.4 genauer untersucht.

5.1.4 Diskussion

In diesem Abschnitt wurden verschiedene, im Visualisierungsumfeld bekannte Techniken zur Darstellung des räumlichen Bezuges auf deren Einsetzbarkeit im Klimaumfeld hin untersucht. Dazu wurde eine Vielzahl von Darstellungstechniken für verschiedene Datenklassen entworfen und als OpenDX bzw. Java-Module umgesetzt. Wichtige Aspekte beim Entwurf waren gemäß des Mantras von Shneiderman den Anwendern aus der Klimaforschung eine Vielzahl von Interaktionsmöglichkeiten an die Hand zu geben, unterstützende Darstellungen des geographischen Bezuges in die Darstellungen einzubetten sowie die Exploration von klimatischen Extremen zu unterstützen. Insbesondere bei der Darstellung von multi-variaten, gestreuten 2D-Klimadaten wird mit der metapherbasierten Visualisierung von Ikonen und deren Anordnung in Anlehnung an Mosaikbilder Neuland beschritten.

Ferner wurden verschiedene im Klimaumfeld bisher wenig gebräuchliche, jedoch im Visualisierungsumfeld wohlbekannte Darstellungstechniken umgesetzt und an die Anforderungen der Anwendung angepasst. So sind z.B. die Darstellung von parallelen Schnitten (vgl. Abb. 5.5d und 5.6b) und die Darstellung von kritischen Punkten unter Kodierung von deren Typ (vgl. Abb. 5.8b) in diesem Umfeld bisher ungebräuchlich. Mit einer Vielzahl von Abbildungen wurden die umgesetzten Techniken illustriert (vgl. weiterhin auch Abb. A.3 im Anhang).

Tabelle 5.1 fasst die im Rahmen dieser Arbeit umgesetzten Techniken und deren wichtigste Eigenschaften zusammen. Dabei werden ähnliche Techniken zu Gruppen zusammengefasst und diesen Techniken die Anzahl darstellbarer Merkmale, unterstützte Aufgaben und Zielstellungen, umgesetzte Interaktionen und Navigationen sowie mögliche Probleme zugeordnet.

5.2 Darstellung von Klimadaten im zeitlichen Bezug

Die Beachtung zeitlicher Zusammenhänge in den Daten und die explizite Darstellung des Parameters Zeit sind in der Visualisierungsliteratur eher unterrepräsentiert (vgl. z.B. Thomas 2005). Für die Klimafolgenforschung ist die Zeit jedoch ein extrem wichtiger Parameter und muss in der Visualisierung dementsprechend beachtet werden. Dazu werden zuerst Anspruch und Probleme in diesem Kontext diskutiert (vgl. Abs. 5.2.1). Im Anschluss daran werden Standardmethoden und deren für Einsatz die Visualisierung von Klimadaten im zeitlichen Kontext diskutiert (vgl Abs. 5.2.2). Daran schließen sich spezielle, zum Teil neue Methoden zur Darstellung des Parameters Zeit an (vgl. Abs. 5.2.3). Abschließend werden die Eigenschaften der vorgestellten Techniken zusammengefasst (vgl. Abs. 5.2.4).

5.2.1 Anspruch und Probleme

Häufig handelt es sich bei Klimadaten um zeitlich **hochaufgelöste Messdaten über sehr langen Zeiträumen**. Eine Beispiel hierfür ist die Potsdamer Reihe, bei der an der Messstation Potsdam über einen Zeitraum von mehr als 100 Jahren stündliche Messungen von Temperatur, Niederschlagswerten und anderen Merkmalen vorgenommen wurden. Neben der Darstellung einzelner Merkmale bei solch massiv großen Datensätzen ist gerade auch die **Untersuchung multi-variater Muster** für die Klimaforscher von Interesse. Herkömmliche Zeitgraphen unterstützen die Analyse solcher komplexer Abhängigkeiten in zeitlichen Merkmalen jedoch nur in bedingtem Maße.

Ferner ergeben sich für die Anwendung verschiedene **Fragestellungen auf verschiedenen Skalen**. So können z.B. zum Beispiel Darstellungen für jährlich akkumulierte Daten nur begrenzt für die Darstellung von Tagesverläufen angewendet werden. Insbesondere der systematische Einfluss von Jahreszeiten und Tag- und Nachtbedingungen spielt für die Visualisierung eine besondere Rolle.

Weiterhin ist auch bei zeitabhängigen Klimadaten die Untersuchung von **extremen klimatischen Verhältnissen** und von **Periodizitäten** eine wichtige Aufgabe. Entsprechend müssen Mechanismen bereitgestellt werden, um solche Bedingungen zu veranschaulichen.

5.2.2 Standardmethoden

Auch zur Darstellung zeitabhängiger Klimadaten werden leicht verständliche Standardmethoden eingesetzt. Grundsätzlich zählen hierzu die drei folgenden Darstellungsarten: Zeitgraphen, Animationen[6] sowie räumliche Darstellungstechniken unter Abbildung der Zeit auf eine räumliche Achse. Diese Techniken wurden im Rahmen der vorliegenden Arbeit umgesetzt und sollen kurz vorgestellt werden.

Zeitgraphen. Zeitgraphen sind die wohl bekannteste Darstellungstechnik für zeitliche Phänomene, wobei von deren Abhängigkeit vom räumlichen Bezug abstrahiert wird (vgl. Abb. 5.12). Insbesondere ergibt sich bei vergleichenden Darstellungen von Zeitgraphen das Problem, dass die dargestellten Merkmale häufig verschiedene Einheiten und/oder Wertebereiche aufweisen. Um Fehl-

[6] Erzeugung mehrerer Bilder unter Variation der Dimension Zeit

Gruppe	Technik	Anz.MM	Aufgaben	Zielstellungen	Interakt. u. Navig.	Probleme
Darstellungen	Interpol. Farb- o. Höhenfelddar.	1	Überblick	Verteilungen	Pan, Zoom	Verdeckungen
von Skalaren	Isolinien o. Farbbänder	1	DoD	Segmentieren	Pan, Zoom	
auf 2D-Gittern	diskrete Farb- o. Höhenfelddar.	1	Details	Identif. u. Lokal. von Werten	Pan, Zoom, Picking	Verdeckungen
(OpenDX)	kombinierte Darstellungen	max 3	Überblick, Details	s.o., Vergleichen	Pan, Zoom, Picking	
	Kugeldarstellung	max 2	Überblick, Details	Untersuchung der Pole	Rotation, Pan, Zoom	
Ikonendarstellungen	aggregierte Darstellung	8 u. mehr	Überblick	Vergleich	Piking	
f. 2D-Gitter(OpenGL)	Einzeldarstellung	8 u. mehr	Details	Vergleich	Piking	
Standardtechniken	farbk. Delauney-Triangulation	1	Überblick	Verteilungen unters.	Pan, Zoom	
für gestreute	farbk. Abb. auf reg. Gitter	1	DoD	regionale Verteilungen unters.	Pan, Zoom	Aggr.-fehler
2D-Daten (OpenDX)	farbk. Voronoizerlegung	1	Details	lokale Gebietsverteilungen	Pan, Zoom	
	farbk. Kreise	1	Details	Identif. u. Lokal. von Werten	Pan, Zoom	
Volumenvisuali-	Quadermethode	1 u. mehr	DoD	Identifizieren, Lokalalisieren	Pan, Zoom, Rot.	Verdeck., Muster
sierungen	kreuzende Schnitte	1	Details	Identif., Lokal., Verteil.	Pan, Zoom, Rot.	
(OpenDX)	parallele Schnitte	1	DoD	Vergl. von Schichten, Verteil.	Pan, Zoom, Rot.	
	Isoflächendarstellung	1	DoD	Segmentieren	Pan, Zoom, Rot.	
	DVR	1	Überblick, Detail	Verteilungen, Segmentieren	Pan, Zoom, Rot.	
Strömungsvisuali-	Liniendarstellungen	1S u. 1V	Überblick, DoD.	Unters. d. Strömungsverhaltens	Pan, Zoom	Abtastprobleme
sierungen (2D)	Pfeile variabler Länge	1S u. 1V	Details	Identif., Lokal., Verteil.	Pan, Zoom, Piking	Verdeckungen
(OpenDX)	Pfeile konstanter Länge	1S u. 1V	Details	Identif., Lokal., Verteil.	Pan, Zoom, Piking	
	Darstellung krit. Punkte	1S u. 1V	Details	Untersuchung der Topologie	Pan, Zoom, Piking	
Mosaikikonen für	gestreutes Layout	6 u. mehr	Details	Verteil.,Identif.,Lokal.,Vergl.	Piking	Verdeckungen
gestreute 2D-Daten	regelmäßiges Layout	6 u. mehr	Überblick	Verteil.,Identif.,Lokal.,Vergl.	Piking	
(Java)	Multires. Layout	6 u. mehr	Überblick, Detail	Verteil.,Identif.,Lokal.,Vergl.	Piking	Aggr.-fehler

Tabelle 5.1: Wichtige Eigenschaften der umgesetzten Raumdarstellungstechniken

Legende:

1S. u. 1V	-	ein skalares und ein vektorielles Merkmal	*Identif.*	-	Identifizieren von Werten,
Anz.MM	-	Anzahl darstellbarer Merkmale	*krit. Punkte*	-	kritische Punkte
Aggr.-fehler	-	Aggregationsfehler durch Durchschnittsbildung,	*Lokal.*	-	Lokalisieren von Werten
DoD	-	Details-on-Demand,	*Rot.*	-	Rotierbarkeit der Darstellung
farbk.	-	farbkodiert,	*Vergl.*	-	Vergleich
Interakt. u. Navig.	-	unterstützte Interaktions- und Navigationstechniken,	*Verteil.*	-	Untersuchung von Werteverteilungen

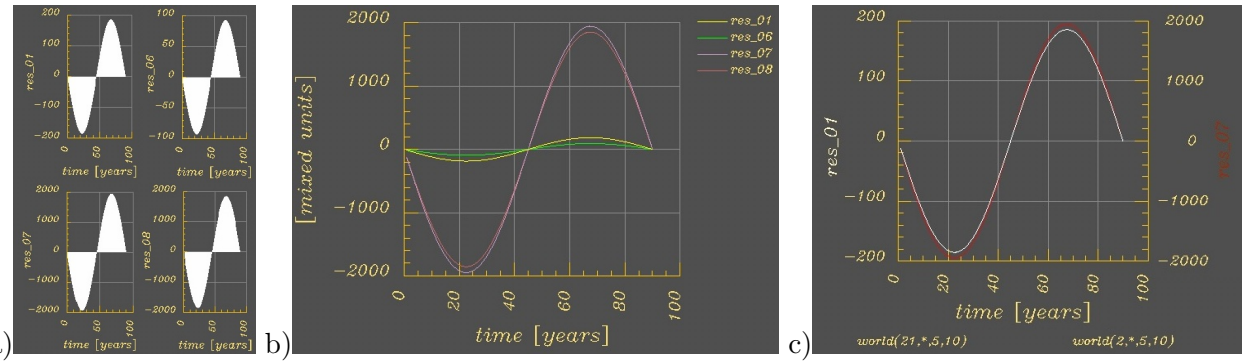


Abbildung 5.12: Darstellungen von Zeitgraphen in SimEnvVis; a) Tabellendarstellung mit vier Einzeldarstellungen; b) Kombinierte Darstellung von vier Graphen; c) Kombinierte Darstellung von zwei Graphen (mit zwei Achsen)

interpretationen bei deren gemeinsamer Darstellung zu vermeiden, können die drei folgenden Strategien verfolgt werden:

1. Um die grundlegenden Verläufe vergleichen zu können und um einen allgemeinen *Überbick* zu erhalten, können alle Merkmale in separaten Graphen dargestellt und zu einer **tabellarischen Darstellung** zusammengefasst werden (vgl. Abb. 5.12a). Durch eine Ausfüllung der Kurve unter der Fläche entstehen so Muster, die einen schnellen Einblick in generelle Trends und Abhängigkeiten ermöglichen.
2. Sollen mehrere Graphen direkt miteinander verglichen werden, können diese in einen **kombinierten Plot** eingezeichnet werden (vgl. Abb. 5.12b). Hierbei muss von den Einheiten der einzelnen Merkmale abstrahiert werden. Durch Filterung von Merkmalen können so Abhängigkeiten je nach Bedarf im Detail untersucht werden (*Details-on-Demand*). Nachteil hierbei ist, dass die Merkmale mit der größten Wertebereichsspanne die Darstellung dominieren (wie auch in Abb. 5.12b).
3. Um diesen Nachteil zu umgehen, kann der Anwender die **Darstellung** zum dritten **auf zwei Merkmale** reduzieren (vgl. Abb. 5.12c), und dabei durch Darstellung zweier Achsen für beide Wertebereiche eine maximale Spreizung der beiden Kurven erreichen. Diese können dann auf Abhängigkeiten im *Detail* hin untersucht werden.

Animation. Handelt es sich um zeitveränderliche Klimadaten mit räumlichem Bezug, werden üblicherweise räumliche Darstellungen gewählt und deren Veränderung über die Zeit animiert. Dies gibt dem Anwender schnell einen *qualitativen Überblick* über grundlegende Trends in den Daten. Nachteil von Animationen ist, dass der Vergleich einzelner Zeitschritte oder die Extraktion einzelner Werte wesentlich erschwert ist. Der Anwender kann lediglich einzelne Zeitschritte anwählen und deren Darstellung mit Darstellungen anderer Zeitschritte auf der Bildebene vergleichen (vgl. auch Abb. A.5 im Anhang).

Abbildung der Zeit auf eine räumliche Achse. Um die eingeschränkte Flexibilität von Animationen - bezüglich des Vergleiches verschiedener Zeitschritte und der Extraktion von *Details* aus den Daten - zu überwinden, werden (Standard-)Darstellungen für räumliche Daten unter Ersetzung einer räumlichen Achse durch die Zeit eingesetzt (vgl. Abb. 5.13). Dies ermöglicht, raum-zeitliche Verteilungen explizit zu untersuchen. Abbildung 5.13a stellt die Zeit in einer 2D-Darstellung zusammen mit der Breite in einer kombinierten Farb- und Isoliniendarstellung dar. Hierbei kann das zyklische Verhalten des Merkmals „synoptischer Wärmetransport“ direkt untersucht und ggf. auftretende Abweichungen der einzelnen Zyklen eingeschätzt werden.

Ferner können zeitveränderliche 2D-Gitter durch Standard-3D-Darstellungen abgebildet werden (vgl. Abb. 5.13b-f). Insbesondere Isoflächendarstellungen sind geeignet, um zeitliche Muster zu explorieren (vgl. Abb. 5.13b,f). Bei stark fluktuierendem Verhalten über die Zeit (wie beim Mee-

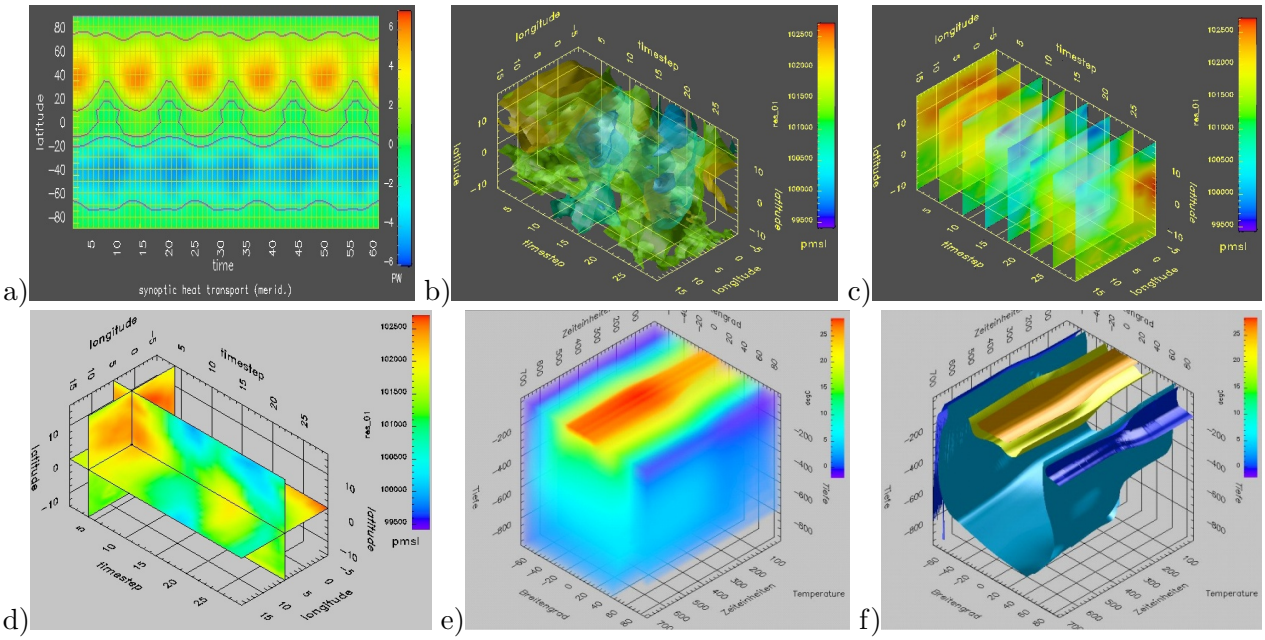


Abbildung 5.13: Darstellungen der Zeit durch Ersetzung einer räumlichen Achse aus SimEnvVis; a) 2D interpolierte Farb- und Isoliniendarstellung (synoptischer Wärmetransport); b) transparente Isoflächendarstellung (Meeresspiegeldruck); c) parallele Schnitte (Meeresspiegeldruck); d) sich kreuzende Schnitte (Meeresspiegeldruck); e) DVR mit geringer Transparenz (Ozeantemperatur); f) Isoflächendarstellung (Ozeantemperatur)

resspiegeldruck) können transparente Isoflächen Informationen über Stabilität von (Werte-)Mustern sowie das Auftreten extremer Bedingungen verfolgt werden (vgl. Abb. 5.13b). Aber auch bei einem stabilen Verhalten können Isoflächen eingesetzt werden, um geringe raum-zeitliche Variationen zu untersuchen. So werden in Abbildung 5.13f die Auswirkungen der globalen Erwärmung für die Temperaturen im Atlantik (vertikaler Schnitt) durch Isoflächen (0°C, 4°C, 20°C und 25°C) dargestellt. In Abbildung 5.13e wird derselbe Datensatz durch ein *direct volume rendering* dargestellt, wobei der Fokus auf den oberen Wasserschichten mit den hohen Temperaturbereichen liegt. Nach Bedarf kann der Anwender hier durch Variation der Transferfunktion andere Ausschnitte aus dem Wertebereich fokussierten, und so in Abbildung 5.13e verdeckte Datenbereiche untersuchen.

Um mehrere Zeitschritte explizit zu vergleichen, bietet sich eine Darstellung paralleler Schnitte entlang der Zeitachse an (vgl. Abb. 5.13c). So können in einer Darstellung auch nicht benachbarte Zeitschritte leicht verglichen werden. Sind ferner einzelne raum-zeitliche Schnitte und deren Zusammenspiel von Interesse, können - analog zur 2D-Darstellung aus Abbildung 5.13a - auch sich kreuzende Schnitte zur Zeitdarstellung eingesetzt werden (vgl. Abb. 5.13d).

5.2.3 Spezielle Methoden zur Darstellung des zeitlichen Bezuges

Die vorgestellten Standardtechniken, welche typischerweise für die Visualisierung von zeitabhängigen Daten im Klimaumfeld eingesetzt werden, sind jedoch nur begrenzt geeignet, die spezifischen Anforderungen bei der Visualisierung von Klimadaten zu adressieren. Insbesondere geraten Standartechniken bei **extrem langen, zeitlich strukturierten Zeitreihen**[7] schnell an die Kapazitätsgrenzen typischer Ausgabegeräte. Da sich bei diesen Zeitreihen die Fragestellungen der Anwendung zumeist auf bestimmte Skalen wie stündlich, täglich, monatlich oder jährlich beziehen, werden bestimmte Sichten auf die Daten generiert. Hierbei spielt vor allem die Datenaggregation mit Hilfe

[7]So enthält beispielsweise die Potsdamer Reihe stündliche Messungen von Temperatur, Niederschlag, relativer Luftfeuchtigkeit, dem Bedeckungsgrad, dem Luftdruck und der Windstärke seit dem Jahre 1893.

statistischer Verfahren eine wichtige Rolle, deren Ergebnisse geeignet dargestellt werden müssen. So werden aus den stündlichen Werten Mittel-, Minimum- oder Maximumwerte für Tage, Monate oder Jahre. Zur Identifikation bestimmter klimatischer Bedingungen (z.B. von heißen Sommern) werden auch weitere Aggregatfunktionen verwendet (z.B. Anzahl heißer Tage mit einer mittleren Temperatur über 25°C).

Auch bei der quantitativen Untersuchung von **Periodizitäten**, dem Vergleich von **Abhängigkeiten mehrerer Merkmale über der Zeit** sowie der Identifikation von **Abweichungen einzelner Zeitschritte** sind den vorgestellten Standardverfahren Grenzen gesetzt. Um sich diesen Herausforderungen zu stellen, sollen im folgenden bekannte Darstellungen aus der Informationsvisualisierung auf deren Anwendbarkeit in diesem Kontext untersucht und neuartige Techniken vorgestellt werden.

Darstellung von zeitlichen Abweichungen räumlicher Daten. Die im vorigen Abschnitt vorgestellten Techniken zur Abbildung der Zeit auf eine räumliche Achse (vgl. Abb. 5.13) sind nicht speziell dafür ausgelegt, die spezifischen Eigenschaften der Zeit einzubeziehen und die damit verbundenen Zielstellungen zu unterstützen. Insbesondere ist für die Anwender von Interesse, wie stark die Daten zu gewissen Zeitpunkten von Interesse von einem Referenzzeitpunkt - unter Einbeziehung der räumlichen Verteilung - abweicht. So geben z.B. die Abbildungen 5.13d-e nur begrenzte Hinweise darauf, in welchen räumlichen Gebieten sich das Klima zu bestimmten Zeitpunkten verändert hat. Besser geeignet zur Beantwortung solcher Fragestellungen ist die Abbildung 5.13c, welche einzelne Zeitscheiben als parallele Ebenen darstellt. Aber auch hier ergibt sich ein relativ hoher kognitiver Aufwand, nicht benachbarte Ebenen miteinander zu vergleichen.

Um diesen Herausforderungen zu begegnen, wurde die Technik **„Differenzmethode“** entworfen und umgesetzt[8] (vgl. Abb. 5.14). Diese bildet - für Daten auf zeitveränderlichen regulären 2D-Gittern - die Zeitschritte auf parallele Schnitte ab und färbt diese, ausgehend von einem gesondert eingefärbten Referenzzeitschritt, nach deren Differenzen zu diesem Referenzzeitschritt ein. Durch Wahl dreier orthogonaler Farbskalen, einer für die positiven Differenzen, einer für die negativen Differenzen sowie einer für die Farbkodierung der Absolutwerte des Referenzzeitschrittes, können so relative Änderungen über die Zeit leicht verfolgt werden, ohne die Absolutwerte dabei aus den Augen zu verlieren.

So werden in Abbildung 5.14a Zunahmen des Meeresspiegeldruckes durch eine Gelb-Schwarz-Skala kodiert, während negative Änderungen durch eine Weiß-Blau-Skala representiert werden. Zusätzlich ist der Startzeitpunkt (als Referenzzeitpunkt), in einer Grau-Skala farbkodiert. Um extreme Klimaveränderungen hervorzuheben, können zusätzlich Transparenzen eingesetzt werden. Dazu können geringe Differenzwerte mit einer größeren Transparenz dargestellt werden, was auftretende Verdeckungen reduziert und so Extremwerte visuell verstärkt (vgl Abb. 5.14a-d).

Ein zweites Beispiel zur Anwendung dieser Methode wird in den Abbildungen 5.14b-d dargestellt. Dort wird die Entwicklung der Temperatur- und Salzgehaltverteilung im Atlantik (als vertikaler Schnitt) dargestellt. Abbildung 5.14b zeigt nach moderaten Temperaturerhöhungen in den ersten Jahren starke Temperaturerhöhungen insbesondere in Oberflächenbereich in den Regionen um 60° nördlicher und südlicher Breite. Ferner sind aber auch leichte Temperaturverringerungen in nördlichen, tieferen Meeresschichten zu konstatieren. Die Abbildungen 5.14c und 5.14d zeigen eine spezielle Sicht auf solche Differenzdaten: die Blickrichtung entlang der Zeitachse. Durch Einsatz der Transparenzen werden dabei Gebiete geringer Veränderung ausgeblendet und enthüllen einen Blick auf die absoluten Daten des Referenzzeitpunktes. Dies sind grüne Gebiete in Abbildung 5.14c und graue bzw. schwarze Gebiete in Abbildung 5.14d. Durch Farbüberlagerungen können Gebiete positiver, negativer und gemischter Anstiege über den gesamten Zeitverlauf auf einen Blick identifiziert, sowie die zeitliche Häufigkeit deren Auftretens anhand der Farbintensität gegenüber der Hintergrundfarbskala abgeschätzt werden.

[8] Die Umsetzung erfolgte im Rahmen eines betreuten Studentenprojektes (Schröder u. Wagenknecht 2003).

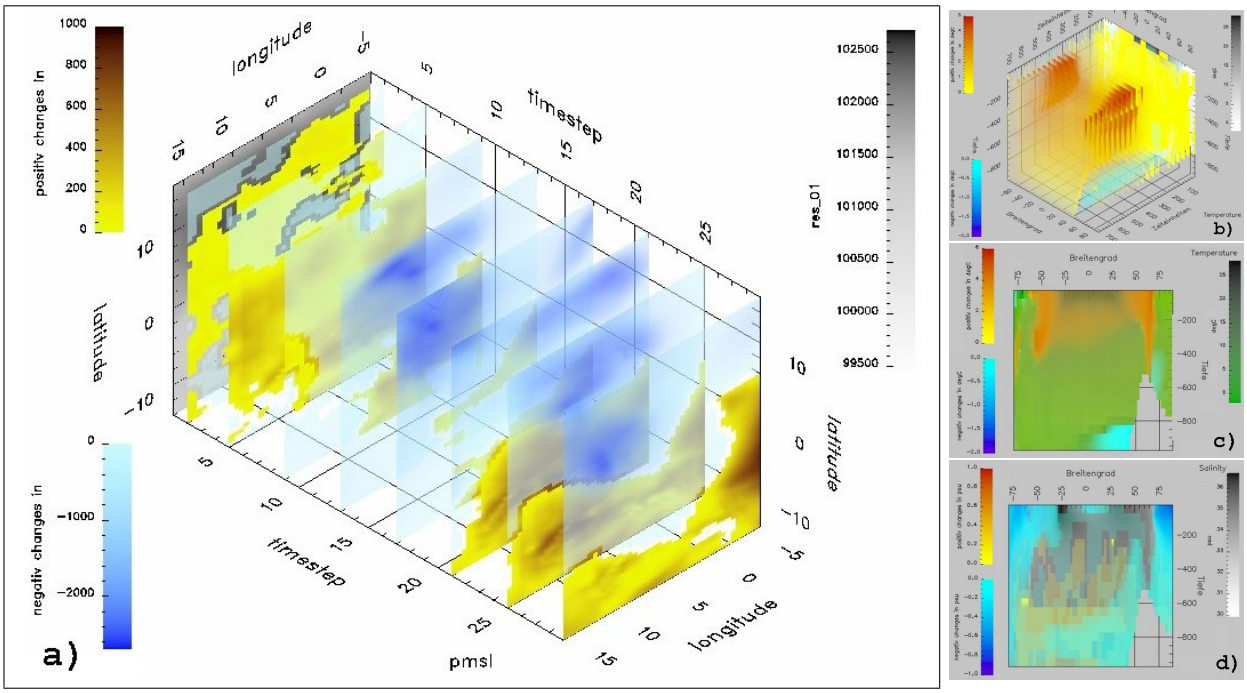


Abbildung 5.14: Darstellungen von zeitlichen Veränderungen mit der Differenzmethode aus SimEnVVis; a) zeitliche Variation des Meeresspiegeldrucks (pmsl); b) zeitliche Variation der Temperatur in einem vertikalen Schnitt im Atlantik; c) wie b, Blickrichtung entlang der Zeitachse; d) wie c, jedoch zeitliche Variation des Salzgehalts

Mit der Differenzmethode wurde also ein neues Werkzeug zur Darstellung zeitveränderlicher Klimadaten auf 2D-Gittern entworfen und umgesetzt. Es unterstützt Vergleiche verschiedener Zeitschritte und erlaubt die detaillierte Untersuchung der zeitlichen Entwicklung von extremen Klimabedingungen. Weiterhin kann die Differenzmethode, wenn auch für die spezielle Darstellung zeitveränderlicher Phänomene ausgelegt, ggf. auch für den Vergleich von räumlichen Ebenen eingesetzt werden.

Auswertung langskaliger Tagesdaten. Um lange Zeitreihen kompakt darzustellen, können pixelbasierte Darstellungen eingesetzt werden (vgl. z.B. Keim u. a. 1993). Diese bilden jeden Datenwert farbkodiert auf genau ein Pixel des Ausgabegerätes ab. So ergeben sich stark verdichtete Bilder, die es ermöglichen, eine große Anzahl von Datenwerten gleichzeitig darzustellen. Im folgenden soll untersucht werden, wie sich solche Darstellungen auch für die kompakte Kodierung von Klimazeitreihen einsetzen lassen.

Abbildung 5.15 zeigt eine pixelbasierte Darstellung der Potsdamer Reihe mit sechs Merkmalen. Über 111 Jahre werden aggregierte Tageswerte auf eine Blau-Weiß-Orange-Skala abgetragen, wobei geringe Werte des jeweiligen Merkmales auf Blau und hohe Werte auf Orange abgebildet werden. Dabei werden die Datenwerte des ersten Jahres 1893 von Januar bis Dezember von links nach rechts auf die erste Pixelzeile, für das Jahr 1894 auf die zweite Zeile u.s.w. abgebildet. So ergibt sich ein schneller *Überblick* über den gesamten Datensatz. Extreme Witterungsbedingungen (in Blau und Orange), können von gemäßigten Bedingungen (in Weiß) unterschieden werden. Weiterhin erlaubt die kompakte Darstellung der Datenwerte einen Bild-zu-Bild-Vergleich der grundlegenden Abhängigkeiten mehrerer Merkmale untereinander (vgl. auch Abb. A.6 im Anhang).

Die in Abbildung 5.15 dargestellte spaltenweise Anordnung ist für die Darstellung von Tagesdaten speziell geeignet, da die Zuordnung von Pixeln zu den zugehörigen Jahren und Monaten so erleichtert wird. Ferner spiegelt sie die durch die Jahreszeiten erzeugten zyklischen Abfolgen geeignet wieder und ermöglicht es so, typische Muster und Abweichungen hiervon zu identifizieren und die Vertei-

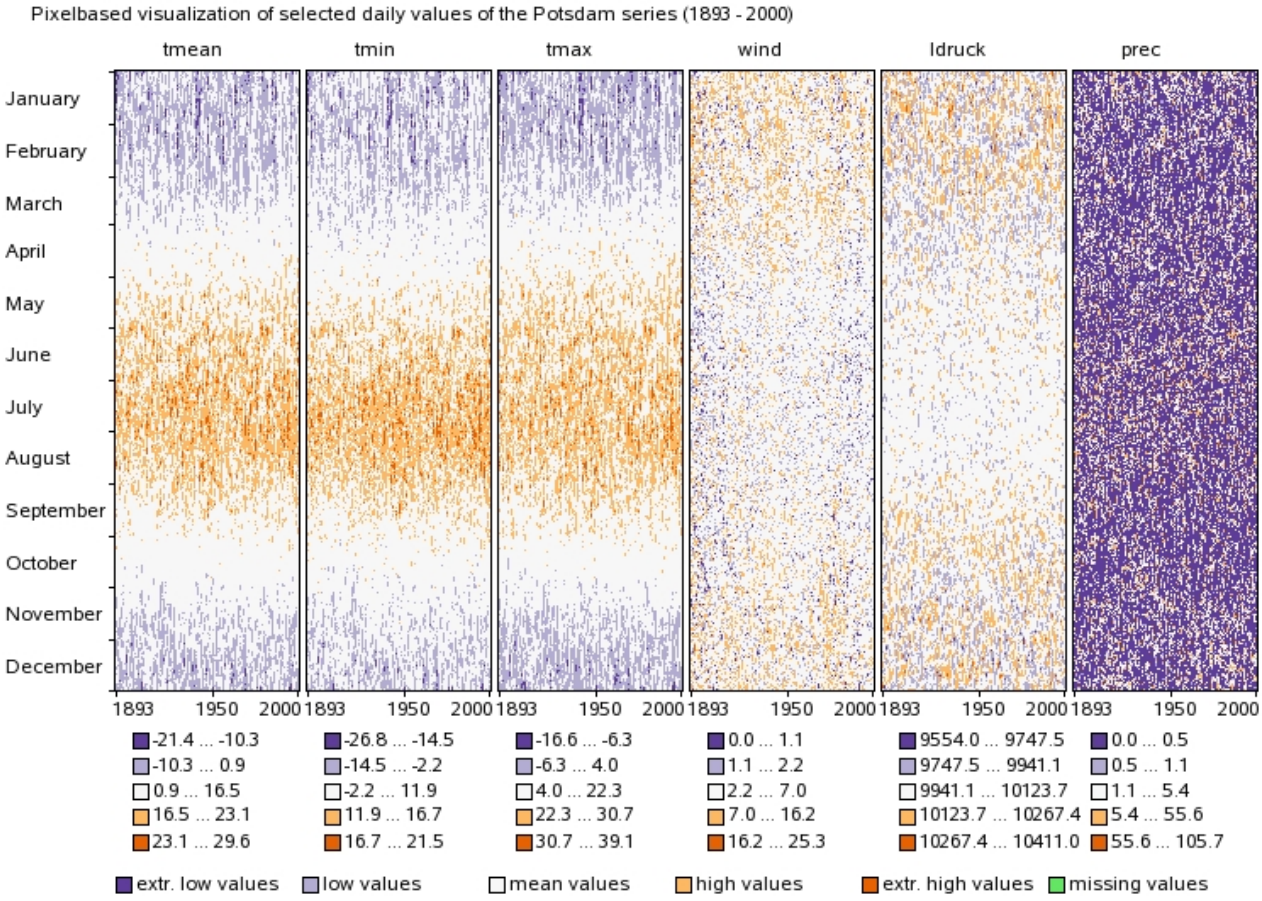


Abbildung 5.15: Pixelbasierte Darstellung der Potsdamer Reihe; Darstellung von sechs Merkmalen für jeden Tag von 1893 bis 2000 in einer Blau-Weiß-Orange-Skala

lung von Ausreißern zu untersuchen. Nachteil von pixelbasierten Darstellungen ist, dass einzelne Werte nur schwer identifiziert und exakt lokalisiert werden können. Alternative Anordnungen (z.B. raumfüllenden Kurven und „recursive pattern"-Techniken in Keim u. a. (1995)) wurden im Rahmen der Arbeit untersucht, sind jedoch aufgrund einer erschwerten Zuordnung der Datenwerte zur Zeitachse und der Unterbrechung jahreszeitlicher Muster nur bedingt für tägliche Klimamessreihen geeignet.

Darüber hinaus bleiben Herausforderungen für weitere Arbeiten. So ist z.B. zu untersuchen ob und bei welchen Klimamessreihen Spiraldarstellungen oder zyklische Darstellungen (vgl. z.B. Weber u. a. 2001) einsetzbar sind, die insbesondere dazu geeignet sind, versteckte Periodizitäten zu identifizieren. Weiterhin müssen erste Versuche mit der „Two-Tone"-Farbabbildung aus Saito u. a. (2005) vertieft werden (vgl. Abb. A.7 und Abb. A.8 im Anhang).

Uni-variate Darstellung jährlicher Klimadaten. Liegt der Fokus der Untersuchung - nicht wie im vorangegangen Abschnitt auf täglichen, sondern - auf jährlichen Phänomenen, steht (bei der gleichen zugrunde liegenden Messreihe) mehr Darstellungsplatz für die Datenwerte zur Verfügung, wodurch eine Lokalisation und Identifikation einzelner Werte vereinfacht wird. Um ferner auch Periodizitäten in solchen Daten explorieren zu können, wurde speziell für die uni-variate Visualisierung von jährlichen Merkmalen die neue „Rechteckmethode" entworfen und umgesetzt (vgl. hierzu auch Nocke u. a. (2004)).

Abbbildung 5.16 zeigt zwei Darstellungen dieser Methode. Zur Repräsentation von aggregierten, extreme Sommer identifizierenden Merkmalen werden dabei Datenwerte auf farbkodierte Quadrate

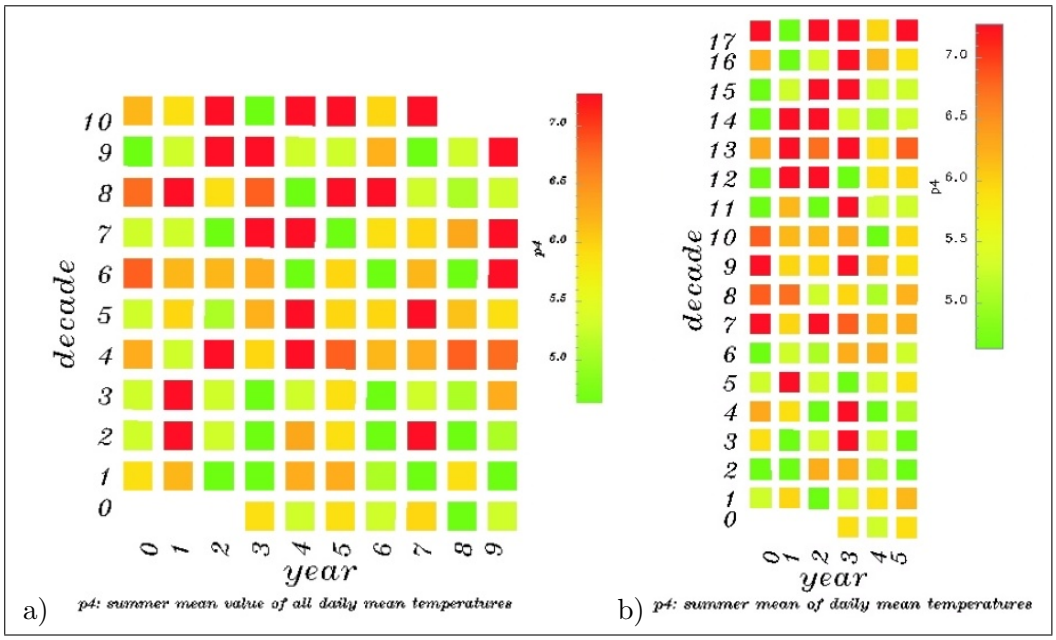


Abbildung 5.16: Darstellung eines Datensatzes zur Analyse von extremen Sommern mit der Rechteckmethode (Merkmal „Sommermittelwert der mittleren Tagestemperaturen“); a) Spaltenanzahl von zehn Jahren; b) Adaptierte Spaltenanzahl von sechs Jahren zur Identifikation von Perioden

abgebildet (Grün-Gelb-Rot-Skala). Diese Quadrate werden dann in einem rechteckigen Layout angeordnet, so dass der erste Zeitschritt (Jahr 1893) auf den Würfel in der linken unteren Ecke abgebildet wird, gefolgt von den nächsten Zeitschritten in der untersten Zeile. Anschließend wird dann beginnend mit dem nächst folgenden Zeitschritt der erste Quadrat der zweiten Reihe farbkodiert, gefolgt von den anderen Zeitschritten in dieser Zeile, und so weiter. Das am weiten rechts dargestellte Quadrat der oberen Zeile enthält den letzten Zeitschritt (Jahr 1997). Eine intuitive Spaltenanzahl für einen solchen Datensatz ist zehn, um die einzelnen Dekaden und die Werte der zugehörigen Jahre leicht lokalisieren zu können. Hierfür wurde ferner ein Offset eingefügt, so dass der erste Zeitschritt nicht notwendigerweise in der ersten Spalte in der untersten Zeile liegt (wie oben beschrieben), sondern (modulo 10) in der dem Jahr entsprechenden Spalte positioniert werden kann (vgl. Abb. 5.16a).

Zu den Interaktionstechniken auf dieser Technik zählen das Picking einzelner Datenwerte sowie die Variation der Spaltenanzahl, um ggf. von der Dekadeneinteilung abweichende Periodizitäten aufdecken zu können. So kann dann jede einzelne Zeile als Periode aufgefasst werden und auftretende zeitlich wiederkehrende Muster auf variierenden Zeitskalen interaktiv exploriert werden. So konnte mit der entwickelten Technik festgestellt werden, dass bei einer Periode von sechs Jahren ein relevantes Muster[9], welches eine Häufung heißer Sommer (in rot) in den ersten vier Spalten ausweißt, während in den letzten zwei Spalten mit Ausnahme des Jahres 1997 kein heißer Sommer auftritt (vgl. Abb. 5.16b). Dieses Muster, welches auch durch ein lokales Maximum in der Autokorrelationsfunktion ausgewiesen ist, gibt Rückschlüsse darüber, dass nach einer Periode von maximal drei heißen Sommern im letzten Jahrhundert wieder mindestens zwei kühlere Sommer folgten.

Für Zeitreihen mittlerer Länge ist die vorgestellte Rechteckmethode ein geeignetes Explorationswerkzeug und erlaubt neben der Lokalisation und Identifikation einzelner Datenwerte, grundlegende Trends sowie Periodizitäten in den Daten zu untersuchen. So ergaben sich für die Anwender für den vorliegenden Datensatz neue Einsichten: neben der grundsätzlichen Zunahme heißer Sommer ließ

[9] nicht nur für das in Abbildung 5.16 dargestellte Merkmal „Sommermittelwert der mittleren Tagestemperaturen“, sondern auch für die anderen fünf Merkmale des Datensatzes

sich insbesondere in den 70er, 80er und frühen 90er Jahren des vergangenen Jahrhunderts eine stabile Periode im Auftreten heißer Sommer feststellen. Darüber hinaus hat sich die Rechteckmethode auch als geeignetes Werkzeug zur Untersuchung von geclusterte Zeitreihen herausgestellt (vgl. Abs. 6.1).

Bei der Darstellung längerer Zeitreihen sind die Möglichkeiten dieser Darstellungstechnik aufgrund des begrenzten Darstellungsplatzes jedoch begrenzt. Zur Lösung dieses Problems kann Zooming & Panning eingesetzt werden, was jedoch eine reduzierte Nutzerorientierung nach sich zieht. Weiterhin ist diese Technik nicht geeignet, mehrere Merkmale miteinander zu vergleichen. Eine mögliche Lösung dieses Problems ist es, analog zur multi-variaten Darstellung räumlicher Daten, anstatt der Rechtecke Ikonen darzustellen, die mehrere Merkmale gleichzeitig kodieren. Im Rahmen dieser Arbeit wurde ein anderer Weg zur Lösung der genannten Herausforderungen beschritten: der Einsatz der Technik „Themenfluss“.

Multi-variate Darstellung jährlicher Klimadaten. Ursprünglich zur Dokumentenvisualisierung entworfen (vgl. Havre u. a. 2002a, b), erlaubt die Technik „Themenfluss“ eine kompakte Darstellung von Zeitreihen, wobei Abhängigkeiten mehrerer Merkmale in einer Darstellung untersucht werden können. In der ursprünglichen Technik wird die Häufigkeit des Auftretens bestimmter Worte für jeden Zeitschritt gezählt und diese zu einer speziellen Art von Balkendiagrammen für jeden Zeitschritt abgebildet (mit der Zeit als zentrierter x-Achse). Durch die Interpolation zwischen den Balken (z.B. über Bezier-Splines) wird der Eindruck eines Flusses mit konstanter Farbe für jedes Wort erzeugt. Dadurch können zeitliche Änderungen und Merkmalsabhängigkeiten intuitiv exploriert werden.

Zur Anwendung auf Klimazeitreihen wurden Tests mit einer vereinfachten Version dieser Technik durchgeführt (vgl. Abb. 5.17). Hierbei werden fünf aggregierte Merkmale zur Identifikation hei-

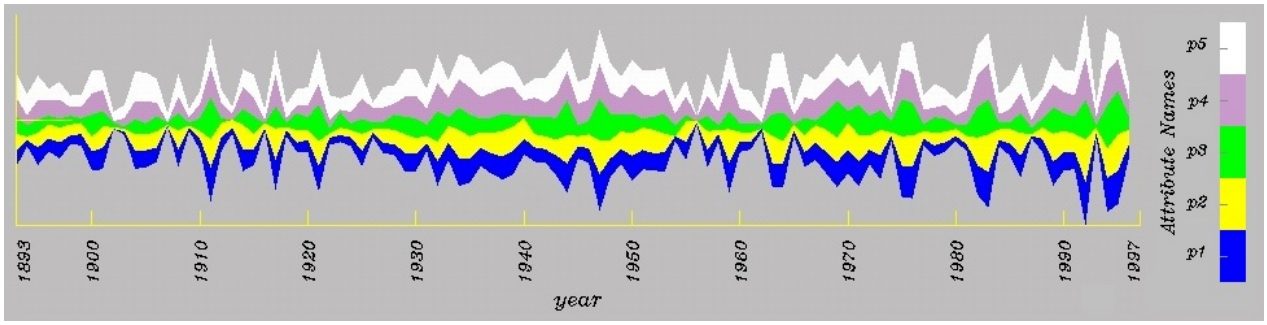


Abbildung 5.17: Vereinfachte Darstellung der Technik Themenfluss zur Darstellung von zeitlichen Trends bei den Sommereigenschaften (fünf aggregierte, extreme Sommer identifizierende Merkmale für die Messstation Potsdam)

ßer Sommer an der Station Potsdam abgebildet, um das Auftreten extremer Sommerbedingungen zu studieren. Die Datenwerte der einzelnen Merkmale werden auf $[0, 1]$-Intervalle normalisiert und anstatt der Dokumentenhäufigkeiten abgebildet. In Abbildung 5.17 werden diese dann unter Anwendung von linearer Interpolation miteinander verknüpft. Hierbei repräsentieren dann Trapeze den „Fluss“ eines Merkmals von einem Zeitschritt zum nächsten. Kleine Parameterwerte oder ein „dünner Fluss“ repräsentieren extrem kalte Sommer, während hohe Parameterwerte oder ein „breiter Fluss“ extrem heiße Sommer wiederspiegeln. Als erster Eindruck aus Abbildung 5.17 ergibt sich damit ein Anstieg heißer Sommer in der zweiten Hälfte des 20. Jahrhunderts. Auch Abhängigkeiten einzelner Merkmale können visuell identifiziert und deren Anteil an allgemeinen Klimaveränderungen untersucht werden. In diesem Fall lässt sich z.B. feststellen, dass das Merkmal p1 („totale Wärme: Summe der täglichen Maximumtemperaturen $\geq 20°C$“, blau) stark mit dem Merkmal p2 („Anzahl heißer Tage“, gelb) verknüpft ist. Dies ist für die Klimaforscher ein Indiz dafür, dass der

Anstieg der totalen Wärme in starkem Maße durch einen Anstieg der täglichen Maximaltemperaturen und weniger durch länger andauernde, moderate Wärmeperioden verursacht wird. Weiterhin können die Klimaforscher anhand dieser Darstellungstechnik auch konstatieren, dass eine Erhöhung der Maximaltemperaturen eine Erhöhung der mittleren Temperaturen nach sich zieht (Merkmal p4, fliederfarben).

Die Technik „Themenfluss“ erlaubt, allgemeine Trends und Merkmalsabhängigkeiten über die Zeit kompakt wiederzugeben. Sie hat großes Potential für die Anwendung im Klimaumfeld; so konnten für den vorliegenden Datensatz detaillierte Untersuchungen durchgeführt werden, die mit typischerweise im Klimaumfeld eingesetzten Techniken wie z.B. Zeitgraphen nicht durchführbar sind. Dies wurde auch in einer gemeinsamen Veröffentlichung mit den Klimaforschern in einer anwendungsspezifischen Zeitschrift dokumentiert (vgl. Böhm u. a. 2004).

Nachteil dieser Technik ist, dass die Merkmale nicht gleich behandelt werden. So erhalten Merkmale in der Mitte des Flusses stärkeres visuelles Gewicht, während die Merkmale an den Rändern mit stärkeren Verzerrungen wahrgenommen werden (Sinustäuschung). Deswegen kann eine detaillierter Vergleich von Werten verschiedener - insbesondere nicht benachbarter - Merkmale erschwert werden, weswegen geeignete Interaktionsfunktionalität zur Vertauschung der Merkmale bereitzustellen ist.

5.2.4 Diskussion

In diesem Abschnitt wurden verschiedene Techniken zur Darstellung des zeitlichen Bezuges auf deren Einsetzbarkeit im Klimaumfeld hin untersucht. Dazu wurden aus dem Visualisierungsumfeld bekannte Techniken umgesetzt und an die Erfordernisse der Anwendung angepasst. Ferner wurden mit der Differenzmethode und der Rechteckmethode zwei einfache, neue Techniken entworfen, die intuitiv sind und die Aufgaben der Anwendung sehr gut unterstützen. So konnten mit den umgesetzten Methoden zum Teil neue Einsichten zu bisher verborgenen Eigenschaften der Daten gewonnen werden.

Gruppe	Technik/Layout	Anz.MM.	Aufgaben	Zielstellungen	Int.u.Nav.	Probleme
Zeitgraphen (OpenDX)	Tabellendarstellung	∼100	Überblick	Verteil.,Identif.,Lokal.	Pan, Zoom	begrenzte Komplexität
	Kombinierte Darst.	max. 6	DoD	Vergl.,Identif.,Lokal.	Pan, Zoom	
	Vergl. 2er Kurven	2	Details	Vergl.,Identif.,Lokal.	Pan, Zoom	
Ersetzung einer räuml. Achse		max 3	Überblick DoD	Verteil.,Identif.,Lokal.	Pan, Zoom Piking	
Differenzmethode(OpenDX)		1	Überblick DoD	Vergl., Verteil., Identif.,Lokal.	Pan, Zoom	
Pixelbasierte Darstellungen (C++/MFC)	zeilenweises Layout	beliebig	Überblick	Verteil., Vergl.,Lokal.		Werte identif.
Rechteckmethode(OpenDX)		1	Überblick DoD	Identif.,Lokal., Trends, Extr.	Pan, Zoom, Perioden	Platzverbrauch
Themenfluss (OpenDX)		∼15	Überblick DoD	Verteil., Vergl. Identif., Lokal., Extr.	Merkmale vertauschen	Wichtung d. Merkmale

Tabelle 5.2: Wichtige Eigenschaften der umgesetzten Zeitdarstellungstechniken
Legende:

Anz.MM - Anzahl darstellbarer Merkmale, *Int.u.Nav.* - Interaktions- u. Navigationstechniken,
DoD - Details-on-Demand, *Lokal.* - Lokalisieren von Werten,
Extr. - Aufspüren von Extremen, *Vergl.* - Vergleich,
Identif. - Identifizieren von Werten, *Verteil.* - Untersuchung von Werteverteilungen

Wichtiger Aspekt beim Entwurf der Techniken war es, den Anwendern eine Vielzahl von Inter-

aktionsmöglichkeiten an die Hand zu geben, um insbesondere die speziellen Eigenschaften der Dimension Zeit zu unterstützen. Hierzu zählen die Exploration und der Vergleich von klimatischen Extremen über die Zeit sowie die Untersuchung von Zyklen in Klimadaten. Tabelle 5.2 fasst die umgesetzten Techniken und deren wichtigste Eigenschaften zusammen.

5.3 Darstellung von Klimadaten im Merkmalsraum

Bisher konzentriert sich die Visualisierung von Klimadaten weitgehend auf deren Darstellung im räumlichen und zeitlichen Bezug. Es hat sich jedoch in anderen Anwendungsbereichen gezeigt, dass auch die Analyse raum-zeitlicher Daten von deren interaktiver Darstellung im (von den abhängigen Variablen aufgespannten) Merkmalsraum profitieren kann. Bisher werden für die Untersuchung von Merkmalsabhängigkeiten in Klimadaten überwiegend einfache Scatterplots eingesetzt, bzw. die Analyse des Merkmalsraumes an den räumlichen oder zeitlichen Bezug gekoppelt. So bleibt diese Analyse zumeist auf bi-variate Abhängigkeiten beschränkt.

Um auch in Raum- und zeitlichen Darstellungen schwer zu identifizierende Abhängigkeiten der Merkmale aufdecken zu können, haben sich im Informationsvisualisierungsumfeld eine Vielzahl von Darstellungs- und Interaktionstechniken etabliert. Im folgenden sollen nun exemplarisch anhand zweier Standardvisualisierungstechniken die Einsatzmöglichkeiten von Darstellungen des Merkmalsraumes bei Klimadaten illustriert werden.

Scatterplot-Matrizen Scatterplot-Matrizen sind eine spezielle Form der Panelmatrizen, in denen matrixförmig verschiedene bi-variate Darstellungen miteinander kombiniert werden, um den m-dimensionalen Merkmalsraum zu veranschaulichen (vgl. Wong u. Bergeron 1997). Bei Scatterplotmatrizen werden im speziellen verschiedene Scatterplots zweier Merkmale kombiniert, in denen jeweils die beiden Merkmale auf zwei senkrechte Achsen abgebildet werden.

Entsprechend der Ausprägungen eines Beobachtungsfalles werden dann Punkte innerhalb der Ebene positioniert. Diese ermöglichen Rückschlüsse über die grundlegende Verteilung der Datenwerte. Abbildung 5.18a zeigt eine solche Scatterplotmatrix für einen atmosphärischen Datensatz mit den Merkmalen Temperatur (t), relative Luftfeuchte (q) sowie den drei vektoriellen Komponenten (u, v und w) der atmosphärischen Strömung. Um Abhängigkeiten höherer Ordnung zu untersuchen, können dabei durch Brushing & Linking spezielle Wertekombinationen von Interesse ausgewählt und in allen Scatterplots farblich hervorgehoben werden (rot).

Nachteil dieser Art der Darstellung ist, dass nicht zu erkennen ist, wie viele Datenwerte durch einen dargestellten Punkt repräsentiert werden. Um diesem Problem zu begegnen, kann Farbe oder Transparenz eingesetzt werden, um die Häufigkeit der durch einen Punkt repräsentierten Beobachtungsfälle darzustellen. Weil die Farben einzelner Punkte nur schwer unterscheidbar sind, können zusätzlich gewisse Gebiete des Merkmalsraumes zusammengefasst und je nach den auftretenden Häufigkeiten gleichmäßig kodiert werden. In den Abbildungen 5.18b und 5.18c wurden die Wertebereiche der Merkmale dafür gleichmäßig in Intervalle unterteilt, was zu rechteckigen homogenen Darstellungsbereichen führt. Zusätzlich können ausgewählte Beoabachtungsfälle (in rot) auch in der Transparenz variiert werden, um auch für sie die Intervallhäufigkeiten untersuchen zu können (vgl. Abb. 5.18c).

Parallele Koordinaten. Im Gegensatz zu Scatterplotmatrizen, wo die Lage eines Beobachtungsfalles im Merkmalsraum durch wiederholte Punktdarstellung in bi-variaten Scatterplots erfolgt, wird dieser bei Streckenzugdarstellungen durch einen Linienzug zwischen beliebig angeordneten Achsen repräsentiert. Ein bekanntes Beispiel hierfür sind die Parallelen Koordinaten, bei denen die Achsen parallel angeordnet werden. Abbildung 5.19 illustriert eine solche Darstellung am Beispiel des Atmosphärendatensatzes.

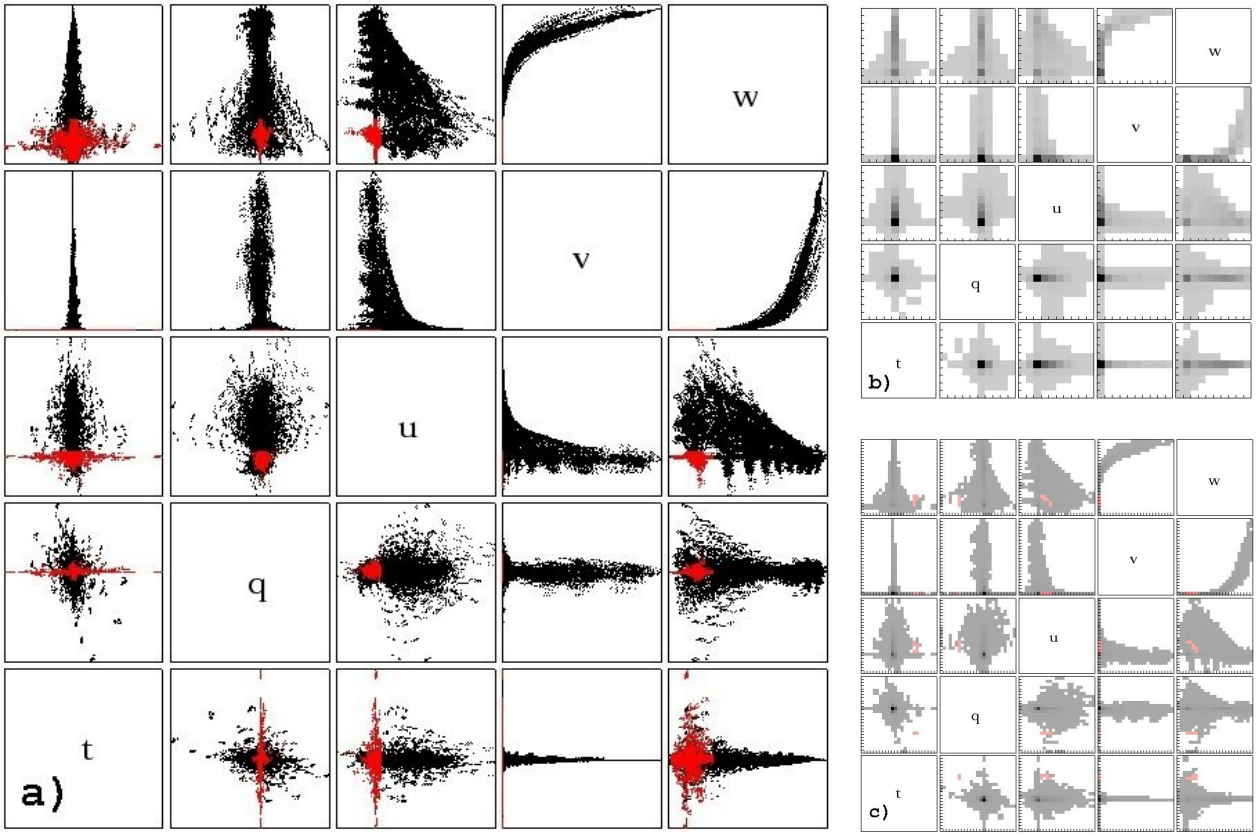


Abbildung 5.18: Scatterplotmatrix-Darstellung einer Atmosphärensimulation aus SimEnvVis; a) punktuelle Darstellungen unter Datenselektion (rot); b) Intervalldarstellung mit Häufigkeiten (größere Intervalle); c) Intervalldarstellung mit Häufigkeiten (feinere Intervalle) und Datenselektion

In der Standardvariante können so kompakt auch eine hohe Anzahl von Beobachtungsfällen und Merkmalen repräsentiert werden (vgl. Abb. 5.19a) sowie Beobachtungsfälle bestimmter Wertebereiche selektiert und die zugehörigen Linienzüge farblich hervorgehoben werden (vgl. Abb. 5.19b). Um in Bereichen hoher Liniendichte weiterhin die Übersicht zu erhalten, können nun analog zu Scatterplotmatrizen ein Alpha-Blending in Kombination mit Intervallbildung durchgeführt werden (vgl. Abb. 5.19c).

Begrenzt werden solche Darstellungen durch die Anzahl gleichzeitig darstellbarer Achsen sowie eine zunehmende Überfrachtung der Darstellung bei sehr großen Datenmengen.

In diesem Abschnitt wurde das Potential von Darstellungen von Klimadaten im Merkmalsraum am Beispiel zweier weit verbreiteter Techniken demonstriert. Gerade durch Interaktionen auf solchen Darstellungen lassen sich neue Erkenntnisse generieren, welche auf in diesem Umfeld üblichen statischen, einzelnen Scatterplots nur begrenzt möglich sind.

Um Häufigkeitsinformationen von Merkmalskombinationen in diesen Darstellungen auch für große Datenmengen, wie sie bei Klimasimulationen auftreten, besser identifizieren zu können, wurden für beide Techniken Methoden zur Intervallbildung in Kombination mit Alpha-Blending vorgestellt. So erhält der Anwender bei Anwendung einer Gruppierung der Wertebereiche (vgl. Abb. 5.18b,c sowie 5.19c,d) einen schnellen Überblick über die Verteilung im Datensatz, kann aber durch Reduktion der Intervalle bis auf Pixelgröße (vgl. Abb. 5.18a sowie 5.19a,b) flexibel Details nachladen. Tabelle 5.3 fasst die Eigenschaften der beiden Techniken zusammen.

Natürlich ergeben sich hier noch eine Vielfalt von Aufgaben für zukünftige Arbeiten. Hierzu zählt

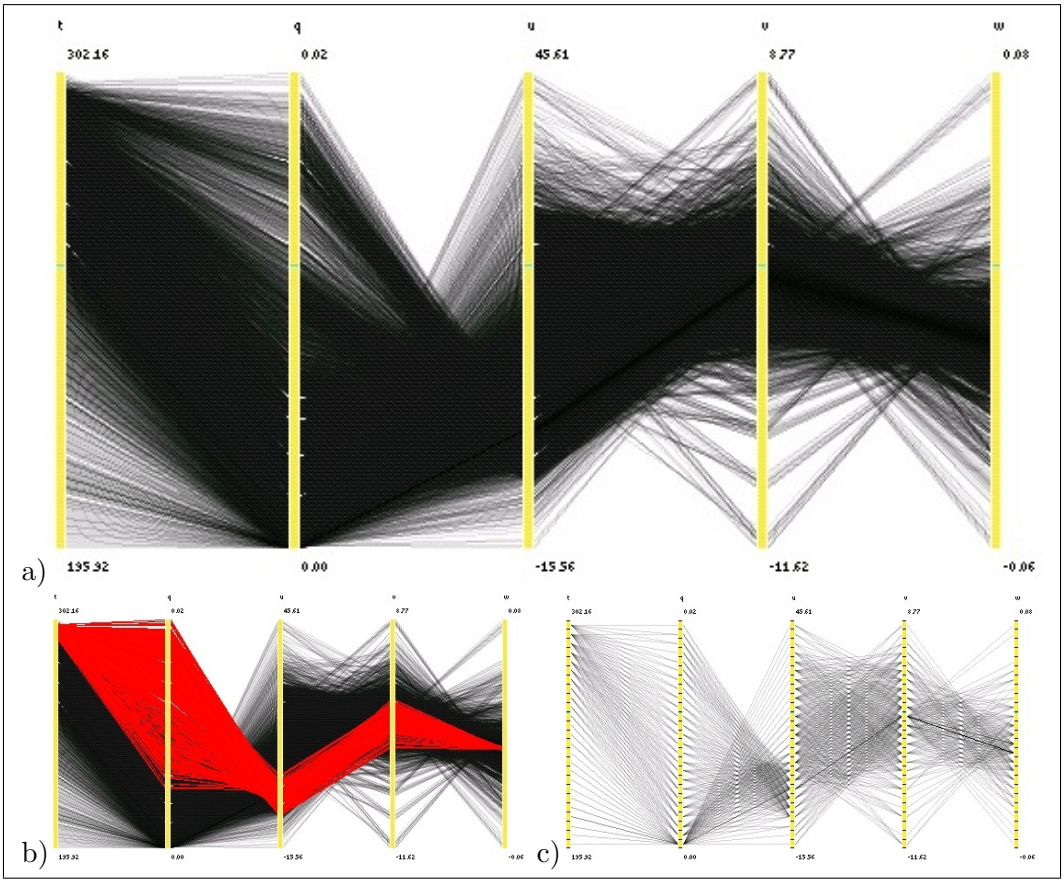


Abbildung 5.19: Parallele Koordinaten-Darstellung einer Atmosphärensimulation aus SimEnvVis; a) Standarddarstellung; b) Standarddarstellung mit Datenselektion; c) Intervalldarstellung

Gruppe	Techn./Lay.	Anz.MM.	Aufgaben	Zielstellungen	Interaktion/Navigation	Probleme
Scatterplot-Matr.	Standard	max 8	Details	Verteil., Identif.,	Merkm. vertauschen,	Häufigkeiten
(QT/OpenGL)	Intervall		DoD	Vergl., Gruppier.	Alpha-Blending,	
Parallele Koord.	Standard	max 15	Details	Verteil., Identif.,	Brushing & Linking,	vis. Clutter
(QT/OpenGL)	Intervall		DoD	Vergl., Gruppier.		

Tabelle 5.3: Wichtige Eigenschaften der umgesetzten Darstellungen im Merkmalsraum
Legende:

Anz.MM - Anzahl darstellbarer Merkmale, *Merkm.* - Merkmale
DoD - Details-on-Demand, *Vergl.* - Vergleich,
Gruppier. - Untersuchung von Gruppierungen, *Verteil.* - Untersuchung von Werteverteilungen
Identif. - Identifizieren von Werten, *vis. Clutter* - Überladung der Darstellung durch

der Einsatz moderner Brushing-Konzepte wie Smooth oder Angular Brushes (vgl. z.B. Doleisch u. Hauser 2002; Chen 2003; Doleisch u. a. 2004) sowie die konsequente Kopplung von Darstellungen im Merkmalsraum mit Darstellungen im räumlichen und/oder zeitlichen Bezug (vgl. hierzu auch Abs. 6.1).

5.4 Vergleichende Visualisierung von Klimadaten auf abweichenden Gittern

In den bisher vorgestellten Ansätzen wurden Visualisierungstechniken beschrieben, welche die Visualisierung einer Datenmenge für sich behandelten. Allerdings reichen solche Darstellungen nicht aus, wenn mehrere Datenmengen miteinander verglichen werden sollen. Insbesondere ist die Validierung von simulierten Modelldaten miteinander oder mit Messdaten in deren räumlichen und zeitlichen Kontext in verschiedenen Anwendungen eine wichtige Aufgabe. Eine solche vergleichende Analyse hat insbesondere zum Ziel, örtliche und/oder zeitliche Regionen mit ähnlichen sowie unähnlichen Werten zu lokalisieren und die Abweichungen und deren Verteilungen untersuchen zu können. Eine breite Palette von Faktoren kann zu Abweichungen in den Daten und den resultierenden Darstellungen führen (vgl. Pagendarm u. Post 1995): verschiedene physische Phänomene, verschiedene Bedingungen (Experiment oder Simulation), Messungenauigkeiten, Rauschen oder numerische Ungenauigkeiten sowie variierende automatische oder Visualisierungsverfahren. Die Visualisierung solcher Daten wird unter dem Begriff *vergleichende Visualisierung* (engl.: comparative visualization) zusammengefasst.

Herausforderungen hierbei sind, dass die Daten

- in verschiedenen Dimensionalitäten (z.B. 2D-Oberflächen-Messungen von Temperaturen und Niederschlagswerten und die 3D-Modellierung von atmosphärischen Bedingungen),
- auf verschieden räumlichen Gittern (z.B. regelmäßig, blockstrukturiert, gestreut),
- in unterschiedlichen zeitlichen Skalen (z.B. Monate und Jahre)
- mit mehreren abhängigen Variablen (z.B. Temperatur und Luftfeuchtigkeit), die gleichzeitig verglichen werden sollen,

vorliegen können. Ein häufiger Ansatz für die vergleichende Analyse solcher Daten ist z.B. im Klimaumfeld, mehrere Bilder nebeneinander zu legen und den Nutzer diese Bild-zu-Bild vergleichen zu lassen. Problem dieser Vorgehensweise ist jedoch der hohe mentale Aufwand, einen Bildpunkt der einen Darstellung mit einem Bildpunkt in einem zweiten Bild zu vergleichen, was bei mehr als zwei Datensätzen in mehr als zwei Darstellungen noch wesentlich erschwert wird. Auch die häufig angewandte Abbildung eines Gitters auf das andere kann – neben dem Problem der Zulässigkeit einer solchen Abbildung – eine Datenreduktion- oder Datenverfälschung verursachen, bei der wichtige Informationen für die vergleichende Analyse verloren gehen können.

Die vergleichende Visualisierung ist in der Visualisierungsliteratur eher unterrepräsentiert. Zwar unterstützen viele Techniken auch den Vergleich von Variablen, jedoch wurde die grundlegende Vorgehensweise von Pagendarm u. Post (1995) nur begrenzt weiterentwickelt. Ferner stammt ein Großteil der Literatur zu diesem Thema aus den neunziger Jahren.

In diesem Abschnitt soll ein neuer Ansatz zur vergleichenden Visualisierung vorgestellt werden, welcher ein allgemeines Vorgehen beschreibt und dadurch eine breite Palette an Fragestellungen und abweichende Datencharakteristika der zu vergleichenden Datensätze unterstützt. Er kombiniert die Vorteile existierender Ansätze zur vergleichenden Visualisierung miteinander und gibt ein konzeptionelles Gerüst vor, wie existierende Techniken für den visuellen Vergleich auch von Daten auf abweichenden Gittern eingesetzt werden können. Zur Demonstration der Leistungsfähigkeit des neuen Ansatzes wurde ein flexibel erweiterbares Framework zum Vergleich von Klimadatensätzen

beispielhaft für reguläre Gitter abweichender Auflösungen umgesetzt.

Der Abschnitt gliedert sich wie folgt: zuerst werden prinzipielle Ansätze der vergleichenden Visualisierung diskutiert (vgl. Abs. 5.4.1). Im Anschluss daran erfolgt eine Problemanalyse (vgl. Abs. 5.4.2), auf der aufbauend der neue Ansatz entworfen wird (vgl. Abs. 5.4.3). Zum Abschluss erfolgt eine Zusammenfassung, welche die Vorstellung eines prototypischen Frameworks sowie umgesetzter Visualisierungsmethoden zur vergleichenden Visualisierung im Klimaumfeld einschließt (vgl. Abs. 5.4.4).

5.4.1 Prinzipielle Ansätze für die vergleichende Visualisierung

Von Pagendarm u. Post (1995) wird bei der *vergleichenden Visualisierung* zwischen zwei Vorgehensweisen bei der Analyse unterschiedlicher Daten unterschieden:

- *image level* - Vergleich auf Basis der bei der Visualisierung generierter *Bild-Daten*,
- *data level* - Vergleich auf der Basis der *Daten im Beobachtungsraum* mit anschließender Visualisierung.

In diesem Abschnitt sollen diese beiden grundlegenden Ansätze vorgestellt und verschiedene Arbeiten aus der Literatur zugeordnet werden. Darüber hinaus werden existierende Visualisierungsverfahren bezüglich ihrer Potenz zur Anwendung in der vergleichenden Analyse, insbesondere beim Auftreten abweichender Gitter, diskutiert.

5.4.1.1 Bildbasierter Ansatz (image level)

Prinzip. Auf dem ***image level*** (vgl. Abbildung 5.20) werden aus den Originaldaten (DS_i) mit Hilfe geeigneter Visualisierungstechniken Bilddaten erzeugt. Diese Bilddaten können dann zum einen in verschiedenen Sichten (z.B. als symmetrische Halbbilder in Pagendarm u. Post 1995) dargestellt werden. Um den hohen mentalen Aufwand beim Vergleich dieser nebeneinander liegenden

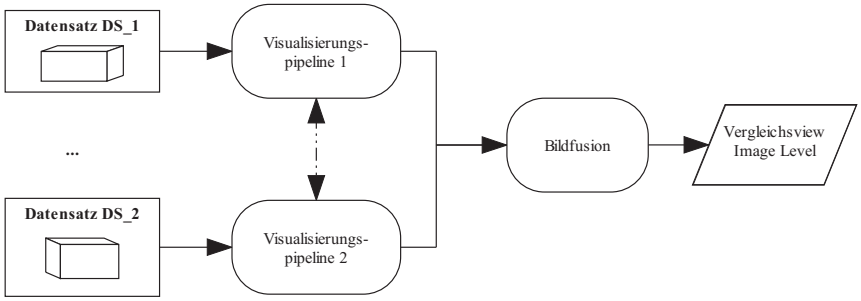


Abbildung 5.20: Vergleichende Visualisierung auf dem *image level* nach Pagendarm u. Post (1995)

Sichten zu reduzieren, können die beiden Darstellungen z.B. mittels Brushing & Linking miteinander gekoppelt werden. Zum anderen können diese Bilddaten miteinander gemischt werden, um ein fusioniertes Bild zu erzeugen. Hierbei können verschiedene Metriken verwendet werden. Diese Fusionierung ermöglicht neben dem Aufdecken von Zusammenhängen in den Daten auch die Untersuchung der Ergebnisse unterschiedlicher Visualisierungstechniken, und damit deren Validierung (vgl. z.B. Pagendarm u. Post 1995; Kim u. Pang 1997; Zhou u. a. 2002a).

Methoden. Zum Vergleich von Visualisierungen auf dem *image level* unterscheiden Zhou u. a. (2002a) drei Vorgehensweisen:

1. wahrnehmungsbasierter Vergleich (Symmetrien, Neben- und Übereinanderlegen von Bildern),
2. Bildraummetriken (Block-Matching z.B. durch Kreuzkorrelation) und

3. Frequenzraummetriken.

Häufig angewendet wird der *wahrnehmungsbasierte Vergleich*, wobei verschiedene Darstellungen nebeneinander oder übereinander gelegt werden. Um den Vergleich der Datensätze in verschiedenen Sichtfenstern zu vereinfachen, können diese mittels – örtlichem und zeitlichen – Brushing & Linking synchronisiert werden (vgl. hierzu auch Ansätze zur „Multiple view visualization“, Robertson 2006). Der *wahrnehmungsbasierte Vergleich* wird insbesondere dann angewendet, wenn simulierte Daten mit Kamerabildern aus Experimenten verglichen werden sollen (Pagendarm u. Post 1995). So werden beispielsweise in der Strömungsvisualisierung Experimentdaten mit Aufnahmen einer realen Kamera verglichen um Modelle zu evaluieren. So kombinieren Pagendarm u. Walter (1995) in einem Windtunnel-Experiment generierte Ölflussbilder mit abstrakten Features aus simulierten Modelldaten wie Oberflächenreibungslinien (engl.: „skin friction lines“), mit 3-dimensionalen Vortex-Kernen und mit Stoßfronten. Diese Art des Vergleichs ermöglicht, auch kleine Abweichungen von Experiment und Simulation an einer Oberfläche zu untersuchen und ggf. Gitterauflösungen des Modells in Regionen ungenauen Modellverhaltens anpassen zu können. Ein analoges Beispiel zur Analyse von Schockwellen findet sich in Pagendarm u. Post (1995), wo die Bildfusion durch eine Überlagerung im RGB-Raum erfolgt.

Über den wahrnehmungsbasierten Ansatz hinaus wurde eine Vielzahl von *Metriken* zum Vergleich von Bildern vorgeschlagen (vgl. z.B. Ahumada 1993, für eine Übersicht). Diese wurden vor allem zur Validierung von (verlustbehafteten) Bildkodierdierungsverfahren entworfen, können z.B. aber auch zur Validierung von realitätsnahen Bildern eingesetzt werden (vgl. Rushmeier u. a. 1995). Für die Validierung von Simulations- und Visualisierungstechniken sowie zum Vergleich der zugrunde liegenden Datensätze sind insbesondere *Bildraummetriken* zur Findung von räumlichen Korrelationen geeignet. So schlagen Pagendarm u. Post (1995) vor, das Korrelationsintegral zwischen beiden Bildern zu verwenden, um räumliche Verschiebungen von Mustern untersuchen zu können. Ein anderes Beispiel für *Bildraummetriken* ist die von West u. Machiraju (1998) vorgeschlagene Multipart-Metrik, die insbesondere gegenüber lokalem Rauschen nicht anfällig ist und neben der Stärke der Abweichung auch ihren Typ bestimmt.

Der Vorteil des *image level*-Ansatzes ist seine Einfachheit und flexible Anwendbarkeit. Beim wahrnehmungsbasierten Vergleich können die menschlichen Wahrnehmungsfähigkeiten direkt eingesetzt werden, und unter Einsatz von Metriken lassen sich auch quantitative Aussagen zum Vergleich von zwei Darstellungen ableiten. So eignet er sich, um simulierte Daten mit durch Kamerabilder aufgenommenen realen Phänomenen zu vergleichen und Datensätze zu vergleichen, bei denen die Beobachtungspunkte den gleichen Beobachtungsraum überdecken, insbesondere bei gleichen oder ähnlichen Gitterstrukturen.

Allerdings ist dieser Ansatz auch in seiner Anwendbarkeit beschränkt, weil er sich auf das Aussehen der resultierenden Bilder konzentriert und vom Inhalt der zugrunde liegenden Daten abstrahiert (vgl. Pagendarm u. Post 1995). Dadurch werden die Möglichkeiten zur vergleichenden Visualisierung - die sich insbesondere auch in den einzelnen Phasen der Visualisierungspipeline ergeben - stark eingeschränkt. So ergeben sich Grenzen seiner Anwendbarkeit, insbesondere wenn die zu vergleichenden Datenmengen sehr groß und/oder einem gewisses Maß an Unsicherheit bzw. Rauschen unterworfen sind (vgl. West u. Machiraju 1998).

Aufgrund seiner Einfachheit eignet sich der *image level*-Ansatz, um schnell einen Überblick über Regionen ähnlichen bzw. unähnlichen Verhaltens zu bekommen. Für die Spezifik komplexer Fragestellungen im Umfeld des Vergleiches von Klimadatensätzen stößt er jedoch schnell an seine Grenzen. So werden die Spezifika abweichender, beliebig im Beobachtungsraum liegender Gitter und der Vergleich darin existierender Muster nicht direkt unterstützt[10]. Auch eine Unterstützung des

[10]So bekommt der Nutzer bsw. keine Unterstützung dabei, ein Ozonloch auf der Südhalbkugel mit einem Ozonloch

Vergleiches mehrerer Merkmale gleichzeitig, wie es eine häufige Anforderung im Klimaumfeld ist, ist mit dem *image level*-Ansatz nur begrenzt möglich.

5.4.1.2 Datenbasierter Ansatz (data level)

Prinzip. Als zweiten Ansatz schlagen Pagendarm u. Post (1995) die Datenintegration auf dem ***data level*** vor (vgl. Abbildung 5.21). Hierbei werden die zu vergleichenden Datensätze in eine ge-

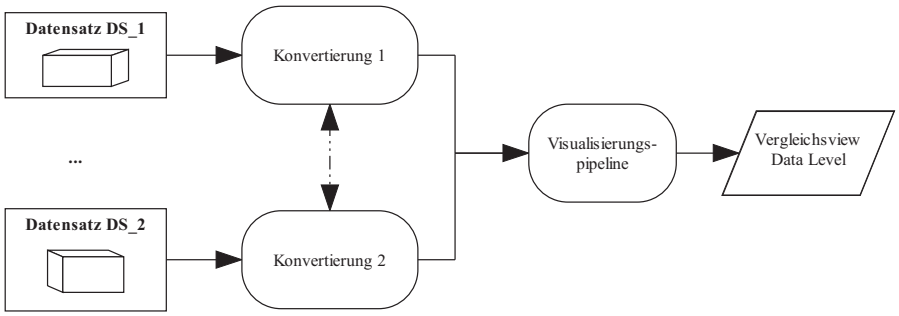


Abbildung 5.21: Vergleichende Visualisierung auf dem *data level*

meinsame Repräsentation überführt und dort durch die selbe Visualisierungspipeline transformiert.

Methoden. Grundsätzlich lassen sich - insbesondere für den Vergleich von Daten auf abweichenden Gittern - drei Ansätze unterscheiden:

1. Überführung der beiden Datensätze auf ein Gitter und anschließende
   - Berechnung einer Metrik auf den Daten und Darstellung der Ergebnisse mit Standardverfahren für uni-variate Daten oder
   - Anwendung von multi-variaten Darstellungen für die transformierten Datensätze,
2. getrennte Durchführung von Musterextraktionsverfahren und die gemeinsame Visualisierung der extrahierten Features und
3. Berechnung und Darstellung von Metriken zum Vergleich der Originalgitter und der darunter liegenden Daten.

Algorithmen zur *Gittertransformation* sind weitverbreitet. Insbesondere werden sie eingesetzt, um komplexe Gitterstrukturen in ein vereinfachtes Gitter (i. allg. regulär) zu überführen, da auf diesem vereinfachten Gitter eine breitere Palette von Visualisierungsverfahren anwendbar ist, und diese häufig eine wesentlich höhere Zeit- und Speichereffektivität aufwesen. Beispiel für die Überführung von gestreuten 2D-Daten auf ein regelmäßiges Gitter finden sich z.B. in Treinish (1994). Ein weiteres Beispiel für den Einsatz von Gittertransformationen findet sich im System Visual Earth (vgl. Ribarsky u. a. 2002b, a). Hier können in einer LOD-Hierarchie verschiedene Datenklassen miteinander kombiniert werden. Insbesondere werden - teilweise sich auch überschneidende - radiale 3D-Radar-Wetter-Daten auf eine regelmäßiges Gitter abgebildet, um sie besser mit Daten aus 2D-Satellitenbildern, 3D-Wettersimulationsdaten und Daten aus anderen Quellen auf regulären Gittern koppeln und visualisieren zu können. Unter Einsatz von Fehlermetriken kann der Nutzer bei der Visualisierung eine für ihn geeignetes LOD auswählen.

Methoden der *Featureextraktion und -visualisierung* ermöglichen es, Muster in den Daten komprimiert darzustellen. Ikonifizierte geometrische Primitive lassen Raum, um verschiedene Features

auf der Nordhalbkugel zu vergleichen, wobei insbesondere verschiedene Koordinatensystemtransformationen erforderlich sind.

gleichzeitig darzustellen (vgl. z.B. Pagendarm u. Walter 1995). Insbesondere können so auch Features unterschiedlicher Datensätze auf abweichenden Gittern in einem Bild dargestellt werden. Diese Vorgehensweise wird insbesondere für Strömungsdaten eingesetzt. Wichtige Strömung-Features können durch

- punkthafte Objekte (z.B. kritische Punkte)
- linienhafte Objekte (Stromlinien, Streichlinien, Reibungslinien, separierende Linien),
- Stromflächen (z.B. Schockwellen, flow-ribbons und separierende Ebenen),
- Flussräume (eng.: flow volumes)

repräsentiert werden (vgl. z.B. Silver u. Zabusky 1993; Pagendarm u. Walter 1995; Crowfis u. a. 2000; Schumann u. Müller 2000; Weinkauf u. a. 2004; Xue u. a. 2004), und lassen Raum für den Vergleich mehrerer Datensätze. In der Volumenvisualisierung lassen sich ebenfalls bestimmte Features extrahieren, bsw. über geeignete Transferfunktionen in Verbindung mit Segmentierungsverfahren im Direct-Volume-Rendering (vgl. z.B. Frühauf 1997; Hauser u. a. 2001; Viola u. a. 2004) oder über die Extraktion von Isoflächen. Aufgrund der platzsparenden Darstellung von Features sowie deren Abkopplung vom darunter liegenden Gittern eignen sich Methoden der Featurevisualisierung im besonderen für die vergleichende Visualisierung (auch abweichender Gitter).

Die *Berechnung und Darstellung von Metriken zum Vergleich der Originalgitter* eignet sich dafür, um zu untersuchen, inwieweit sich das gewählte Gitter auf die Genauigkeit der Simulation auswirkt. So stellen bsw. Trapp u. Pagendarm (1996) die Abweichungen von Gittergeometrien beim Design von Flugzeugen durch euklidische Differenzen der resultierenden Oberflächen und Abweichungen der resultierenden Krümmungen dar. Darüber hinaus können, um quantitative Aussagen über die Daten abzuleiten, die bereits für den *image level* diskutierten Metriken direkt auf den Datensätzen angewendet werden, wobei ggf. eine Koordinatenanpassung erforderlich sein kann.

Der Datenvergleich auf dem *data level* ist ein flexibler Ansatz zum Untersuchung von Datensätzen, da er sowohl Daten als auch enthaltene Muster in den Vergleichsprozess einbezieht, und dabei sowohl qualitative als auch quantitative Aussagen ermöglicht. Vorteil dieses Ansatzes ist, dass Ungenauigkeiten und Fehlerquellen - im Unterschied zum *image level*-Ansatz - während des eigentlichen Visualisierungsprozesses reduziert werden können (vgl. Pagendarm u. Post 1995).

Problem speziell bei *Gittertransformationen* ist, dass sie leichte Datenverfälschungen beinhalten und dadurch das Vertrauen der Anwender in die Visualisierung reduzieren können. Frühauf (1997) zählt einige Probleme auf, die bei der Gitterabbildung auf ein reguläres Gitter entstehen können:

- es entstehen Bereiche im Gitter, für die keine Daten vorliegen,
- durch die Interpolation bei der Gitterabbildung werden Extreme in Richtung des Mittelwertes verschoben und
- um die in hoher Auflösung modellierten Bereiche gut abzutasten, bedarf es eines feinen Gitters (gemäß der Abtastrate nach Nyquist) im gesamten Datenraum.

Der Vergleich auf dem *data level* bietet eine verbreiterte Grundlage, um zwei oder mehrere Datensätze auch in ihrer Struktur vergleichen zu können. Dies erweitert insbesondere die Möglichkeiten zur Einbeziehung der Gitter(-geometrie) in den Vergleichsprozess, was gerade auch im Klimaumfeld erforderlich ist. Die Grenze des *data level*-Ansatzes stellt seine Konzentration auf den Vergleich vor der Visualisierungspipeline dar. Die Daten müssen, folgt man der Definition von Pagendarm u. Walter (1995), vor der Ausführung der Visualisierungspipeline in ein einheitliches Bezugssystem überführt werden, bei dem insbesondere vorher bestehende Gitterabweichungen eliminiert werden müssen. Dies schließt streng genommen eine Fusion der Datenquellen erst im Mapping und/oder Rendering aus.

Ferner bezieht der Ansatz neuere Konzepte zur Datenvisualisierung nicht ausdrücklich ein. Hierzu gehört im besonderen die Definition und spezielle Darstellung von Regionen von Interesse bzw. die Anwendung von Fokus & Kontext-Mechanismen. Auch Konvertierungen der zugrunde liegenden Koordinatensysteme zum Vergleich auch räumlich und zeitlich unterschiedlich gelegener Muster, wie sie im Klimaumfeld z.B. beim Vergleich von Zirkulationen eingesetzt werden können, müssen hier konzeptuell untersetzt werden.

5.4.1.3 Ansätze zur Fusionierung im Mapping/Rendering

Die beiden vorgestellten Ansätze führen die vergleichende Visualisierung entweder über Fusionierung der Ausgangsdaten vor dem Mapping/Rendering (*data level*) oder mittels Fusionierung der Bilddaten nach Ausführung der Visualisierungspipeline (*image level*) durch. Gerade den visuellen Vergleich von Datensätzen mit abweichenden Gitter ist diese Vorgehensweise jedoch zum Teil nicht flexibel genug, denn sie bezieht eine Fusionierung direkt im Mapping bzw. Rendering nicht mit ein. Insbesondere ist eine unverfälschte gemeinsame Darstellung der abweichenden, zugrunde liegenden Gitter und die damit verbundene Darstellung deren Werteverteilung bei den beiden vorgestellten Ansätzen nur begrenzt möglich.

In der Literatur wird die Fusionierung im Mapping/Rendering nicht in den Kontext der vergleichenden Visualisierung gesetzt. Im folgenden soll dies getan werden und grundlegende Vorgehensweisen zur vergleichenden Visualisierung im Mapping und Rendering aufgelistet werden. Insbesondere für die Darstellung von Daten auf abweichenden Gittern können eine Vielzahl bekannter, nicht speziell für die vergleichende Visualisierung entworfene Darstellungstechniken genutzt werden bzw. haben Potential, hierfür angepasst zu werden. Für die vergleichende Visualisierung von Datensätzen auf abweichenden Gittern sind im besonderen solche Techniken relevant, deren Vorgehensweise nicht direkt auf der Gitterstruktur arbeitet[11] oder die es erlauben, visuelle Attribute zweier Merkmale auch bei abweichenden Gitterstrukturen im Mapping/Rendering zu fusionieren[12].

Grundsätzlich sollen drei Vorgehensweisen beim Einsatz von Standardtechniken bei der vergleichenden Visualisierung von Daten auf abweichenden Gittern unterschieden werden:

- Einsatz uni-variater Darstellungstechniken bei
  - separater Abbildung der Daten auf visuelle Attribute und deren anschließender Fusionierung im Mapping/Rendering (z.B. Farbe, Isolinien, Höhe oder Ikonen in 2D) oder
  - bildraumbasierte Abtastung des durch die abweichenden Gitter überspannten Raumes, aufsammeln der beteiligten Daten und deren Fusionierung im Rendering (z.B. mit bivariaten Farbskalen im Direct Volume Rendering),
- Einsatz von bi- oder multi-variaten Darstellungstechniken (für gleiche Gitter, vgl. z.B. Crowfis u. Max 1992; Pang u. Freeman 1996; Gelin 2002) bzw. von Visualisierungstechniken für Tensordaten (vgl. z.B. Hotz u. a. 2004; Helgeland u. a. 2004) und
- Einsatz von Darstellungstechniken für Unsicherheiten (vgl. z.B. Pang u. a. 1997; Djurcilov u. a. 2001; Griethe u. Schumann 2005) durch gemeinsame Kodierung von Absolutwerten und Abweichungen.

[11] Zum Beispiel sind Direct Volume Rendering Objektraumverfahren eher weniger geeignet, da sie bei der gleichzeitigen Darstellung mehrerer Merkmale (z.B. über 2D-Transferfunktionen) ein gleiches Gitter erfordern.

[12] Zum Beispiel kann in einem 2D Datensatz ein Merkmal des einen zu vergleichenden Datensatzes auf eine Farbfläche abgebildet werden, während das Merkmal des anderen Datensatzes (mit abweichendem Gitter) eine Deformation dieser Fläche bewirkt (Abbildung auf die Höhe). Bei dieser Art der Fusionierung im Mapping ist im allgemeinen keine vorverarbeitende Gitterüberführung notwendig, jedoch können aufgrund der verschiedenen Gitter Abtastprobleme auftreten, die z.B. eine Neutriangulierung der originalen Farbfläche erfordern können.

5.4.1.4 Diskussion

Pagendarm u. Post (1995) stellen fest, dass sich ein großer Anteil der Literatur zur vergleichenden Visualisierung in Arbeiten verschiedener Anwendungen findet (z.B. Strömungsmechanik), und nur wenige Arbeiten in der Visualisierungs-Community bekannt sind. Diese Aussage von 1995, insbesondere für 3D skalare und vektorielle Daten auf gleichen Gitterstrukturen ausgesprochen, hat auch heute noch Gültigkeit.

Grundsätzlich werden in der Literatur zwei Vorgehensweisen unterschieden: Datenfusion auf dem *image level* oder auf dem *data level*. Beim Vergleich auf dem *image level* werden Bilder über- oder nebeneinander gelegt sowie durch Bildmetriken miteinander verglichen. Im besonderen ist dieses Verfahren geeignet, um mit Kameras aufgenommene reale Phänomene mit simulierten Daten oder durch verschiedene Renderingverfahren generierte Darstellungen zu vergleichen. Nachteil dieses Vorgehens ist es, dass der kognitive Aufwand zum Vergleich nebeneinander liegender Bilder hoch ist, während Differenzbilder die ursprünglichen Werte nicht mit einbeziehen. Die Analyse von Daten auf abweichenden Gitter ist mit dem *image level*-Ansatz nur sehr begrenzt möglich.

Beim Vergleich auf dem *data level* werden die Daten in einem Konvertierungsschritt in einen gemeinsame Repräsentation überführt und anschließend gemeinsam in der Visualisierungspipeline transformiert. Dies bietet mehr Möglichkeiten des Datenvergleichs bereits im Beobachtungsraum, wie die Extraktion und Darstellung von Mustern, den Datenvergleich mithilfe von Metriken und/oder die Konvertierung in ein gemeinsames Gitter[13]. Damit kann der *data level*-Ansatz auch variierende Gitter wesentlich flexibler einbeziehen. Insbesondere sind Ansätze aus der Feature-Visualisierung geeignet, für Daten auf abweichenden Gittern eingesetzt zu werden, da die extrahierten Muster platzsparend wichtige Datencharakteristika kodieren und somit das Verdeckungsproblem reduzieren.

Während für 1D- oder 2D-Gitter bereits Verfahren zur vergleichenden Darstellung auch abweichender Gitter existieren, ist die Darstellung von abweichenden 3D-Gittern (ohne Gittertransformation) noch weitgehend offenes Forschungsfeld. Insbesondere die Fusionierung erst im Mapping und/oder Rendering bleibt zu untersuchen. Hier haben eine Vielzahl von (Standard-)Visualisierungstechniken, die bisher nicht für die vergleichende Visualisierung eingesetzt wurden, Potenz zur Anwendung in diesem Kontext. Beispiel hierfür sind Darstellungstechniken für Unsicherheiten, da diese im Kern ähnliche Zielstellungen verfolgen wie die vergleichende Visualisierung, nämlich Abweichungen im Kontext der Datenwerte zu untersuchen. Darüber hinaus müssen existierende Ansätze zur vergleichenden Visualisierung ausgebaut werden, den Nutzer auch bei beliebigen Koordinatentransformationen sowie bei der Definition von Bereichen von Interesse zu unterstützen, um beliebig im Beobachtungsraum liegende Phänomene fokussiert vergleichen zu können.

5.4.2 Problemanalyse

Im Umfeld der Analyse von Klimadaten ergeben sich eine Vielzahl von Szenarien, in denen eine vergleichende Visualisierung helfen kann, die Fragestellungen der Anwender - insbesondere bei der Modellevaluierung - zu beantworten. Aufgrund der heterogenen Datenquellen (z.B. gestreute 2D-Messstationen und simulierte Daten auf regulären 3D-Gittern) ergeben sich dabei besondere Herausforderungen. In der einschlägigen Visualisierungsliteratur fehlt bisher eine allgemeine Betrachtung der aufgrund solcher heterogenen Daten auftretenden Problemstellungen für eine vergleichende Visualisierung. In diesem Abschnitt sollen die auftretenden Probleme aus allgemeiner Sicht systematisiert werden. Diese Problemanalyse bildet die Basis, die speziellen Anforderungen bei der vergleichenden Visualisierung von Klimadaten einzuordnen und einzugrenzen (vgl. Abs. 5.4.3.1).

[13] Allerdings ist eine Abbildung der Gitter aufeinander oder auf ein reguläres Gitter mit Datenverfälschungen verbunden, hat jedoch den großen Vorteil einer breiten Palette von existierenden Techniken.

Des Weiteren werden bei der vergleichenden Visualisierung die Anforderungen im Vergleich zur Visualisierung einzelner Merkmale aufgrund des erhöhten Datenaufkommens verschärft. Insbesondere betrifft dies den begrenzten Darstellungsplatz und die begrenzten Wahrnehmungsfähigkeiten des Anwenders, aber auch erhöhte Zeit- und Speicheranforderungen. Die Herausforderungen sollen im folgenden systematisiert (vgl. Abs. 5.4.2.1) und anschließend die wichtigsten Zielstellungen bei der vergleichenden Visualisierung aufgrund des State-of-the-Art zusammengetragen werden (vgl. Abs. 5.4.2.2).

5.4.2.1 Herausforderungen und Probleme

Komplexität der Daten und der beteiligten Gitter. Bei steigender Komplexität der Gitter sowie bei anwachsender Größe der Datensätze steigt die Komplexität des visuellen Vergleiches. Speziell wird die Darstellung von Merkmalen, Gitterpunkten, Gitterlinien und/oder Gitterzellen erschwert. Entsprechend bedarf es neuer Strategien, um die verfügbare Darstellungsfläche für die erhöhte Menge darzustellender Daten effektiv zu nutzen und dabei spezielle Interaktionstechniken einzubeziehen.

Im folgenden sollen wichtige Einflussfaktoren zur Charakterisierung der Komplexität der vergleichenden Visualisierung aufgelistet und näher erläutert werden. So können beim Vergleich mehrerer Datensätze alle Elemente des gitterbeschreibenden Tupels $DG = (V, L, C, M, I)$ (vgl. S. 6) voneinander abweichen. Weiterhin müssen aber auch allgemeine Eigenschaften des Beobachtungsraumes und der Merkmale mit einbezogen werden.

- **Die Anzahl der zu vergleichenden Datensätze**: Üblicherweise beschränkt sich die vergleichende Visualisierung auf zwei Datensätze. Grundsätzlich kann es sich jedoch auch um mehr als zwei handeln, wobei typische (räumliche) Darstellungstechniken in einem Bild jedoch schnell an ihre Grenzen stoßen.
- **Die Dimensionalität des Beobachtungsraumes, in dem sich die Gitterpunkte V befinden**: Typischerweise liegen die Gitterpunkte in ein-, zwei oder drei Dimensionen vor. Darüber hinaus können der Beobachtungsraum aber auch mehr als drei unabhängige Variablen aufweisen (z.B. bei Multi-Run-Experimenten). Bereits bei 3D-Daten ist der Vergleich im räumlichen Kontext aufgrund von Verdeckungen erschwert. Für höherdimensionale Räume verschärft sich dieses Problem noch.
- **Art des Beobachtungsraumes:** Typischerweise liegen die Daten im euklidischen Raum vor. Darüber hinaus kann es sich aber auch um andere Arten von Räumen handeln. Beispiele hierfür sind der Frenquenzraum oder gekrümmte Räume bei relativistischen Simulationen. Je nach der Art des Raumes und des die Daten repräsentierenden Phänomens können verschiedene **Interpolationsverfahren** eingesetzt werden. Insbesondere werden diese durch den **Wirkungskreis der Gitterpunkte** bestimmt, welcher angibt, wie weit die Datenwerte eines Gitterpunktes Geltung auf den umliegenden Beobachtungsraum haben. Man unterscheidet punktuelle[14], lokale[15] und globale[16] Geltungsbereiche. Der Wirkungskreis sowie die Interpolationsverfahren schränken die in Frage kommenden Visualisierungsmethoden auch in der vergleichenden Visualisierung stark ein.
- **Der Gittertyp**: Der Gittertyp wird durch die Gittertopologie aufgrund der Gitterpunkte V, der Gitterlinien L und der Gitterzellen C sowie der Lage der Gitterpunkte zueinander bestimmt. So wird zwischen regulären, blockstrukturierten, kurvilinearen, unstrukturierten und hybriden Gittern unterschieden. Insbesondere für reguläre 2D- und 3D-Gitter existieren eine

[14] Die Wirkung eines Datenwertes gilt ausschließlich für den Gitterpunkt.
[15] Die Wirkung eines Datenwertes beschränkt sich auf eine Umgebung um den Gitterpunkt
[16] Ein Datenwert eines Gitterpunktes wirkt auf den gesamten Beobachtungsraum.

Vielzahl geeigneter Visualisierungstechniken, die z.T. auch für die vergleichende Visualisierung eingesetzt werden können. Deswegen ist auch bei der vergleichenden Visualisierung ein häufiger Ansatz, beteiligte Gitter auf ein reguläres Gitter zu überführen, und den Vergleich auf diesem Gitter visuell oder über Nutzung einer Metrik durchzuführen. Die dabei auftretenden Fehler erfordern jedoch neue Vorgehensweisen.

- **Die Gitterauflösung**: Bei Messung oder Simulation können je nach Erfordernis verschiedene Gitterauflösungsstufen verwendet werden, um die den Daten zugrunde liegenden Phänomene geeignet wiederzugeben. So kann die Gitterauflösung in Abhängigkeit von Modell und/oder Simulationsverfahren stark variieren. Spezielle Techniken, die die grundlegende Struktur des Gitters ausnutzen, und dabei verschiedene Gitterauflösungen vergleichen, sind erforderlich, um solche Variationen zu untersuchen. Insbesondere gehört hierzu auch die Dimension Zeit. So können sich abweichende Datensätze auch in der Zeitabtastung unterscheiden. Dabei kann (bei Simulationen in Abhängigkeit der Schrittweitensteuerung), zwischen uniformen und variierenden Schrittweiten unterschieden werden. Entsprechend ergeben sich bei ungleichen Zeitabtastungen erhöhte Anforderungen an die vergleichende Visualisierung, insbesondere falls eine Interpolation zwischen den Zeitschritten nicht erlaubt ist.
- **Anzahl zu vergleichender Merkmale** M: Je nach der Anzahl an Merkmalen pro Beobachtungspunkt wird in uni-variate (ein Merkmal), bi-variate (zwei Merkmale) und multi-variate Darstellungen (mehrere Merkmale) unterschieden. Bei einer vergleichenden Visualisierung kann es mit zunehmender Anzahl an zu vergleichenden Merkmalen schnell zu Überfrachtungen der Darstellung führen. Techniken zum Vergleich multi-variater Datenmengen wurden bisher kaum untersucht.
- **Der Datentyp der Merkmale** M: Je nach Art der beteiligten Datensätze können skalare, vektorielle oder tensorielle Merkmale verglichen werden. Sollen mehrere Merkmale gleichzeitig verglichen werden, kann auch eine gemischte Darstellung mehrerer Datentypen erforderlich sein. Insbesondere die vergleichende Darstellung eines skalaren oder eines vektoriellen Merkmals bzw. eine Kombination dieser beiden wurde bisher in der Literatur untersucht.
- **Art und Umfang des Wertebereiches der Merkmale** M: Vor allem bezieht dies den Skalentyp[17] der Merkmale sowie die Kardinalität der Wertebereiche ein. Während die Skalentypen der beteiligten Merkmale die einsetzbaren Techniken einschränken, wird insbesondere bei steigender Kardinalität des Wertebereiches auch der visuelle Vergleich komplexer.

Tabelle 5.4 fasst wichtige Datencharakteristika, die sich direkt auf die Komplexität des visuellen Vergleichs auswirken, zusammen.

Dabei sind die Ausprägungen der Datencharakteristika nach der Komplexität des Problems aufsteigend von links nach rechts sortiert. So wird eine Darstellung mehrerer Datensätze mit mehreren vektoriellen Merkmalen auf unstrukturierten Gittern eine höhere Komplexität als eine Darstellung zweier skalarer Größen auf abweichenden, aber strukturierten Gittern haben.

Darüber hinaus ergeben sich Herausforderungen, wenn nicht die gesamten Datensätze, sondern spezielle Phänomene wie z.B. Zyklonen oder Ozonlöcher verglichen werden sollen. So können diese in Koordinatensystemen mit abweichenden Koordinatenursprung und/oder Achsenausrichtung vorliegen. Dann müssen dem Nutzer geeignete Interaktionswerkzeuge an die Hand gegeben werden, um eine geeignete Überlagerung der Räume bzw. Phänomene zu unterstützen. Ein spezieller Aspekt bei einer solchen Überlagerung ist, dass sich die zu vergleichenden Gitter nur teilweise überdecken. Dies erschwert insbesondere einen Vergleich auf dem *image level*, da die Zuordnung von Bildpunkten in den einzelnen Bildern erschwert wird.

Verdeckungen und Mehrdeutigkeiten. Mit steigender Komplexität der zu vergleichenden Daten - insbesondere bei 3D-Datensätzen - kann es zu Überdeckungen und Mehrdeutigkeiten in den

[17] Grundsätzlich werden nominale, ordinale, diskrete und kontinuierliche Skalentypen unterschieden.

Anzahl der Datensätze	2	N			
Anzahl der Dimensionen	1D	2D	3D	nD	
Art des Raumes	Euklidischer R.	Frequenzraum	...		
Wirkungskreis	punktuell	lokal	global		
Gittertyp	strukturiert	hybrid	unstrukturiert		
Gitterauflösung	gering	hybrid	hoch		
Zeitabhängigkeit	nein	ja			
Zeitabtastung	uniform	variierend			
Anzahl der Merkmale	univariat	bivariat	multivariat		
Skalentyp der Merkmale	Skalar	Vektor	Tensor		

Tabelle 5.4: Datencharakteristika mit Einfluss auf den Schwierigkeitsgrad der vergleichenden Visualisierung

Daten kommen. Techniken zur vergleichenden Visualisierung müssen Mehrdeutigkeiten vermeiden sowie Verdeckungen in Bereichen von Interesse minimieren. Insbesondere müssen dabei

- das Problem des begrenzten Darstellungsplatzes und
- das Verdeckungsproblem bei der Darstellung von
  - beteiligten Merkmalen,
  - Gitterpunkten und
  - Gitterlinien bzw. Gitterzellen

  einzeln und in Kombination

gelöst werden.

Weitere Herausforderungen und Probleme. Über die genannten Probleme hinaus ergeben sich bei der vergleichenden Visualisierung erhöhte Anforderungen an *Speicherbedarf* und *Rechenzeit*, da es sich hier um ein erhöhtes Datenaufkommen handelt. Um gerade bei 3D-Darstellungen die vergleichende Visualisierung echtzeitfähig zu halten, müssen vergleichende Verfahren hardwarenah umgesetzt werden.

Da eine typische Vorgehensweise bei der vergleichenden Visualisierung die Überführung der Daten auf ein gemeinsames Gitter ist, ergeben sich weiterhin Probleme durch die *Interpolation* der Daten. Um eine Unterabtastung zu vermeiden, müssen höheraufgelöste Gitter als die Originalgitter verwendet und auftretende Unsicherheiten in die Darstellung mit einbezogen werden. Dies führt zu erhöhten Anforderungen an die vergleichende Visualisierung.

5.4.2.2 Zielstellungen bei der vergleichenden Visualisierung

In der Literatur zur vergleichenden Visualisierung lassen sich zwei grundlegende Zielstellungen identifizieren: *Vergleichen bzw. Unterscheiden der Merkmale* und *Vergleichen bzw. Unterscheiden der Gitter*.

Vergleichen bzw. Unterscheiden der Merkmale M. Hierbei lassen sich die folgenden, aufeinander aufbauenden Zielstellungen identifizieren:

1. Lokalisieren von räumlichen und zeitlichen Regionen gleichen (bzw. ähnlichem) und/oder abweichendem Werten und Werteverteilungen (vgl. z.B. Pagendarm u. Post 1995; West u. Machiraju 1998; Deines u. a. 2006; Woodring u. Shen 2006),
2. Quantifizieren der Stärke der Abweichungen (vgl. z.B. West u. Machiraju 1998; Deines u. a.

2006; Woodring u. Shen 2006),

3. Vergleich und Unterscheidung von räumlich/zeitlichen Mustern in den Daten (vgl. z.B. Pagendarm u. Walter 1995; Pagendarm u. Post 1995; Woodring u. Shen 2006),
4. Gruppierung/Klassifikation zum Einsatz beim Vergleich (vgl. z.B. Zhou u. a. 2002a).

Vergleichen bzw. Unterscheiden der Gitter:

5. Lokalisation und Vergleich der Koordinaten der Gitterpunkte V (vgl. z.B. Trapp u. Pagendarm 1996),
6. Vergleich und Unterscheidung der Gittertopologie und der Gittergeometrie (vgl. z.B. Pang u. Freeman 1996; Trapp u. Pagendarm 1996).

Kombination von Merkmals- und Gittervergleich. Häufig sind die beiden Zielstellungen miteinander gekoppelt, so dass gleichzeitig Gitter und Werte miteinander verglichen werden sollen. So ist die Beantwortung der Fragestellung, ob eine lokale Abweichung in den Werteverteilungen auf der Abweichung von Gitterauflösungen, Gitterstrukturen- oder unterschiedlichen Interpolationsverfahren beruht, häufig von Bedeutung.

Weitere Aufgaben und Zielstellungen. Neben diesen sind auch allgemeine Aufgaben wie Überblick und Details-on-Demand und Zielstellungen wie Identifizieren einzelner Werte – die im Umfeld nicht-vergleichender Visualisierung auftreten – weiterhin von Belang (vgl. Abs. 7.2).

5.4.3 Entwurf eines neuen Ansatzes zur vergleichenden Visualisierung

Im folgenden soll nun aufbauend auf den allgemeinen Betrachtungen der vorangegangenen Abschnitte ein neuer Ansatz zur vergleichenden Visualisierung unter der Einbeziehung abweichender Gitter vorgestellt werden, der insbesondere dafür ausgerichtet ist, den Datenvergleich von Klimadatensätzen zu unterstützen.

5.4.3.1 Problemeingrenzung

Die im vorangegangenen Abschnitt vorgestellte allgemeine Problemstellung bei der vergleichenden Visualisierung ist sehr komplex. Ein allgemeingültiger Lösungsansatz hierfür ist schwer zu entwickeln und würde über die Zielstellung der vorliegenden Arbeit hinausgehen. Darum soll im folgenden eine Eingrenzung auf zwei im Klimaumfeld typische Szenarien vorgenommen werden: Vergleich zweier regulärer oder blockstrukturierter 2D- oder 3D-Gitter (aus Simulationen) untereinander oder mit gestreuten Messdaten.

Entsprechend soll für den anschließenden Entwurf die allgemeine Problemstellung aus Sicht der Daten aus Tabelle 5.4 auf diese Spezialfälle eingeschränkt werden. Ferner sollen die beiden Gitter eine beliebige Lage im euklidischen Raum haben, und dabei lediglich skalare Merkmale mit lokalem bzw. globalem Wirkungskreis betrachtet werden.

Insbesondere sind hierbei drei Fälle für die Anwendung von Interesse: der visuelle Vergleich

- zweier Simulationen auf gleich gelegenen, aber unterschiedlich aufgelösten Gittern,
- eines simulierten Datensatzes auf einem regulärem oder blockstrukturiertem Gitter mit gestreuten Messpunkten sowie
- beliebig im Beobachtungsraum gelegener Muster, z.B. klimatische Phänomene wie zyklonale Strömungen.

Im folgenden soll nun ein neuer, flexibler und interaktiver Ansatz entworfen werden, der auf die beschriebenen Anforderungen der Anwendung zugeschnitten ist, darüber hinaus aber auch für die vergleichende Visualisierung in anderen Anwendungsgebieten eingesetzt werden kann.

5.4.3.2 Ansatz zur vergleichenden Visualisierung

Für die vergleichende Visualisierung von Klimadaten, im besonderen unter Einbezug von Daten auch mit abweichenden Gitterstrukturen, ergeben sich spezielle Anforderungen. Die isolierte Durchführung der Standardansätze zum Vergleich auf dem *image* und *data level* reichen hierzu nicht aus (vgl. Abs. 5.4.1). Deswegen wird im folgenden ein neuer Ansatz vorgestellt, der die beiden genannten Vorgehensweisen in einem einheitlichen Schema vereint und zusätzlich noch um weitere Verfahren erweitert. Als neuer Weg wird hier eine Fusionierung auf dem *Mapping/Rendering level* eingeführt.

So ergibt sich ein mächtiges Werkzeug, dessen Komponenten beliebig mit Funktionalität unterlegt werden können. Dabei wird die Visualisierungspipeline nach dos Santos u. Brodlie (2004) (Datenanalyse, Filterung, Mapping und Rendering) für die vergleichende Visualisierung erweitert. Abbildung 5.22 zeigt das Grundschema des neuen Ansatzes.

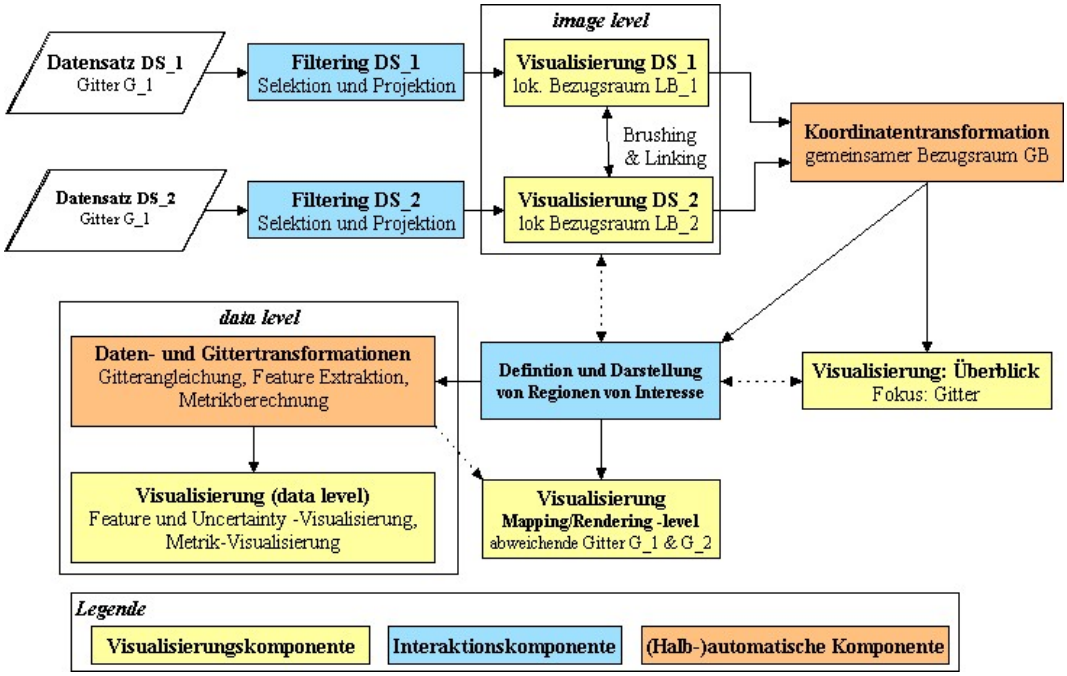


Abbildung 5.22: Überblick über den neuen Ansatz zur vergleichenden Visualisierung

In einem ersten Schritt kann der Nutzer die beiden Datensätze interaktiv *filtern*, was insbesondere die Selektion von Merkmalen und die Projektion des Beobachtungsraumes auf einen geeigneten Unterraum einschließt. Dann können die beiden Datensätze separat dargestellt und auf dem *image level* miteinander vergleichen werden. Wichtiger neuer Schritt ist die sich daran anschließende *Koordinatentransformation*, welche die beiden Gitter in einen gemeinsamen Bezugsraum überführt. Basierend auf dieser Transformation kann dann eine *Überblicksdarstellung* der Datensätze im neuen Bezugsraum erfolgen. Hierbei liegt der Fokus darauf, die Gitter und ihre gegenseitige Lage zu untersuchen. Basierend auf dieser *Überblicksdarstellung* oder den Einzelndarstellungen der beiden Datensätze folgt dann eine interaktive Spezifikation eine *Regionen von Interesse*, die es erlauben, auf die aktuell relevanten Bereiche des Beobachtungsraumes zu fokussieren. Basierend auf den derart eingeschränkten Datensätzen können dann zum einen *Daten- und Gittertransformationen* mit anschließender *Visualisierung der transformierten Daten* erfolgen. Dies entspricht dem Vergleich auf dem *data level*. Zum anderen können aber auch die untransformierten Daten der Ausgangsgitter erst im *Mapping/Rendering-Schritt* fusioniert werden, was neue Möglichkeiten zur Darstellung der Gitter ermöglicht und bei der vergleichenden Visualisierung ein neuer Ansatz ist.

Im folgenden sollen nun diese einzelnen Schritte untersetzt werden. Dabei sollen sowohl Standardherangehensweisen einbezogen und kurz beschrieben und neue Konzepte ausführlicher diskutiert werden.

Filtering (DS_1 und DS_2). Das interaktive Filtering dient der Reduktion der Datenmenge auf ein handhabbares Maß sowie der nutzergesteuerten Zielausrichtung des Analyseprozesses. Vor einer ersten vergleichenden Datenvisualisierung steht hierbei die Selektion der zu vergleichenden Merkmale sowie die Einschränkung auf gewisse Schnitte im Beobachtungsraum im Vordergrund. So ist die Selektion von geeigneten Merkmalen stark anwendungsabhängig, und kann ggf. durch automatische Maße bzw. Metriken unterstützt werden, um z.B. Merkmale mit besondereren Abweichungscharakteristika zu bestimmen und in eine Vorauswahl einzubeziehen.

Visualisierung (DS_1 und DS_2) auf dem *image level*. Hierbei können Standardmethoden für die Darstellung einzelner Daten eingesetzt und miteinander gekoppelt werden (vgl. Abs. 5.4.1.1). Ziel ist es, einen schnellen Eindruck über die beteiligten Datensätze zu erhalten und ggf. in ihren vorliegende Muster bestimmten Regionen zu ordnen zu können (vgl. auch Abb. 5.28).

Koordinatentransformation. Sollen nun diese einzeln dargestellten Datensätze zusammengefügt werden, müssen deren Koordinatensysteme je nach Zielstellung geeignet transformiert werden. Insbesondere sind die drei folgenden Transformationsarten relevant:

1. *Transformation der Gitter in ein gemeinsames Weltkoordinatensystem:* Häufig sind verschiedene Datensätze in abweichenden Koordinatensystemen definiert (z.B. Länge von 0 bis 360° vs. von -180° bis +180°). Zum Vergleich müssen diese Koordinatensysteme angeglichen werden. Im allgemeinen kann dies durch eine Transformationsmatrix beschrieben werden und metadatenbasiert automatisch erfolgen. Liegen entsprechende Informationen nicht vor, muss der Nutzer bei der Angleichung der Koordinatensysteme unterstützt werden (z.B. bei Translationen und Skalierungen).

2. *Koordinatentransformationen zum Vergleich örtlich/zeitlich auseinander liegender Phänomene:* Um beliebige Phänomene in ihren räumlichen Eigenschaften miteinander vergleichen zu können, müssen diese übereinander gelegt werden (alle affinen Abbildungen sind hier einsetzbar). Hierbei können auch Methoden der Bildverarbeitung eingesetzt werden, um den Nutzer bei der Findung einer geeigneten Transformation zu unterstützen.

3. *Transformationen der Daten in andere Räume:* Hierunter fallen beliebige, insbesondere nichtlineare Transformationen der Daten. Ein Beispiel hierfür ist die Abbildung in den Fourierraum, um die Daten nicht in ihren örtlichen, sondern in ihren spektralen Eigenschaften zu vergleichen.

Um den Nutzer hierbei zu unterstützen bedarf es geeigneter automatischer Methoden und Interaktionstechniken.

Visualisierung Überblick. Insbesondere ist das Ziel der Überblicksdarstellung, die Lage der Gitter zueinander im gemeinsamen Beobachtungsraum zu untersuchen und zu vergleichen. Dies umfasst die Lage der Gitterpunkte, -linien und -zellen. Der Anwender bekommt einen Überblick über abweichende Gittertopologie und -geometrie. Darüber hinaus können hier auch Merkmalswerte übersichtsartig dargestellt werden, was jedoch im besonderen bei 3D-Gittern aufgrund der Datendichte und damit einhergehender Verdeckungen nur einen eingeschränkten Wertevergleich erlaubt. Im folgenden sollen nun - bisher in der Literatur nicht umfassend untersucht - grundlegende Vorgehensweisen bei der gemeinsamen Darstellung der abweichender Gitter diskutiert werden.

In Abhängigkeit von Lage der Gitter, Blickpunkt und spezifiziertem Bereich von Interesse können bei beliebig ausgerichteten Gittern allgemein drei Verdeckungsbereiche unterschieden werden: (1) Überlagerungsbereich, (2) Übergangsbereich und (3) verdeckungsfreier Bereich. Abbildung 5.23 illustriert diese Bereiche anhand zweier regelmäßiger 2D- und zweier regelmäßiger 3D-Datengitter.

Wenn die zugrunde liegenden Gitterstrukturen explizit dargestellt werden sollen, um Rückschlüsse

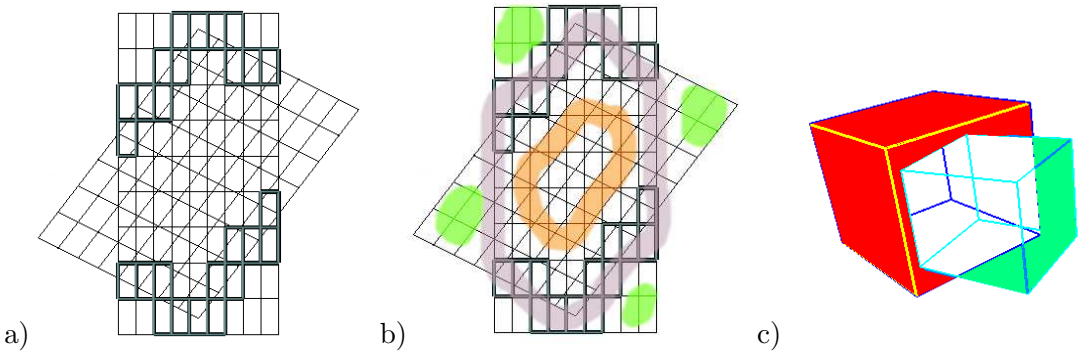


Abbildung 5.23: Illustration von Gitterüberlagerungen a) Überlagerung zweier regelmäßiger 2D-Gitter, Übergangszellen von Gitter 1 hervorgehoben b) Illustration der Verdeckungsbereiche (grün: verdeckungsfreie Bereiche; braun: Übergangsbereich; orange: Überlagerungsbereich) c) Illustration der Verdeckungsbereiche (ohne Übergangsbereich) bei zwei regelmäßigen 3D-Gittern (rot: verdeckungsfreier Bereich des Gitter 1; grün: verdeckungsfreier Bereich des Gitter 2; weiß: Überlagerungsbereich Gitter 1 und 2)

über die daraus resultierende Datenverteilung zu erhalten, müssen die folgenden Probleme adressiert werden:

- *eine eindeutige Zuordnung der zur Kodierung der Gitterstrukturen verwendeten Primitive zu ihrem Gitter*, z.B. durch Farbkodierung (vgl. z.B. Abb. 5.29)
- *eine geeignete Behandlung im Falle der Überdeckung der Primitive*: Die gegenseitige Überdeckung der Primitive variiert nach Lage und Struktur der zugrunde liegenden Gitter sowie nach dem gewählten Blickpunkt (in 3D). Hierbei muss der Nutzer dabei unterstützt werden, Darstellungen unter Reduktion von Verdeckungen zu generieren (z.B. durch eine Vorauswahl geeigneter Blickpunkte unter Maximierung sichtbarer Gitterpunkte) und abweichende Gitterstrukturen gesondert hervorzuheben.
- *ein der Wichtigkeit angemessener Platzverbrauch*: Problem hierbei ist es, geeignete visuelle Kodierungen zur Akzentuierung und Deakzentuierung der Gitterstrukturen in Abhängigkeit von deren Relevanz einzusetzen (z.B. Punkt- und Strichstilmittel wie Größe, Stärke, Transparenz oder Weglassen in bestimmten Regionen; vgl. auch Abb. 5.29)

Definition und Darstellung von Regionen von Interesse. Die *interaktive Festlegung* von Bereichen von Interesse kann sowohl in den Darstellungen auf dem *image level* eingebunden als auch in der Überblicksansicht erfolgen. Für eine leicht handhabbare Interaktion ist z.B. eine *Definition* einer rechteckigen Region von Interesse in mehreren 2D-Sichten auf dem *image level* geeignet (vgl. auch Abb. 5.27 links). Die resultierende Region kann dann in der *Überblicksdarstellung* im 3D-Raum dargestellt werden (vgl. Abb. 5.29c).

Darüber hinaus kann der Anwender auch durch automatische Verfahren bei der Findung interessanter Regionen - im Sinne einer Voreinstellung - unterstützt werden (z.B. über vergleichende Metriken, vgl. z.B. West u. Machiraju (1998)).

Daten- und Gittertransformation. Diese dienen vor allem dazu, die Daten im Sinne des *data level* in eine gemeinsame Repräsentation zu überführen. Standardvorgehensweisen sind *Gitterangleichung*, *Metrikberechnung* und *Featureextraktion* (vgl. Abs. 5.4.1.2).

Die *Gitterangleichung* durch Transformation der beteiligten Gitter auf ein gemeinsames Gitter ermöglicht, auch Standardverfahren einzusetzen bzw. diese unter leichten Modifikationen an die Bedürfnisse der vergleichenden Visualisierung anzupassen. Für die vergleichende Visualisierung im Klimaumfeld sind insbesondere die zwei folgenden Vorgehensweisen typisch:

- Abbildung der Gitter auf ein regelmäßiges Gitter (vgl. hierzu Abs. 5.4.1.4).
- Abbildung auf gestreute Messstationen (vgl. z.B. Böhm 1999; Böhm u. a. 2004).

Über diese Standardverfahren hinaus können bei der Gittertransformation auch weitere Techniken eingesetzt werden, um die Möglichkeiten bei der vergleichenden Visualisierung zu erweitern:

- Einsatz automatischer Verfahren, um die nachfolgenden Schritte Mapping und Rendering zu unterstützen: Ausnutzung der Gitter- und/oder Merkmalscharakteristika zur Hervorhebung oder Abschwächung (z.B. die Einteilung des Raumes in verschiedene Verdeckungsbereiche (vgl. S. 85) oder die Bestimmung geeigneter Kamerapositionen unter Minimierung von Verdeckungen)
- Einsatz von Level-of-Detail-Techniken: hierarchische Organisation der zu vergleichenden Daten zur Darstellung der Daten in verschiedenen Detailgraden (Details-on-Demand); auch zur Bewältigung der Datenquantität
- Einsatz von Fokus & Kontext-Techniken: veränderte Darstellung von Gitterpunkten, -linien und/oder -zellen (insbesondere von deren Position, aber auch deren Renderingstil), um *Darstellungsplatz für bestimmte Gitterbereiche von Interesse zu gewinnen*, während der Kontext komprimiert dargestellt wird (z.B. Explosionszeichnungen aus Strothotte u. Schlechtweg (2002); Mullerworth (2004); Bruckner u. Gröller (2006), Linsentechniken aus Bier u. a. (1993); Ropinski u. Hinrichs (2004); Griethe u. a. (2005) oder angepasste Renderingstile aus Strothotte u. Schlechtweg (2002); Ropinski u. Hinrichs (2004); Islam u. a. (2004); Singh u. Silver (2004); Viola u. a. (2004); Chen (2006))

Visualisierung auf dem *data level*. Hierbei werden die Ergebnisse der oben beschriebenen Transformationen visualisiert. Insbesondere schließt dies extrahierte räumliche Muster (vgl. Abb. 5.30) und auf ein Gitter transformierte Datensätze ein (vgl. Abs. 5.4.1.2), wobei z.B. bi-variate Darstellungstechniken (Kodierung der Absolutwerte beider Datensätze; in Abb. 5.31a) oder Darstellungstechniken für Unsicherheiten (zur Kodierung der Absolutwerte eines Datensatzes in Kombination mit den Differenzinformationen, z.B. Abb. 5.31c) eingesetzt werden können.

Visualisierung auf dem *Mapping/Rendering level*. Beim visuellen Vergleich auf dem *data level* müssen die Daten vor dem Mapping/Rendering in eine gemeinsame Repräsentation überführt werden, wodurch es zu Datenverfälschungen kommen kann. Außerdem werden vielfältige Möglichkeiten zur Fusionierung erst im Mapping[18] bzw. zu einem getrennten Mapping und der Fusionierung im Rendering[19] nicht mit einbezogen. Gerade bei der Visualisierung abweichender Gitter, wo aufgrund der verschiedenen Gitterstrukturen verschieden kodiert werden kann, ermöglichen diese Arten der Fusionierung neue, expressive Darstellungen. Im folgenden sollen allgemeine Lösungsstrategien zur Fusionierung von Merkmalen auf dem *Mapping/Rendering level* diskutiert werden.

Grundlegend ist zu untersuchen, welche Arten von visuellen Attributen und deren Parametrisierungen in der vergleichenden Visualisierung von Merkmalen auf abweichenden Gitter eingesetzt werden können, und dabei Verdeckungen und Mehrdeutigkeiten zu reduzieren. Dazu sollen die folgenden zwei Strategien genauer untersucht werden: zum einen kann das Wissen ausgenutzt werden, dass es sich bei den Daten um vergleichbare (klimatische) Phänomene mit oft nur geringen Abweichungen handelt, was ermöglicht, platzsparende visuelle Kodierungen unter *Reduzierung von Verdeckungen* zu generieren. Zum anderen können spezielle *Informationen über Lage und Typ der beteiligten Gitter* bei der visuellen Abbildung ausgenutzt werden.

Zur *Reduktion von Verdeckung* – die insbesondere bei 3D-Gittern ein großes Problem darstellt – lassen sich die folgenden Strategien identifizieren:

[18] Beim **Mapping** können die zu vergleichenden Merkmale auf abweichende visuelle Attribute wie Farbe, Form oder Größe abgebildet werden, was einen visuellen Vergleich ermöglicht.

[19] Werden die zu vergleichenden Merkmale beim Mapping auf gleiche visuelle Attribute abgebildet, und kann der Vergleich durch verschiedene Renderingstile- und -parameter erfolgen.

- unterschiedliche Behandlung von Bereichen mit großer Abweichung[20] und von Bereichen mit geringer Abweichung[21],
- platzsparenden Repräsentationen wichtiger Eigenschaften durch Einsatz kompakter Primitive (analog zum *data level*) wie punktuelle Objekte (z.B. Glyphen an den Gitterpunkten, kritische Punkte), linienhafte Objekte (z.B. Isolinien, Stromlinien) und flächenhafte Objekte (z.B. Isoflächen, Schnittebenen oder separierende Flächen).

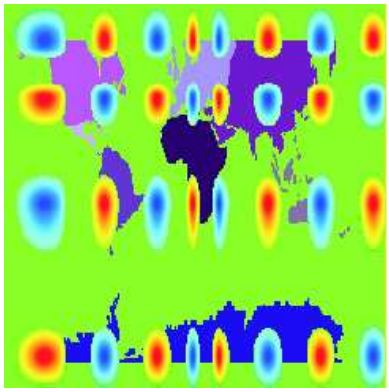


Abbildung 5.24: Farbkodierte Visualisierung von Extrembereichen eines Gitters über einem zweiten Gitter mit Land-See-Maske

Durch die Fokussierung auf wichtige Datencharakteristika (z.B. die Abweichung oder die Absolutwerte) können gewisse Raumbereiche auf diese Art hervorgehoben oder abgeschwächt werden, um dadurch die Verdeckung zu reduzieren und/oder den Nutzer auf gewisse Aspekte in den beteiligten Datensätzen lenken. So können bsw. Lücken im Definitionsbereich der Daten (z.B. Land- und Seemasken) ausgenutzt oder die Daten nur in Bereichen von Werteextremen dargestellt werden (vgl. Abb. 5.24).

Zu den *Informationen über Lage und Typ der beteiligten Gitter* gehören die drei Bereiche Überlagerungsbereich, Übergangsbereich und überlagerungsfreier Bereich. Die Information über diese Bereiche kann ausgenutzt werden, um im Überlagerungsbereich visuelle Attribute zu verwenden, die Verdeckungen reduzieren, z.B. durch Einsatz von punkt- oder linienhaften Objekten oder transparenten Flächen im Überlagerungsbereich und flächenhafter Merkmalskodierung in den überlagerungsfreien Bereichen (vgl. Abb. 5.25).

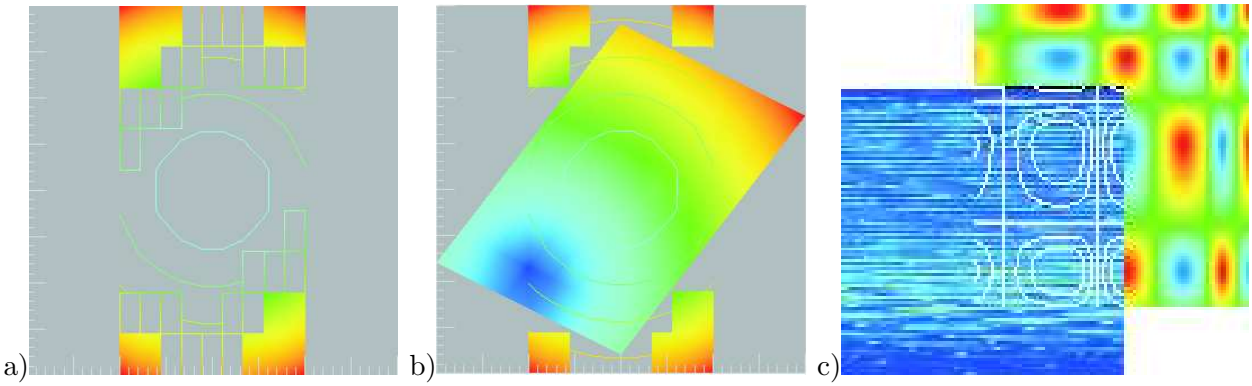


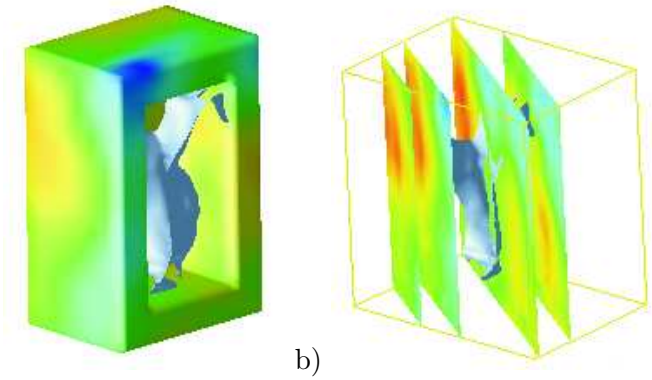
Abbildung 5.25: Illustration verschiedener Renderstile für zwei sich überlagernde 2D-Gitter: a) Isoliniendarstellung im Überlagerungsbereich, farbkodierte Gitterkanten im Übergangsbereich und flächenhaft interpolierte Farbdarstellung im verdeckungsfreien Bereich (lediglich für Gitter 1); b) Gitter 2 vollständig farbkodiert dargestellt, Gitter 1 ist abgeschwächt dargestellt (Isoliniendarstellung im Überlagerungsbereich und Übergangsbereich und flächenhaft interpolierte Farbdarstellung im verdeckungsfreien Bereich; c) Alternative Einfärbung der Isolinien

Bei Verwendung von Farbe kann sich in sehr ähnlichen Bereichen des Überlagerungs- sowie des Übergangsbereiches zu Kontrastproblemen der verschiedener farbkodierten 2D-Objekte kommen (vgl. Abb. 5.25b) . Alternative ist es, die Datensätze unter Reduktion des Kontrastproblems im Übergangsbereich auch farblich abweichend zu kodiert werden, was jedoch den direkten Vergleich der Werte der beiden Datensätze erschwert (vgl. Abb. 5.25c).

[20] hervorgehobene, auch komplexere Kodierungen in diesen Bereichen,

[21] durch Reduktion der Komplexität der Visualisierung in diesen Bereichen, z.B durch Darstellung der Merkmale nur eines Gitters oder durch abgeschwächte Darstellung mittels Transparenz

Analog können auch für 3D-Darstellungen verschiedene Mappings unter Ausnutzung von Gitterbereichen oder platzsparenden Primitiven verwendet werden (vgl. Abb. 5.26).



a) b)

Abbildung 5.26: Illustration verschiedener Renderstile für zwei verschachtelte 3D-Gitter: a) äußeres Gitter im verdeckungsfreien Bereich als opake Farbkodierung sowie Isoflächendarstellung im Überlagerungsbereich; b) Kombinierte Darstellung von Schnitten und Isoflächen verschiedener Datensätze

5.4.4 Diskussion

In diesem Abschnitt wurde ein neuer Ansatz zur vergleichenden Visualisierung von abweichenden Gittern entworfen, der über bisherige Ansätze hinausgeht, diese jedoch mit integriert. Neben Standardverfahren zur vergleichenden Visualisierung (für gleiche Gitter) wurden auch die Potentiale existierender Visualisierungstechniken zum Einsatz in diesem Kontext untersucht. Damit wurde die Basis gelegt, um die vielfältigen Aufgaben beim Vergleich von Klimadaten zur visuell gestützten Evaluation von Klimamodellen zu unterstützen, die sich bisher weitgehend auf das Nebeneinanderlegen mehrerer, zumeist ungekoppelter Darstellungen beschränkt.

Im Rahmen einer Diplomarbeit (vgl. Kaeding 2006) und einer Veröffentlichung (Nocke u. a. 2007) konnte gezeigt werden, dass der vorgestellte Ansatz breite Einsatzmöglichkeiten für die vergleichende Analyse von Klimadaten hat. Im folgenden soll das dabei entstandene Framework anhand von Screenshots am Beispiel des visuellen Vergleichs zweier Klimasimulationen eines Regionalmodells (CLM) mit abweichenden Gitterauflösungen vorgestellt werden.

Abbildung 5.27 zeigt die grundlegende Fensterstruktur des Frameworks: in den vier linken Fenstern können die Datensätze auf dem *image level* untersucht, Koordinatentransformationen durchgeführt sowie Regionen von Interesse spezifiziert werden. Im mittleren, oberen Fenster zeigt eine Überblicksdarstellung die beteiligten Gitter sowie die aktuelle Region von Interesse. Im mittleren, unteren Fenster erfolgt für die Gitterpunkte der Region von Interesse eine detaillierte Wertedarstellung der beiden Datensätze. Im Parameterdialog auf der rechten Seite kann ein Filtering der Merkmale sowie eine Auswahl bestimmter Sichten sowie eine Parametrisierung der einzelnen Visualisierungsfenster erfolgen.

Für den Vergleich auf dem *image level* wurden farb- und Isolinien-kodierte Minimal-, Maximal- oder Mittelwertdarstellungen der achsenparallel ausgerichteten regulären Gitter in Drauf- und Seitenansicht umgesetzt (vgl. Abb. 5.28). Hierdurch erhält der Anwender erste Indizien über Abweichungen in den beteiligten Datensätzen.

Um einen Überblick über die beteiligten Gitter zu bekommen, können in der Überblicksdarstellung die Gitterlinien verschiedenfarbig kodiert ausgegeben (Abb. 5.29a), die eingefärbten Gitterlinien auf drei senkrecht aufeinanderstehende Ebenen projiziert (Abb. 5.29b) oder die Gitterpunkte und Gitterlinien in der Region von Interesse hervorgehoben werden (Abb. 5.29c). Weiterhin kann in

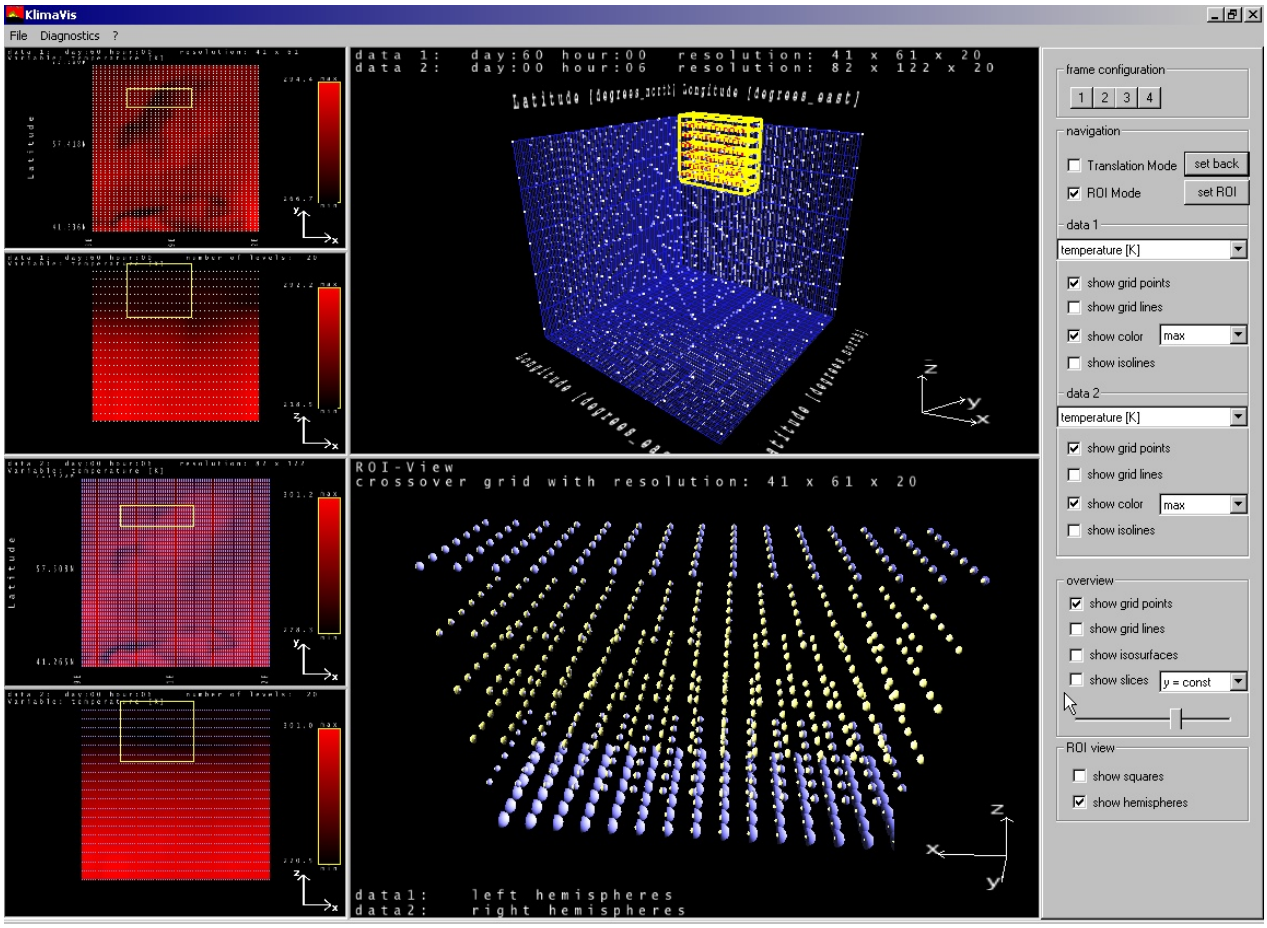


Abbildung 5.27: Screenshot des Frameworks zur vergleichenden Visualisierung

der Überblicksdarstellung auch ein erster Vergleich auf dem *data level* erfolgen: der Vergleich von extrahierten Isoflächen (vgl. Abb. 5.30). Um zwei sich durchdringende Isoflächen zu vergleichen, können entweder die Gitter gegeneinander verschoben werden (vgl. Abb. 5.30a) oder die Abstände der beiden Isoflächen explizit auf eine der Flächen farbkodiert abgebildet werden (vgl. Abb. 5.30b).

Bei der Darstellung der Region von Interesse können ferner die Datenwerte der beiden Gitter an jedem Gitterpunkt direkt miteinander verglichen werden (vgl. Abb. 5.31). Dies erfolgt hier durch 3D-Ikonen[22] (Abb. 5.31a), durch explizit berechnete Distanzen, die auf farbkodierte Billboards abgebildet werden (Abb. 5.31b) oder durch eine Kombination aus beiden Ansätzen (Abb. 5.31c).

Mit dem umgesetzten, prototypischen Framework konnten für den neuen Ansatz zur vergleichenden Visualisierung erste beispielhafte Techniken umgesetzt und auf Klimadatensätzen getestet werden. Darüber hinaus verbleiben jedoch eine Vielzahl von Herausforderungen für weiterführende Forschungsarbeiten. Neben der systematischen Untersuchung möglicher Abbildungsvarianten von Merkmalen und Gittern auf beliebige visuelle Attribute sowie auf verschiedene Renderingstile steht auch der Vergleich mehrerer Merkmale auf abweichenden Gittern in einem Bild und der Vergleich von Gittern aus Simulationsdaten mit gestreuten Messstationsdaten noch aus.

[22] Jeder der beiden Datenwerte wird auf den Radius einer Halbkugel abgebildet. Diese beiden Halbkugeln werden an der Schnittebene aneinander gelegt. Die Zuordnung zu den Gittern erfolgt durch eine für jedes Gitter einheitliche Farbe.

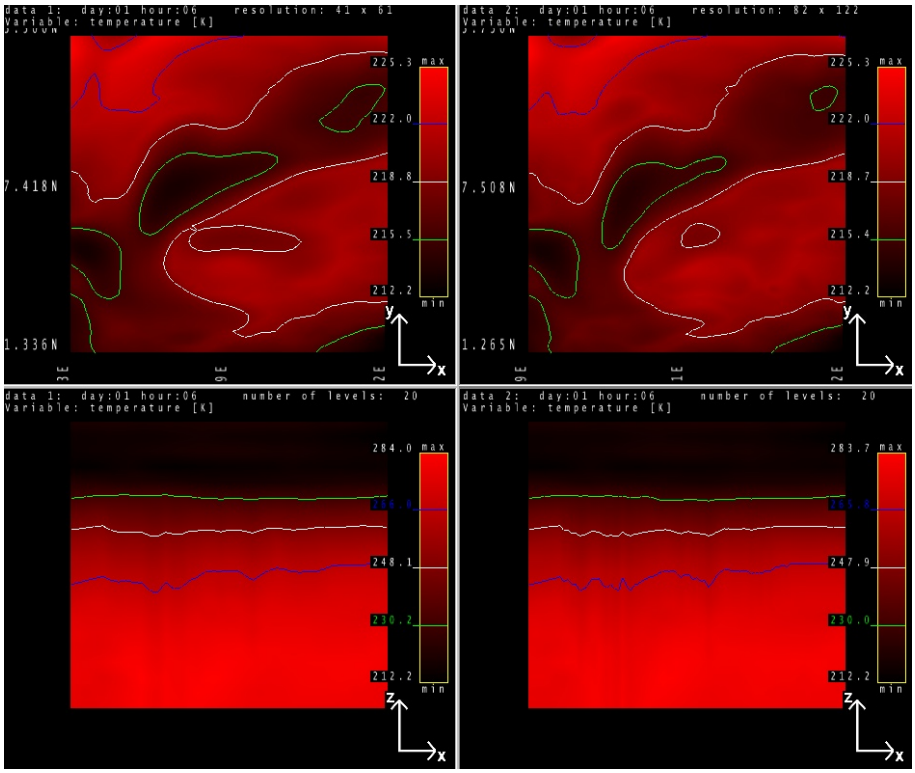


Abbildung 5.28: *Image level*-Vergleich zweier Regionalmodell-Simulationsläufe

5.5 Zusammenfassung

Mit diesem Kapitel wurde eine wichtige Grundlage für das visuelle Data Mining auf Klimadaten gelegt. Für die Visualisierung ergaben sich dabei besondere Herausforderungen aufgrund einer breiten Paletten von Datensätzen und zugehörigen Aufgabenstellungen der Klimaforscher. Entsprechend wurde eine Vielzahl von Visualisierungstechniken z.T. neu entworfen bzw. an den Kontext angepasst und umgesetzt. Hierzu gehören sowohl Standardtechniken, die für die speziellen Anforderungen der Klimaforschung erweitert und geeignet parametrisiert wurden und neue Techniken, die speziell auf die Charakteristika der Klimadaten und den Anwendungskontext zugeschnitten wurden.

Neben den in diesem Umfeld üblichen räumlichen Darstellungen wurde auch die Anwendbarkeit spezieller Darstellungstechniken für den zeitlichen Bezug der Daten sowie für die Darstellung des Merkmalsraumes untersucht. Um den Einsatz von im Anwendungsumfeld bisher unbekannten Techniken zu erleichtern, wurden insbesondere neue, leicht verständliche, metapherbasierte Darstellungstechniken entwickelt und/oder umgesetzt sowie den Anwendern durch hohe Interaktivität neue Möglichkeiten der Datenanalyse eröffnet. Ferner wurde für den Vergleich von Klimadatensätzen auf abweichenden Gittern ein neuer Ansatz entworfen und dessen Potential an beispielhaften Darstellungstechniken untersucht.

Neben dem Einsatz weiterer aus dem Visualisierungsumfeld bekannter Techniken für die vielfältigen Datensätze in diesem Umfeld verbleibt im besonderen das Problem, den Anwender bei der Auswahl und Parametrisierung aus der Vielfalt vorgestellter Techniken zu unterstützen. Darüber hinaus muss im folgenden untersucht werden, inwieweit die vorgestellten Techniken auch in Kombination mit automatischen Berechnungsverfahren - wie der im Klimaumfeld häufig angewendeten Clusteranalyse - einsetzbar sind oder inwieweit für die Verknüpfung mit automatischen Verfahren neue Methoden und Vorgehensweisen erforderlich sind.

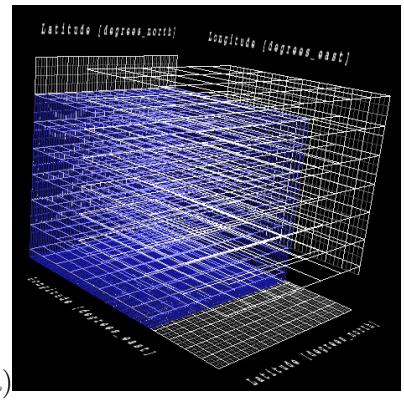
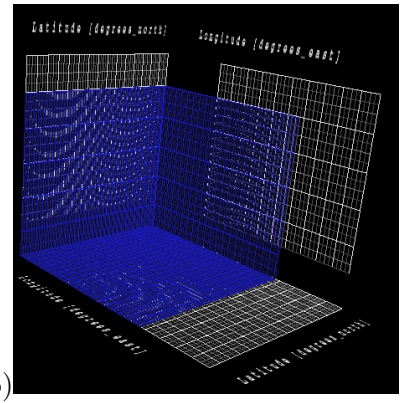
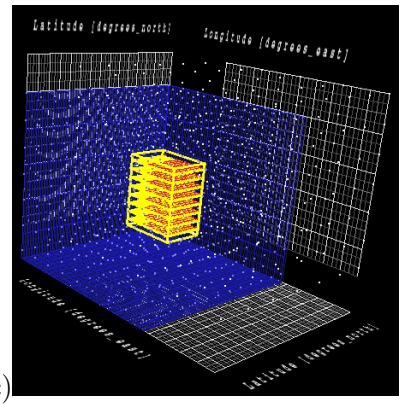
a)  b)  c) 

Abbildung 5.29: Vergleichende Überblicksdarstellung mit eingefärbten Gitterlinien (a), projizierten Gitterlinien (b) und Gitterpunkten, Gitterlinien mit einer hervorgehobenen Region von Interesse (c)

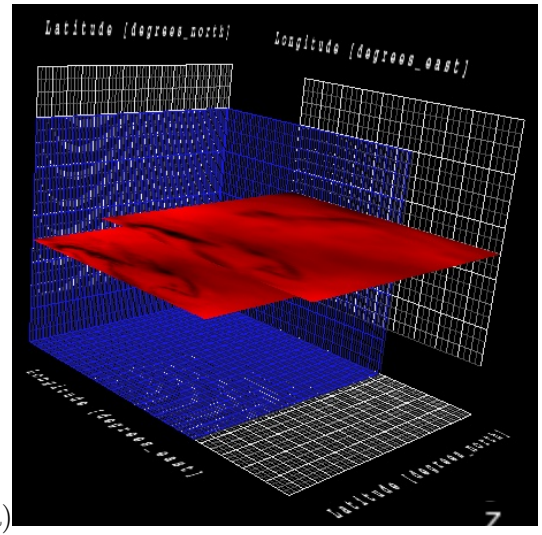
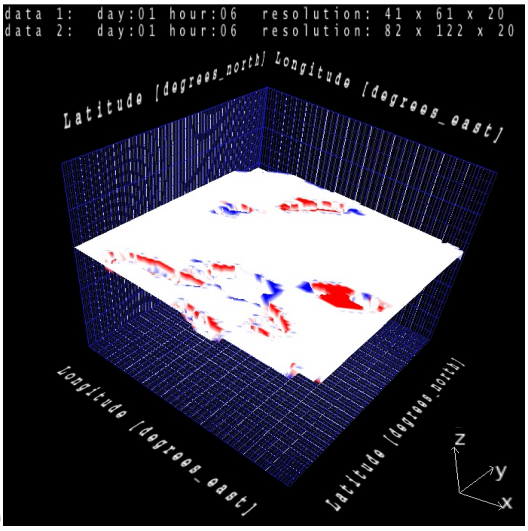
a)  b) 

Abbildung 5.30: Überblicksdarstellung mit Vergleich zweier Isoflächen: a) Verschiebung der Isoflächen zur Vermeidung von Verdeckungen; b) Farbkodierung einer Isofläche mit der punktweisen Distanz (in z-Richtung) der beiden Isoflächen

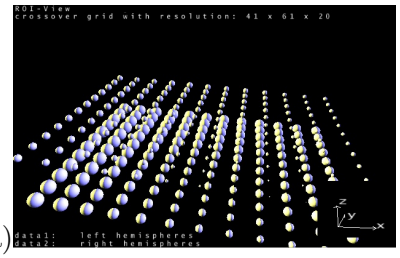
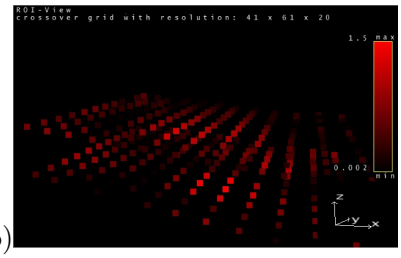
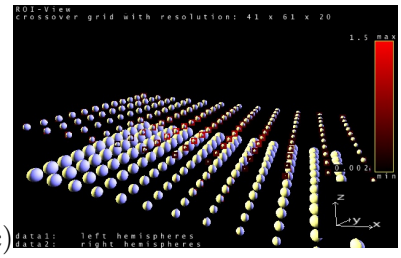
a)  b)  c) 

Abbildung 5.31: Vergleichende Visualisierung eines Datenausschnittes in der Region von Interesse: a) ikonifizierte Halbkugeldarstellung; b) Differenzdarstellung mit farbkodierten Billboards; c) Kombination aus a) und b)

Kapitel 6

Visuelles Data Mining auf Klimadaten

Mit den im vorangegangenen Kapitel vorgestellten Visualisierungstechniken wurde eine breite Basis gelegt, um Klimadaten darzustellen. Vorteil beim Einsatz solcher Visualisierungstechniken ist, dass sie eine intuitive, schnelle Aufnahme auch großer Datenmengen durch die Anwender ermöglichen. Dazu beziehen sie in einem interaktiven Prozess die außerordentlichen menschlichen Wahrnehmungsmechanismen und das kreative Potential der Nutzer in die Datenanalyse ein. Allerdings sind sie nur begrenzt geeignet, wenn quantitative Aussagen über die Daten gemacht werden sollen (z.B. Hypothesentestung). Auch bei der Handhabbarmachung insbesondere sehr großer Datenmengen z.B. durch die Reduktion der Daten, durch Vorsortierung oder durch die Extraktion dominanter Eigenschaften ist der alleinige Einsatz von Visualisierungstechniken oft nicht ausreichend. Für diese Aufgaben wurde eine breite Palette an automatischen Data Mining Methoden aus verschiedenen Gebieten wie z.B. Statistik und künstlicher Intelligenz entwickelt. Automatische Verfahren zeichnen sich dadurch aus, dass sie sehr schnell auch komplexe numerische Berechnungen und Suchvorgänge auf den Daten durchführen können, und dabei sehr genaue Ergebnisse liefern. Allerdings muss das Problem hierfür exakt spezifiziert werden können.

Um die Vorteile sowohl von Visualisierungstechniken als auch von automatischen Verfahren zu verbinden, werden diese im visuellen Data Mining miteinander verknüpft. Allerdings sind bei dieser Kopplung noch immer viele Probleme ungelöst. Als zweiten Schwerpunkt dieser Arbeit soll deswegen die Verknüpfung verschiedener automatischer Verfahren mit Visualisierungsmethoden vor dem Hintergrund der speziellen Problemstellungen der Klimaforschung diskutiert werden.

So ergeben sich spezielle Herausforderungen bei der Verzahnung von **Clusteranalyse und Visualisierung** sowie von **Hauptkomponentenanalyse und Visualisierung**, wenn die Daten im räumlichen oder zeitlichen Kontext vorliegen. Des Weiteren konzentrieren sich bisherige Arbeiten in diesem Umfeld auf die enge Verfahrenskopplung in der **explorativen Analyse**. Eine Kopplung von Visualisierungstechniken und automatischen Data Mining Methoden auch in der **konfirmativen Analyse** wurde bisher kaum untersucht. Erst der kombinierte Einsatz von visuellen und automatischen Methoden sowohl in der Exploration als auch in der Konfirmation legt die Basis, um den gesamten Prozess von **Klimamodellbildung, -simulation und -evaluation** effektiv zu unterstützen.

Häufig werden Visualisierungstechniken zur - zumeist statischen - Ausgabe der Resultate automatischer Berechnungen eingesetzt. Dies beschränkt jedoch die breite Palette an Kopplungsmöglichkeiten von visuellen und automatischen Methoden:

- Darstellung der Ergebnisse von automatischen Methoden:
  - visuelle Veranschaulichung des Vorgehens: Verstehen der Verfahren,
  - visuelle Veranschaulichung der Ergebnisse: Interpretation der Resultate,

- Kombinierte Darstellung der Originaldaten und der Ergebnisse automatischer Methoden,
  - Anreicherung von Darstellungen der Originaldaten mit zusätzlichen Informationen
  - Verdeutlichung der Auswirkungen bei der Darstellung von Ergebnissen automatischer Verfahren auf die Interpretation der Daten,
- Unterstützung von automatischen Methoden bei der Auswahl und Parametrisierung von Visualisierungsmethoden,
- Unterstützung von Visualisierungsmethoden bei der Auswahl und Parametrisierung von automatischen Methoden.

Durch Einbeziehung dieser Schritte in das visuelle Data Mining wird dieses zu einem vernetzten, iterativen und interaktiven Prozess, und geht weit über die reine Darstellung der Resultate automatischer Verfahren hinaus.

Dieses Kapitel gliedert sich wie folgt: zuerst wird die Kopplung von VDM-Verfahren zur Analyse von Klimadaten am Beispiel von *Visualisierung und Clusteranalyse* (Abs. 6.1) und von *Visualisierung und Hauptkomponentenanalyse* (Abs. 6.2) untersucht. Im Anschluss daran wird die Unterstützung von *Modellbildung, -simulation und -evaluation* durch Methoden des VDM am Beispiel von Klimamodellen diskutiert (Abs. 6.3). Abschließend werden die Ergebnisse zusammengefasst (Abs. 6.4).

6.1 Visualisierung und Clusteranalyse auf Klimadaten

Ein wichtiges Mittel zur Untersuchung von Klimadaten ist der Einsatz der Clusteranalyse (vgl. z.B. Böhm 1999; Kücken u. a. 1999; Böhm u. a. 2004). Die Clusteranalyse fasst ähnliche Datenobjekte zu Clustern zusammenzufassen, während unähnliche Datenobjekte in verschiedene Cluster einsortiert werden (vgl. z.B. Bock 1974, für eine Übersicht). Die sich ergebenden Gruppen von Objekten haben in Abhängigkeit von den eingesetzten Ähnlichkeits- bzw. Distanzmaßen[1] und den Eigenschaften des Clusterverfahrens homogene Eigenschaften. Im Unterschied zu Klassifikationsverfahren, wo die Anzahl und Eigenschaften der Zielklassen bereits feststeht, werden die Eigenschaften der Zielcluster und je nach Verfahren auch deren Anzahl, erst bei der Ausführung des Clusterverfahrens bestimmt. Damit ermöglicht die Clusteranalyse auch die Aufdeckung bisher in den Daten unbekannter Muster, weswegen sie auch als unüberwachtes Lernverfahren bezeichnet wird. Grundsätzlich unterscheidet man in disjunkte[2], nichtdiskjunkte[3] und hierarchische Clusterverfahren[4]. Daneben haben sich Verfahren etabliert, die den Informationsraum auf einen 2D- oder 3D-Raum abbilden und dabei die Ähnlichkeit der Datenobjekte weitgehend erhalten. Hierbei wird eine Anordnung der Daten durchgeführt, deren Visualisierung Schlüsse über in den Daten auftretende Gruppierungen und deren Eigenschaften ermöglichen (vgl. z.B. Multidimensional Scaling Verfahren).

Die disjunkte Clusteranalyse erlaubt es, multivariate Daten auf ein nominales Merkmal, die Clusterzugehörigkeit, zu reduzieren. Dadurch werden wesentliche, im Informationsraum auftretende Muster identifiziert und quantifiziert. Insbesondere können dadurch komprimierte Darstellungen erzeugt werden, was z.B. den im Klimaumfeld wichtigen Vergleich größerer, multivariater Datensätze vereinfacht. Weiterhin können die Ergebnisse *hierarchischer Clusterungen* mit speziellen Darstellungstechniken für hierarchische Strukturen dargestellt oder als Basis für *Fokus & Kontext* und *Information Hiding* eingesetzt werden, um große, geclusterte Datenmengen zu veranschaulichen.

[1] Ähnlichkeits- und Distanzmaßen werden auch unter dem Oberbegriff *Proximitätsmaße* zusammengefasst.
[2] Alle Datenobjekte fallen in genau einen Cluster.
[3] Datenobjekte können auch zu mehreren Clustern gehören.
[4] Es wird eine Hierarchie von Clustern aufgebaut, deren Wurzel alle Datenobjekte umfasst und deren Blätter die einzelnen Objekte repräsentieren.

Im folgenden werden Herausforderungen und Einsatzmöglichkeiten bei der Kopplung von Visualisierungstechniken und Clusterverfahren (Abs. 6.1.1) und verschiedene Lösungsvorschläge (Abs. 6.1.2 - 6.1.7) diskutiert.

6.1.1 Anspruch, Herausforderungen und Zielstellungen

In der Literatur findet sich eine Vielzahl von Ansätzen zur Visualisierung von Clusterungen. Verschiedene Data Mining Systeme (z.B. MineSet in Brunk u. a. (1997)) stellen Clusterverfahren und Visualisierungstechniken zur graphischen Darstellung ihrer Ergebnisse bereit (vgl. z.B. Westphal u. Blaxton 1998, für einen Überblick). Weiterhin können nahezu alle Techniken zur Darstellung multivariater Daten auch zur Clusterdarstellung eingesetzt werden, z.B. Hierarchiedarstellungstechniken zur Visualisierung hierarchischer Clusterungen.

Allerdings sind bei einer engen Kopplung von Clusteranalyse und Visualisierung vielfältige Probleme zu lösen. So existieren nur wenige Ansätze, die neben der Darstellung der Cluster vor allem auch ein verbessertes Verstehen der Clusterverfahren und deren Ergebnisse sowie die Ausnutzung der durch die Cluster gewonnenen Informationen bei der Interaktion und der Parametrisierung der Visualisierungstechniken mit einbeziehen. Die Herausforderung speziell bei der Analyse von geclusterten Klimadaten besteht darin, auch große Clusteranzahlen im räumlichen und zeitlichen Bezug darzustellen und die Interpretation der Clusterergebnisse zu verbessern.

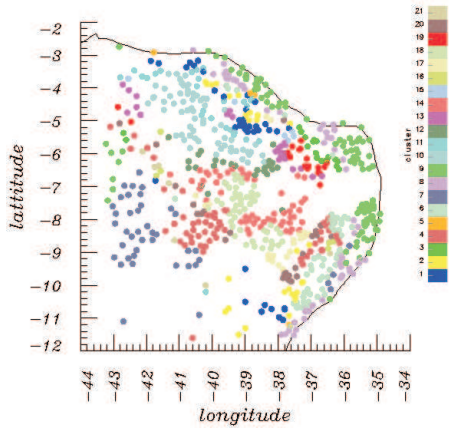


Abbildung 6.1: Farbdarstellung der Clusterzugehörigkeit für den geclusterten Maisanbaudatensatz in Brasilien

So muss bei der Darstellung von (disjunkten) Clustern die nominale **Clusterzugehörigkeit** der einzelnen Objekte geeignet abgebildet werden. Typischerweise werden nominale Merkmale neben Position und Textur auf den **Farbton** abgebildet. Hierbei ergibt sich ein allgemeines Problem, welches nicht auf die Darstellung von Clustern beschränkt ist (vgl. Brewer 1999), dass nämlich ähnliche Farben eine Ähnlichkeit der assoziierten farbkodierten Objekte suggerieren. Zum einen besteht das Problem, dass typische Farbskalen wie die Regenbogenfarbskala nicht uniform wahrgenommen werden. Zum anderen ergibt sich gerade bei einer Clusterung mit vielen Clustern die Herausforderung, gut unterscheidbare Farben zu finden, die, falls dies nicht durch die zugrunde liegenden Daten gestützt wird, keine Ähnlichkeiten von Clustern suggerieren. Abbildung 6.1 zeigt einen solchen Problemfall, wo bei einer Clusterzahl von 21 diese zwar unterschieden, aber die Darstellung nicht ohne die Verwendung ähnlicher Farben auskommt. Entsprechend müssen bei der Clusterdarstellung geeignete Farbkodierungen verwendet werden, um Missverständnisse zu vermeiden.

Neben der Darstellung der Clusterzugehörigkeit ist eine wichtige Aufgabe, die **Eigenschaften der Cluster** wie z.B. Clustermittelwerte[5] oder die Clusterhomogenität mit in die Darstellung einzubeziehen. Die Einbindung dieser Eigenschaften in die Clustervisualisierung verbessert die Interpretation, die Evaluation und den Vergleich der Clusterverfahren und der dabei eingesetzten Proximitätsmaße sowie die Untersuchung der resultierenden Cluster in Kombination mit den zugrunde liegenden Daten. Bisher finden sich hierzu in der Literatur nur wenige Ansätze, zum Beispiel in der Kalenderdarstellung (vgl. van Wijk u. van Selow 1999) oder in zylindrischen Ikonen auf einer Hierarchiedarstellung (vgl. Kreuseler u. a. 2003).

[5] Clustermittelwerte, oder auch Clustercentroide, entsprechen dem Mittelwert aller Vektoren der zu einem Cluster gehörigen Objekte.

Eine weitere Herausforderung stellt die Darstellung von **Clustern in ihrem räumlichen oder zeitlichen Bezug** dar, was wesentlich zum Verständnis der Daten beiträgt und bisher in der Literatur nur begrenzt untersucht wurde. Gerade bei der Analyse von Klimadaten ist eine visuelle Assoziation von Clustern mit den zugehörigen Zeitschritten oder geographischen Regionen besonders wichtig.

Zielstellungen speziell aus Sicht der Klimafolgenforschung sind, mit Hilfe von Clusteranalyse und Visualisierung das Verständnis von natürlichen klimatischen Prozessen zu verbessern, die Qualität von Klimamodellsimulationen einzuschätzen und dominante Eigenschaften klimatischer Systeme zu identifizieren.

6.1.2 Kodierung der Clusterzugehörigkeit

Abbildung von Clusterzugehörigkeiten auf Farbe. Die Clusterzugehörigkeit ist ein nominales Merkmal, da zwischen den einzelnen Clustern keine Ordnung definiert ist. Nach Mackinlay (1986) sind für die Abbildung nominaler Merkmale insbesondere die visuellen Attribute Position, Farbe und Textur geeignet. Im Kontext der Visualisierung von geclusterten Klimadaten wird die Position typischerweise verwendet, um den räumlichen oder zeitlichen Bezug der Daten zu veranschaulichen, weswegen sie nur eingeschränkt zur Abbildung der Clusterzugehörigkeit geeignet ist. Auch Texturen eignen sich zur Abbildung nominaler Merkmale (vgl. auch Bertin 1983), bedürfen aber einer größeren Ausdehnung in der Darstellung, erlauben eher die Abbildung globaler und lokaler Eigenschaften und weniger eine Identifikation einzelner Datenwerte. So sind sie z.B. beschränkt geeignet, Gebiete auf Karten Clustern zuzuordnen, stoßen jedoch bei vielen, kleinen Gebieten schnell an ihre Grenzen. Deswegen konzentriert sich diese Arbeit auf den *Einsatz von Farben* für die Darstellung disjunkter Clusterungen (vgl. Nocke u. a. 2004).

Die Forschung zur expressiven und problemspezifischen Farbkodierung hat eine lange Tradition in der Kartographie und Visualisierung (vgl. z.B. MacEachren 1994; Brewer 1999; Kalvin u. a. 2000; Kindlmann u. a. 2002). Die Auswahl einer geeigneten Farbabbildung hängt sowohl von den Datencharakteristika als auch von den Zielstellungen der Analyse ab, und die Anwender werden auch von gängigen Visualisierungssystemen wie dem OpenDX nur teilweise dabei unterstützt (vgl. z.B. Bergman u. a. 1995). So ergeben sich verschiedene Probleme, wenn die - z.B. im OpenDX als Standard voreingestellte - Regenbogenfarbskala für die Darstellung der Clusterzugehörigkeiten eingesetzt wird (vgl. Abb. 6.2a):

- Unterschiede bei den wahrgenommenen Farbdifferenzen (einige benachbarte Farben werden ähnlicher als andere wahrgenommen),
- abweichende Helligkeiten (einige Farben werden intensiver wahrgenommen als andere, selbst wenn sie die selbe Sättigung haben),
- generelle Probleme bei der wahrnehmbaren Unterscheidung von (in diesem Fall 13) verschiedenen Farben (wenn mehr Cluster vorliegen als gut unterscheidbare Farben der selben Helligkeit), und
- Suggestion einer Ordnung der (nominalen) Clusterzugehörigkeiten (z.B. schon aufgrund einer geordneten Farblegende).

Zur Vermeidung dieser Probleme, wurde im Rahmen dieser Arbeit eine speziell auf die Farbabbildung von Clustern zugeschnittene Farbskala entworfen. Dazu wurde die Standardregenbogenskala einerseits angepasst, so dass die Farbdifferenzen besser wahrnehmbar sind, und andererseits umgeordnet, um den Eindruck einer Clusterordnung zu vermeiden. Abbildung 6.2 vergleicht die OpenDX Standardregenbogenfarbskala (Abb. 6.2a) mit einer abgepassten Farbskala (Abb. 6.2b), mit einer umgeordneten Farbskala (Abb. 6.2c) sowie einer isoluminanten Farbskala (Abb. 6.2d).

Trotz der linearen Interpolation im Farbraum in Abbildung 6.2a ist deutlich zu erkennen, dass

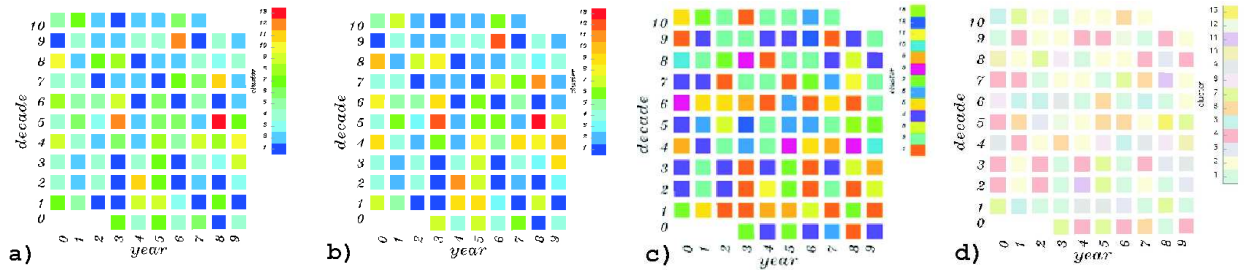


Abbildung 6.2: Darstellung des geclusterten Sommerdatensatzes der Potsdamer Reihe mit der Rechteckmethode: a) Standardregenbogenfarbskala (rot-blau); b) angepasste Regenbogenfarbskala mit Verstärkung der farblichen Differenzen; c) umsortierte Darstellung (basierend auf dem gesamten Farbkreis); d) isoluminante Farbdarstellung

türkise, grüne und gelbgrüne Farbdifferenzen (vgl. Cluster 3 bis 9 in Abb. 6.2a) wesentlich schlechter zu unterscheiden sind als Farbunterschiede im roten und blauen Farbbereich (vgl. Cluster 1 bis 3 und 10 bis 13 in Abb. 6.2a). Aus diesem Grund wurden die Farbabstände zwischen der Hauptfarben angepasst (Vergrößerung die Abstände zwischen Rot und Gelb und zwischen Blau und Türkis, und Verringerung der Abstände bei den anderen Farben; Abb. 6.2b). Um weiterhin den Eindruck einer Ordnung der Cluster zu vermeiden, werden die Farben der auf der Legende auf dem Farbkreis (mit gesättigten Farben) automatisch umsortiert. Dazu wurde eine Funktion entwickelt, die den Farbkreis derart abtastet, dass (1) zwei aufeinander folgende Farben einen maximalen Abstand zueinander haben und dass (2) jede neue Farbe eine maximale Distanz zu allen bisher gewählten Farben hat. Abbildung 6.3 veranschaulicht die umgesetzte Abtastungsfunktion zur Farbumsortierung, die nur in uniformen Farbräumen effektiv einsetzbar ist. Die Abtastung erfolgt derart, dass möglichst viele, gut unterscheidbare, gesättigte Farben ausgewählt werden.

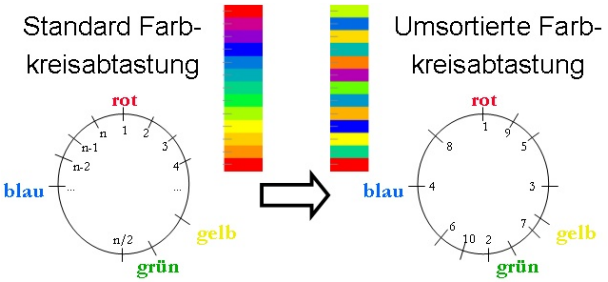


Abbildung 6.3: Veranschaulichung der Abtastfunktion des Farbkreises für Clusterzugehörigkeiten mit gestauchtem Farbkreis (links) und der Illustrierung der Umsortierungsfunktion (rechts)

Um die hierbei auftretenden Helligkeitsunterschiede abzumildern, kann der Farbkreis zum einen mit einem Start-Offset abgetastet werden, so dass die hervorstechenden Farbtöne wie rot und blau zumindest bei kleineren Clusterungen nicht gewählt werden (vgl. Abb. 6.2c). Eine andere Möglichkeit zur Vermeidung von Helligkeitsunterschieden ist es, direkt eine isoluminante Farbskala zu verwenden (vgl. Abb. 6.2d, oder auch auch Kindlmann u. a. (2002)). Allerdings ist die Unterscheidbarkeit hier auf wenige Cluster beschränkt und es kommt zu Kontrastproblemen mit dem Hintergrund.

Bei steigender Clusterzahl, treten jedoch schnell ähnliche Farben auf, die nicht in den Daten vorhandene Ähnlichkeiten zwischen den Clustern suggerieren. Um dieses Problem bei großen Clusterzahlen zumindest zu beherrschen, können neben der beschriebenen Farbwahl auf dem Farbkreis auch ungesättigte vordefinierte Farben zur Kodierung mit einbezogen werden (vgl. Abb. 6.1). Dies erlaubt eine individuelle, problemangepasste Einstellung der Farben für Präsentationsaufgaben, wobei Cluster mit spezifischen Eigenschaften bewusst hervorgehoben werden können.

Mit den vorgestellten Farbabbildungen stehen den Anwendern geeignete Kodierungsmöglichkeiten von Clusterzugehörigkeiten für disjunkte Clusterungen zur Verfügung, die eine gute Identifikation von Clustern unter Vermeidung von Fehlinterpretationen erlauben. Allerdings lässt sich bei der Abbildung der Clusterzugehörigkeiten auf Farbe nicht vermeiden, bei steigenden Clusteranzahlen ungewollte Clusterähnlichkeiten zu suggerieren.

Hierarchieerzeugung und -darstellung für hierarchische Clusterungen. Hierarchische Clusterverfahren erzeugen einen binären Baum von ineinander verschachtelten Clustern, das so genannte Dendrogramm, in dem jedem Cluster die zugehörigen Heterogenitätswerte zugeordnet werden[6]. Ein Vorteil solcher Verfahren ist, dass mit dem Dendrogramm wesentlich detailliertere Informationen über Abhängigkeiten von Clustern vorliegen, als es bei einfachen disjunktiven Clusterungen der Fall ist. Nachteil typischer hierarchischer Clusterverfahren (z.B. agglomerativer Verfahren wie Ward oder Single Linkage) ist, dass sie aus statistischer Sicht keine optimale Clusterung erzeugen müssen, da einmal gefällte Entscheidungen bei der Zuordnung eines Clusters bzw. Objekts zu einem Cluster in einem späteren Schritt nicht wieder rückgängig gemacht werden können (greedy Verfahren).

Für die Ergebnisse hierarchischer Clusterungen lassen sich Darstellungstechniken für Dendrogramme und beliebige Hierarchien verwenden. Hierbei ist es für die Visualisierung eine große Unterstützung, wenn der Anwender die Größe der *darzustellenden Hierarchie interaktiv steuern* kann.

Unter Angabe einer Heterogenitätsschranke lassen sich beliebige Schnitte innerhalb des Dendrogramms definieren, wodurch eine disjunkte Clusterung entsteht. Alle dabei extrahierten Cluster weisen eine gemäß der Schranke zugesicherte Homogenität auf. Dies gibt dem Anwender ein flexibles Werkzeug an die Hand, um im Sinne von Details-on-Demand verschiedene Sichten auf die Daten zu erzeugen. Hierbei kann auch die Anzahl gleichzeitig darzustellender Cluster beschränkt werden, indem z.B. nur bestimmte Cluster einer Sicht weiter verfeinert werden. So können die oben diskutierten Probleme bei der Darstellung disjunkter Clusterungen unter Farbkodierung einer großen Anzahl an Clustern vermieden werden. Es ist allerdings nicht immer einfach, geeignete Heterogenitätsschranken festzulegen, die eine „gute Clusterung“ bestimmen. Andererseits ergibt sich bei der Darstellung des vollständigen Dendrogramms das Problem, dass diese bei größeren Datensätzen schnell unübersichtlich werden, da die Anzahl der darzustellenden Verzweigungen im Baum der Anzahl der Objekte entspricht.

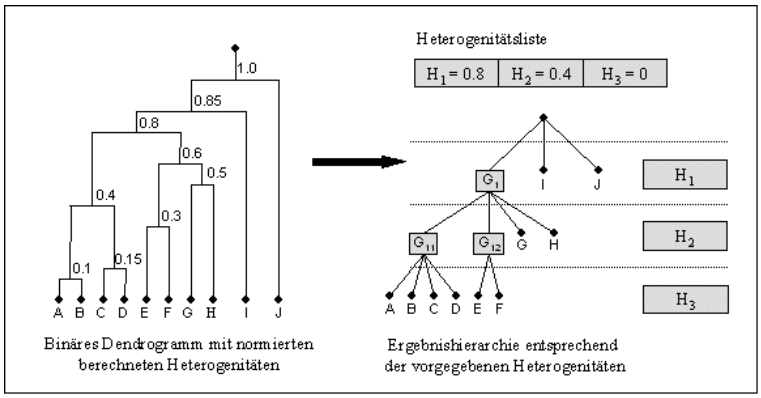


Abbildung 6.4: Algorithmus zur Erzeugung eines n-nären Hierarchiebaumes aus einem indiziertem Dendrogramm mit 10 Objekten durch Vorgabe der Heterogenitätsstufen $0.8, 0.4$ und 0

Deshalb wurde im Rahmen dieser Arbeit ein Algorithmus entworfen und umgesetzt[7], der es erlaubt, Dendrogramme in n-näre Hierarchien mit flexibler Baumtiefe umzuwandeln, wobei die Heterogenitätsschranken geeignet automatisch bestimmt und interaktiv verändert werden können (vgl. hierzu auch Nocke 1999; Kreuseler u. a. 2003). Dabei wird die Hierarchie auf wesentliche Bestandteile komprimiert, was den Einsatz von Standardbaumdarstellungstechniken wesentlich erleichtert.

Das Dendrogramm wird hierbei in mehrere Hierarchieebenen eingeteilt, in denen disjunkte Klassifikationen entsprechend einer Heterogenitätsstufenliste erzeugt werden. Abbildung 6.4 zeigt beispielhaft, wie mit diesem Algorithmus aus einem indizierten Dendrogramm mit 10 Objekten $O_1, \ldots, O_N$ ein Hierarchiebaum mit den drei Heterogenitätsstufen $0.8, 0.4$ und 0 aufgebaut wird. Dabei liegen in der Hierarchiestufe H_3 ausschließlich Einzelobjekte vor, weil nur diese eine Heterogenität von 0 besitzen. In der nächsthöher gelegenen Stufe 0.4 wurden alle die Cluster und Objekte eingegliedert, die eine maximale Heterogenität kleiner gleich 0.4 besitzen. Alle Unterklassen dieser Klassen wurden aufgrund ihrer geringeren

[6] Ein Dendrogramm mit Heterogenitätswerten wird auch *indiziertes Dendrogramm* genannt.

[7] in Zusammenarbeit mit Mathias Kreuseler

Heterogenität nicht in den neuen Baum übernommen. Hiervon ausgeschlossen sind die Klassen, die unterhalb oder auf der nächsttieferen Heterogenitätsstufe liegen. Nach dem selben Prinzip wurden Klassen und Objekte auf der Heterogenitätsstufe 0.8 eingefügt.

Verallgemeinert folgt der Algorithmus den folgenden Schritten:

1. Erzeuge die Wurzel des neu zu erzeugenden Baumes.
2. Wähle aus der Heterogenitätsstufenliste die erste Heterogenitätsstufe aus.
3. Wähle die beiden Sohnknoten der Dendrogrammwurzel aus.
4. Untersuche für die beiden aktuellen Knoten, ob ihre Heterogenitätswerte bereits kleiner gleich der aktuellen vorgegebenen Heterogenitätsstufe sind.
5.
   - Fall 1: Füge für die Knoten, die unterhalb der Schranke liegen, einen neuen Sohnknoten im Baum ein. Merke dir den Vaterknoten des untersuchten Knotens.
   - Fall 2: Für die Knoten, die oberhalb der Schranke liegen, untersuche wiederum deren Sohnknoten mit Schritt 4.
6. Wiederhole 4., bis alle gefundenen Sohnknoten unterhalb der aktuellen Hierarchiestufe liegen (dies geschieht spätestens, wenn die Klassen nur noch ein Objekt enthalten)
7. Wähle aus der Heterogenitätsstufenliste die nächste Heterogenitätsstufe aus und fahre mit Schritt 4 ausgehend von den in Schritt 5.1 gemerkten Vaterknoten fort.
8. Wenn alle Elemente der Heterogenitätsstufenliste abgearbeitet sind: STOPP des Algorithmus.

Um den Anwender bei der Findung einer geeigneten Heterogenitätsstufenliste zu unterstützen, wurde basierend auf dem Elbow-Kriterium das folgende Verfahren benutzt: (1) Durch Standardvorgabe oder Nutzereingabe wird eine geeignete Anzahl N von Hierarchiestufen vorgegeben. (2) Dann werden alle Heterogenitätswerte des indizierten Dendrogramms in eine geordnete Liste eingetragen. (3) Darauf wird eine neue, geordnete Liste aufgebaut, in welche die Differenzen von je zwei benachbarten Heterogenitäten der ersten Liste eingefügt werden. (4) Anschließend werden die ersten N Differenzen der zweiten Liste gewählt und die zugehörigen Heterogenitätswerte des unteren zugehörigen Heterogenitätswertes aus der ersten Liste in die Heterogenitätsliste eingetragen. (5) Abschließend wird der Wert 0 hinzugefügt, falls er noch nicht in der Liste vorliegt.

Durch dieses auf dem Elbow-Kriterium basierende Verfahren werden die N stabilsten disjunkten Clusterungen aus dem Dendrogramm extrahiert und lassen sich nun visualisieren. Abbildung 6.5

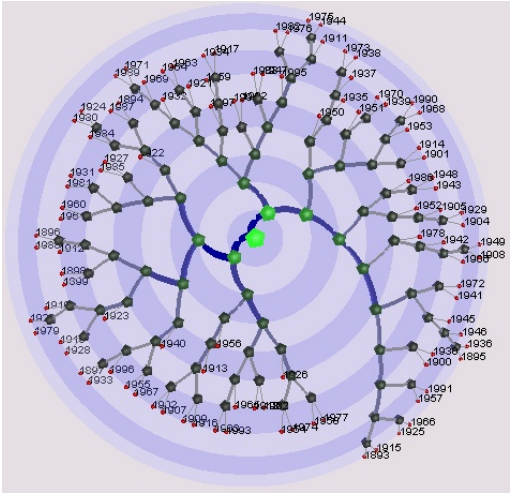
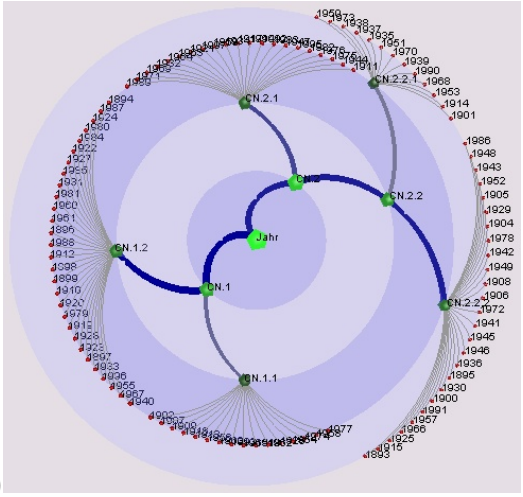
a)  b) 

Abbildung 6.5: Illustration des hierarchisch geclusterten Sommerdatensatzes der Potsdamer Reihe mit der Technik MagicEyeView: a) Dendrogramm; b) Hierarchie mit 4 Stufen

illustriert, wie am Beispiel des Datensatzes extremer Sommereigenschaften mit 105 Objekten (Jahren) die komplexe Dendrogrammdarstellung (Abb. 6.5a) zu einer übersichtlichen Hierarchie mit vier Hierarchiestufen (Abb. 6.5b) umgewandelt werden konnte.

Mit diesem Algorithmus wurde dem Anwender ein Werkzeug an die Hand gegeben, welches es erlaubt, flexibel Sichten auf die geclusterten Daten zu erzeugen, und damit die darzustellende Clusterhierarchie je nach Bedarf in verschiedenen Detailierungsstufen anzuzeigen. Damit eröffnen sich neue Möglichkeiten zur Visualisierung der Cluster, z.B. der Einsatz von Hierarchiedarstellungen, die für große Dendrogramme nicht geeignet sind.

Visuelle Hervorhebung oder Abschwächung von Clustern. Um durch die Cluster erzeugte Strukturen insbesondere im räumlichen und zeitlichen Bezug genauer zu untersuchen, ist es sinnvoll, im Laufe des Explorationsprozesses bewusst einzelne Cluster hervorzuheben oder andere zurückzunehmen bzw. sogar vollständig auszublenden. Hierzu lassen sich u.a. die folgenden Techniken verwenden:

- vergrößerte Darstellung des Clusters von Interesse (vgl. z.B. Abb. 6.8),
- Hinzufügen zusätzlicher Primitive, z.B. durch Umrahmung aller Objekte eines Clusters von Interesse (vgl. z.B. Abb. 6.6 und Abb. 6.12),
- Variation der Farbintensität, z.B. durch Einsatz von Farbton oder Helligkeit (vgl. z.B. Abb. 6.1) zum Akzentuieren oder Deakzentuieren von Clustern,
- Ausblenden von Clustern, z.B. durch Zusammenfalten von Teilbäumen eines Dendrogramms.

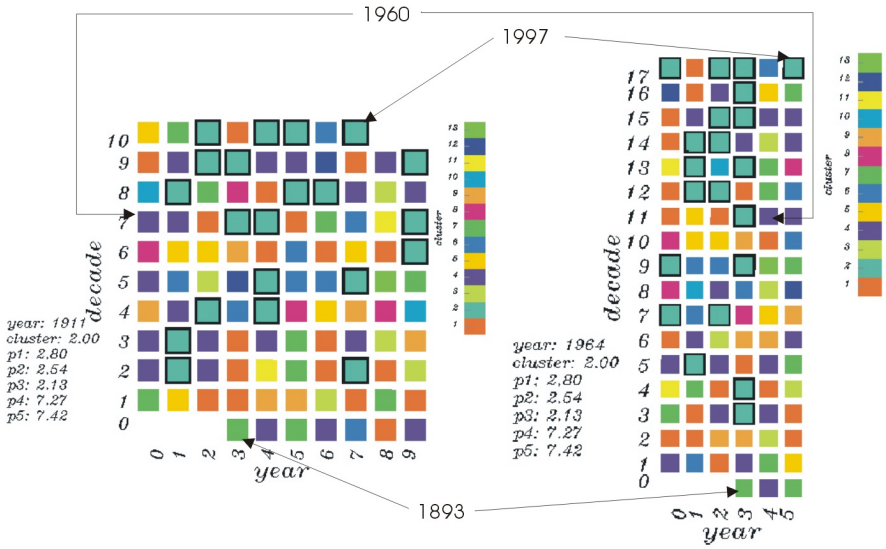


Abbildung 6.6: Darstellung der Clusterzugehörigkeiten mit der Rechteckmethode unter Hervorhebung eines Clusters; links: Periode von 10 Jahren; rechts: Periode von 6 Jahren

Abbildung 6.6 zeigt, wie bei der Rechteckmethode interaktiv ein bestimmtes Jahr (1911 in der linken und 1964 in der rechten Darstellung in Abb. 6.6) ausgewählt und alle Jahre, die zu diesem Cluster gehören, durch eine schwarze Umrahmung hervorgehoben sowie die Clustermittelwerte textuell ausgegeben werden. Das durch diese Hervorhebung erzeugte Muster kann nun analysiert werden, und z.B. durch Veränderung der Periode (Anzahl an Spalten) können versteckte Periodizitäten aufgedeckt werden (vgl. Abb. 6.6 links, wo in den letzten 2 Spalten mit Ausnahme des Jahres 1997 keine Objekte dieses Clusters liegen, was auf eine Periode hinweist; vgl. hierzu auch Nocke u. a. (2004)).

6.1.3 Kodierung der Clustereigenschaften

Typischerweise konzentrieren sich Ansätze zur Clustervisualisierung auf die Darstellung der Clusterzugehörigkeiten. Dabei werden die zugrunde liegenden Eigenschaften der Cluster häufig nicht dargestellt. So lässt sich die durch die Clusterung beschriebene Struktur erkennen, jedoch ist das Entstehen dieser Struktur und ihrer Eigenschaften schwer nachzuvollziehen. Deswegen ist es wichtig, Eigenschaften wie *zentrale Punkte* oder *Clusterheterogenitäten* in die Darstellung mit einzubeziehen. Im folgenden sollen nun Ansätze zur Darstellung von Clustereigenschaften diskutiert werden, woran sich im nächsten Abschnitt kombinierte Darstellungen von Clusterzugehörigkeiten und Clustereigenschaften anschließen. Dabei werden zumeist bekannte Darstellungstechniken für Daten im räumlichen und zeitlichen Bezug sowie im Merkmalsraum an die Bedürfnisse der Clusterdarstellung angepasst.

Um die Eigenschaften der zu einem Cluster gehörigen Objekte einzuschätzen, und damit einen grundlegenden Eindruck über die Lage des Clusters im Informationsraum zu gewinnen, werden für das Cluster typische Objekte bestimmt (vgl. z.B. Bock 1974). Diese lassen sich dann als Clusterrepräsentant in der Visualisierung einsetzen. Im allgemeinen werden hierzu *zentrale Punkte* verwendet, die entweder durch das arithmetische Mittel der zu dem Cluster gehörigen Vektoren[8] oder durch Auswahl eines zentralen Objektes in der Nähe des arithmetischen Mittels bestimmt werden[9].

Bei Klimadaten handelt es sich im allgemeinen um Merkmale mit kontinuierlichem Skalentyp, so dass auch die *zentralen Punkte* Vektoren dieses Typs bilden. Grundsätzlich lassen sich dementsprechend die visuellen Attribute Position, Länge, Winkel, Orientierung, Fläche, Volumen, Helligkeit, Sättigung und Farbton einsetzen (vgl. Mackinlay 1986).

Neben speziellen Techniken für die Darstellung von Clustern können hierbei die im vorangegangenen Kapitel vorgestellten Darstellungstechniken für kontinuierliche Klimadaten eingesetzt werden. Durch die Visualisierung der zentralen Punkte des Clusters anstelle der ursprünglichen Datenwerte ergibt sich dabei eine verstärkte Strukturierung der Darstellung, wobei z.B. Extremwerte in den Daten verstärkt werden. Das Verständnis dessen, dass es sich hierbei um *zentrale Punkte* von Clustern handelt, kann durch spezielle Legenden verbessert werden.

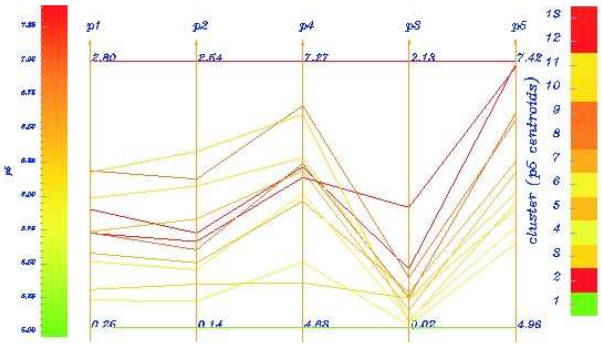


Abbildung 6.7: Parallele Koordinatendarstellung der zentralen Punkte des heiße Sommer indizierenden Datensatzes (Potsdamer Reihe); Farbkodierung basierend auf dem Merkmal p5

(Standard-)Darstellung zentraler Punkte. Zentrale Punkte von Clustern lassen sich als multivariate Daten auffassen und können dementsprechend mit den hierfür entwickelten Techniken dargestellt werden. Durch die Ersetzung der Originaldatenwerte durch die wenigen zentralen Punkte ergeben sich so sehr übersichtliche Darstellungen, die wesentliche Muster eines gesamten Datensatzes bei Vermeidung von Verdeckungen erreichen.

Abbildung 6.7 illustriert am Beispiel von Parallelen Koordinaten die Darstellung zentraler Clusterpunkte. Jeder Streckenzug repräsentiert hierbei das Element des zentralen Punktes eines einzelnen Clusters. Neben der üblichen einfarbigen Kodierung (typischerweise in schwarz) wurden hier zwei alternative Einfärbungsmethoden speziell für Cluster untersucht: Farbkodierung der Li-

[8] bei Merkmalen mit kontinuierlichen Skalentyp

[9] Haben die Cluster im Informationsraum keine runde oder ovale Form, können, um zu vermeiden, dass Objekte außerhalb des durch den Cluster überdeckten Raum genutzt werden, alternativ zu den *zentralen Punkten* auch so genannte *Kernpunkte* an Stellen hoher Objektdichte als Clusterrepräsentanten ausgewählt werden.

nienzüge (1) nach den Clusterzugehörigkeiten (vgl. Abb. A.10 rechts im Anhang) bzw. (2) nach einem einzelnen Merkmal (vgl. Abb. 6.7). Die Farbkodierung der Streckenzüge nach einem einzelnen Merkmal erlaubt es, auch die Abhängigkeiten nicht benachbarter Achsen zu untersuchen. Zum besseren Verständnis der Cluster wurden zusätzlich eine Legende zur Darstellung der Farbkodierung des aktuell ausgewählten Merkmals (Abb. 6.7 links) sowie eine Legende des entsprechenden Elementes des zentralen Punktes aller Cluster (Abb. 6.7 rechts) dargestellt.

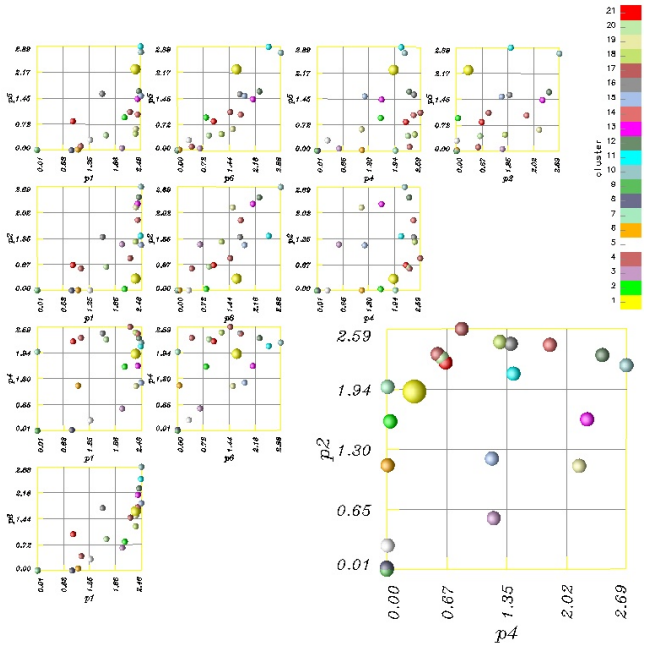


Abbildung 6.8: Scatterplot-Matrix-Darstellung von farbkodierten, zentralen Clusternpunkten (obere Dreiecksmatrix & vergrößerter Plot von Interesse); bras. Maisdatensatz; Hervorhebung eines Clusters (Nr. 1)

Als zweites Beispiel für die multivariate Darstellung zentraler Punkte zeigt Abbildung 6.8 eine Scatterplot-Matrix, in welcher die zentralen Clusterpunkte durch farbkodierte Kreise repräsentiert werden, wodurch sich die Cluster separat identifizieren lassen. Hierbei können einzelne Cluster und Scatterplots ausgewählt und vergrößert dargestellt werden (Plot p2-p4 sowie Cluster 1 (gelb)).

Räumliche Darstellung von zentralen Punkten. Abbildung 6.1 auf Seite 95 illustriert die räumliche Verteilung der Cluster(-zugehörigkeiten) des brasilianischen Maisanbaudatensatzes. So ziehen sich z.B. die Cluster 8 und 9 (helles violett und grün) entlang der Küste, während Cluster 10 (türkis) eher ein homogenes Gebiet im inneren des Landes überdeckt. Um nun die Eigenschaften dieser Cluster besser verstehen zu können, müssen die Anteile der einzelnen Elemente der zentralen Punkte im räumlichen Kontext untersucht werden. Als erste einfache Variante kann hierzu die selbe Technik wie in Abbildung 6.1 verwendet werden, welche das aktuelle Element der zentralen Punkte auf kleine farbkodierte Kreise abbildet (visuelles Attribut *Farbton*).

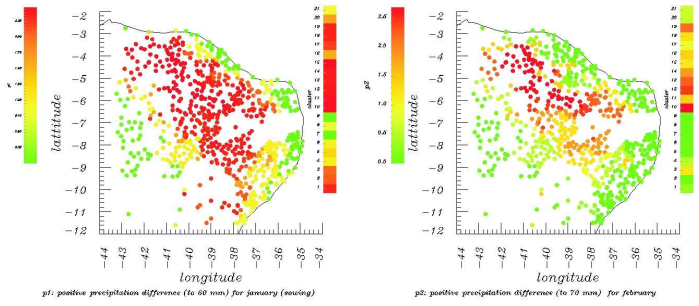


Abbildung 6.9: Darstellung zweier Vektorelemente der zentralen Punkte von gestreuten, geclusterten 2D-Klimadaten mit separater Clusterlegende

Abbildung 6.9 zeigt den *image level*-Vergleich der Elemente „monatlicher Niederschlag Januar“ (links) und „monatlicher Niederschlag Februar“ (rechts). Abbildung 6.9 enthält jeweils auf der linken Seite des Bildes zwei Farblegenden, welche die Farbkodierung des aktuellen Elementes bzw. des dargestellten Merkmales enthält. Zusätzlich zeigen zwei Farblegenden auf der rechten Seite jedes Bildes den Datenwert des aktuell darzustellenden Vektorelementes des zentralen Punktes für jeden Cluster, welche auch für die farbkodierten Kreise aller Objekte des zugehörigen Clusters verwendet werden.

So bekommt der Anwender z.B. einen Eindruck, warum die Küstenregionen zwei unterschiedlichen Clustern zugeordnet wurden und welche Merkmale den größten Einfluss auf die Stärke von Dürren haben (vgl. hierzu auch Nocke u. a. 2004). Insbesondere die rechten Farblegenden ermöglichen dabei einen schnellen visuellen Überblick über die aktuell dargestellten Clustereigenschaften.

Durch die Darstellung der farbkodierten Kreise an den Positionen der Messstationen werden zwar Detailinformationen zu den Messstationen vermittelt, es bleibt jedoch schwer, den Einflussbereich der zentralen Punkte einzuschätzen. Hierfür können Voronoi-tesselierte farbkodierte Flächen oder interpolierte Darstellungen verwendet werden (wie in Abb. 5.4c-f, vgl. auch Nocke u. a. (2003) und Nocke u. a. (2004)).

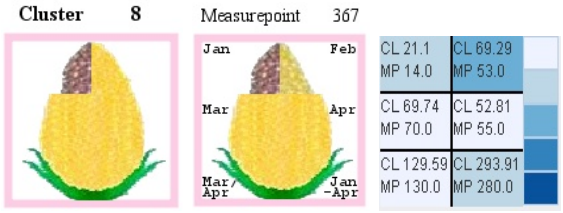


Abbildung 6.10: Legende zur vergleichenden Maisikonendarstellung von geclusterten und ungeclusterten Messstationsdaten; a) zentraler Punkt des Clusters; b) originale Messstationswerte; c) textuelle Darstellung der numerischen Werte sowie farbkodierte Differenzdarstellung

Allerdings müssen für den visuellen Vergleich mehrerer Elemente der zentralen Clusterpunkte mehrere Bilder verwendet werden, was die Untersuchung von Details erschwert. Um eine Übersicht über die zentralen Punkte in einer Darstellung zu bekommen sowie auftretende Abhängigkeiten leichter zu untersuchen, sind Ikonen besonders geeignet. So lassen sich die in Abschnitt 5.1.3 vorgestellten metapherbasierten Ikonen auch für die (räumliche) Darstellung von Clustereigenschaften einsetzen (visuelle Attribute *Form, Größe* und *Farbton*), vgl. auch Nocke u. a. (2005)). Insbesondere kann der Nutzer hierbei durch geeignete Legenden unterstützt werden, die Eigenschaften des Clusters mit den Eigenschaften zugehöriger Objekte zu vergleichen. Abbildung 6.10 zeigt eine solche Legende für den Maisdatensatz. Für eine ausgewählte Messstation wird die zugehörige Ikone (Abb. 6.10b) neben dem zentralen Punkt des zugehörigen Clusters dargestellt (Abb. 6.10a). Somit können auftretende Abweichungen dieser Messstation zum „mittleren Verhalten“ des Cluster qualitativ untersucht werden. In Abbildung 6.10 wird deutlich, dass die Maisanbaubedingungen für Februar zwischen ausgewählter Messstation und zugehörigem Cluster abweichen. Um diese Abweichung auch quantitativ auswerten zu können, wurde eine dritte Legende hinzugefügt (Abb. 6.10c). Diese stellt die den Ikonendarstellungen zugrunde liegenden Werte dar und kodiert deren Differenzen durch einen farbigen Hintergrund (Weiß-Blau-Skala). So erhält der Anwender die Möglichkeit, durch interaktive Auswahl mehrerer Messstationen eines Clusters dessen Heterogenität bezüglich einzelner Merkmale einzuschätzen.

Zeitliche Darstellung von zentralen Punkten. Zur Darstellung von geclusterten Zeitreihen lässt sich der Parameter Zeit wie jede andere Variable auffassen, womit Standardtechniken zur Zeitdarstellung und auch Techniken für multivariate Daten eingesetzt werden können (z.B. die Rechteckmethode, vgl. Nocke u. a. (2004)). Wichtig für die effektive Analyse ist es, dass der Parameter Zeit dabei in der Darstellung hervorgehoben wird. Dies leistet zum Beispiel die Technik Themenfluss[10] (vgl. Abb. 6.11). Hier wird diese zur Visualisierung einer geclusterten Klimazeitreihe (Potsdamer Reihe) eingesetzt, wobei die Originaldatenwerte durch die Werte der zentralen Punkte des jeweiliges Clustern ersetzt wurden. Vergleicht man diese Darstellung mit der Themenflussdarstellung der ungeclusterten Originaldaten (vgl. Abb. 5.17), können die durch die Clusterung verursachten Vereinfachungen in ihrem zeitlichen Verlauf eingeschätzt werden. So werden in diesem Beispiel durch die Clusterung extreme Sommer stärker betont.

6.1.4 Gekoppelte Darstellung von Clusterzugehörigkeit und -eigenschaften

In den bisher vorgestellten Clusterdarstellungen wurden entweder die Clusterzugehörigkeiten oder die Clustereigenschaften im räumlichen oder zeitlichen Bezug separat abgebildet. Eine Verknüpfung dieser beiden Clusterinformationen erfolgte lediglich über eine separate Legende, durch das Nebeneinanderlegen der einzelnen Darstellungen oder eine Farbkodierung einzelner Cluster, was einen

[10] vgl. hierzu auch eine gemeinsame Veröffentlichung mit den Klimaforschern: Böhm u. a. (2004)

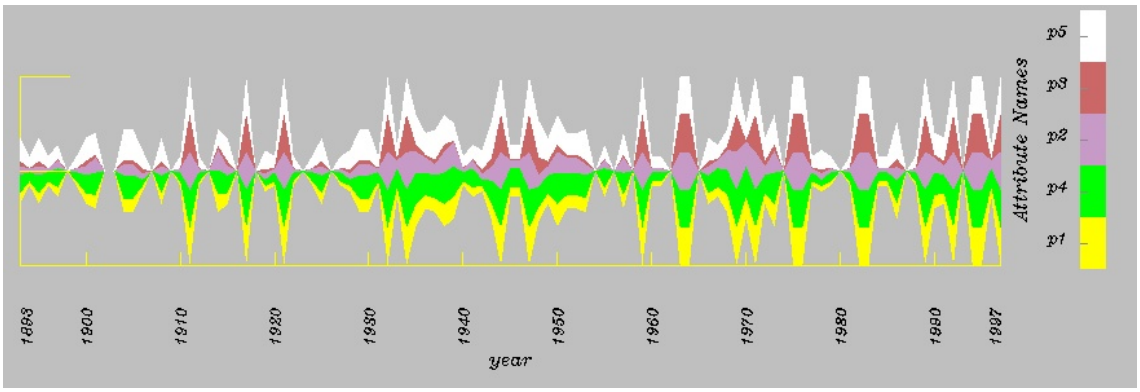


Abbildung 6.11: Themenflussdarstellung zur Darstellung von zeitlichen Trends bei den Sommereigenschaften, wobei die Sommer geclustert wurden und die Originalwerte durch die zentralen Punkte ersetzt wurden

erhöhten kognitiven Aufwand bei der Zuordnung von Clustern und deren Eigenschaften erfordert. Deswegen wurden die beiden folgenden Koppelungsmöglichkeiten untersucht:

- gemeinsame Darstellung von Clusterzugehörigkeiten und Clustereigenschaften (in einer Ansicht)
- getrennte Darstellung von Clusterzugehörigkeiten und Clustereigenschaften (mehrere Ansichten, Kopplung durch Brushing)

Gemeinsame Darstellung von Clusterzugehörigkeiten und Clustereigenschaften in einer Ikone. Eine kompakte Art, die Cluster zusammen mit deren Eigenschaften darzustellen ist es, diese in einer Ikone zu kombinieren. Exemplarisch wurde dies für den Maisanbaudatensatz in Brasilien umgesetzt (vgl. Baalcke 2005; Nocke u. a. 2005). Hierfür bietet es sich an, die metapherbasierte, sechs Merkmale darstellende Maisikonene (vgl. Abs. 5.1.3) zur Kodierung der zentralen Punkte einzusetzen, und die Clusterzugehörigkeit auf den rechteckigen Hintergrund abzubilden. Bei Anordnung der so zusammengesetzten Ikonen im geographischen Kontext erhält der Anwender sowohl eine allgemeine Übersicht über die räumliche Verteilung der Cluster, und behält dabei deren Eigenschaften im Blick. Abbildung 6.12 zeigt, wie solche Ikonen in verschiedenen Anordnungen platziert wurden (gestreutes (Abb. 6.12a), reguläres (Abb. 6.12b) und Multiresolution-Layout (Abb. 6.12c)). Bei Auswahl einzelner Objekte werden alle zugehörigen Objekte des selben Clusters durch

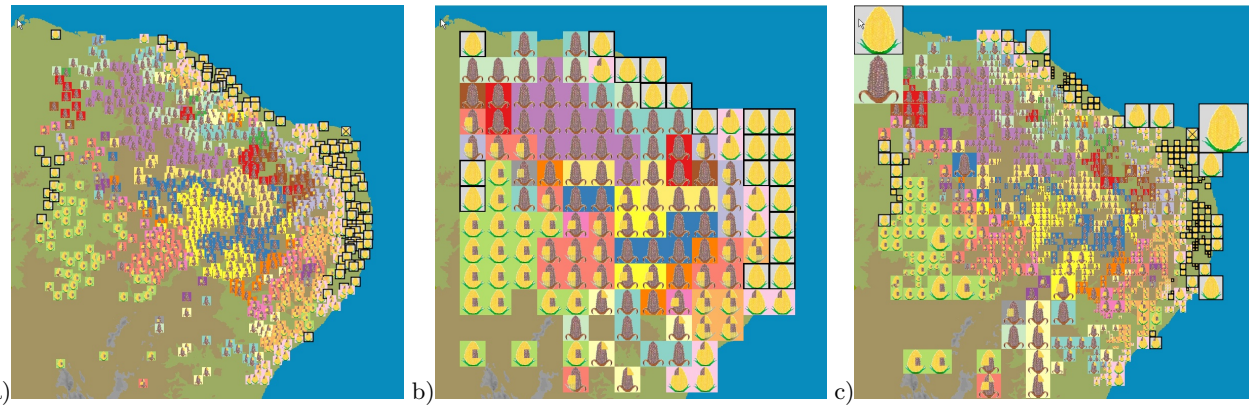


Abbildung 6.12: Metapherbasierte Ikonendarstellungen für gestreute, geclusterte 2D-Klimadaten mit zentralen Punkten (Maisikone) und Clusterzugehörigkeit (Ikonenhintergrundfarbe); a) gestreutes Layout; b) reguläres Layout; c) Multi-resolution Layout

einen schwarzen Rahmen hervorgehoben, was die Untersuchung der räumlichen Verteilung eines Clusters erleichtert (z.B. in Abb. 6.12: Cluster 9 mit Messstationen vorwiegend in Küstenlage). Sollen

Lage und Verteilung der Cluster mit darunter liegenden geographischen Informationen in Beziehung gesetzt werden, kann der Platzbedarf der Ikonen reduziert werden, indem die Clusterzugehörigkeit lediglich auf die farbige Umrandung jeder Ikone abgebildet wird (vgl. Abb. 6.22).

Getrennte Darstellung von Clusterzugehörigkeiten und Clustereigenschaften. Die bisher vorgestellten Darstellungen kombinieren Clusterzugehörigkeit und -eigenschaften in einem Plot. Dies ermöglicht kompakte Darstellungen, beschränkt aber die gleichzeitige Darstellbarkeit dieser Eigenschaften im räumlichen/zeitlichen Bezug und im Merkmalsraum. Hierfür eignen sich über Brushing & Linking gekoppelte Darstellungen.

Ein Beispiel für die Kopplung zweier Clusterdarstellungen ist eine Kombination der Rechteckmethode mit einer Parallelen Koordinaten-Darstellung (vgl. Abb. A.10 im Anhang). Dabei kann der Anwender ein Jahr von Interesse auswählen, was in der Rechteckdarstellung zu einer Hervorhebung aller anderen Jahre dieses Clusters führt, und zudem in einer separaten Parallelen Koordinaten-Darstellung eine Hervorhebung des Steckenzuges des zugehörigen zentralen Punktes bewirkt.

Ein zweites Beispiel für eine solche Kopplung bei der Visualisierung von Klimadaten ist die speziell für die Bedürfnisse von Klimadaten angepasste Kalender-Cluster-Visualisierung[11] (nach van Wijk u. van Selow (1999), vgl. Nocke u. a. (2003)). So zeigt die linke Seite von Abbildung 6.13 eine

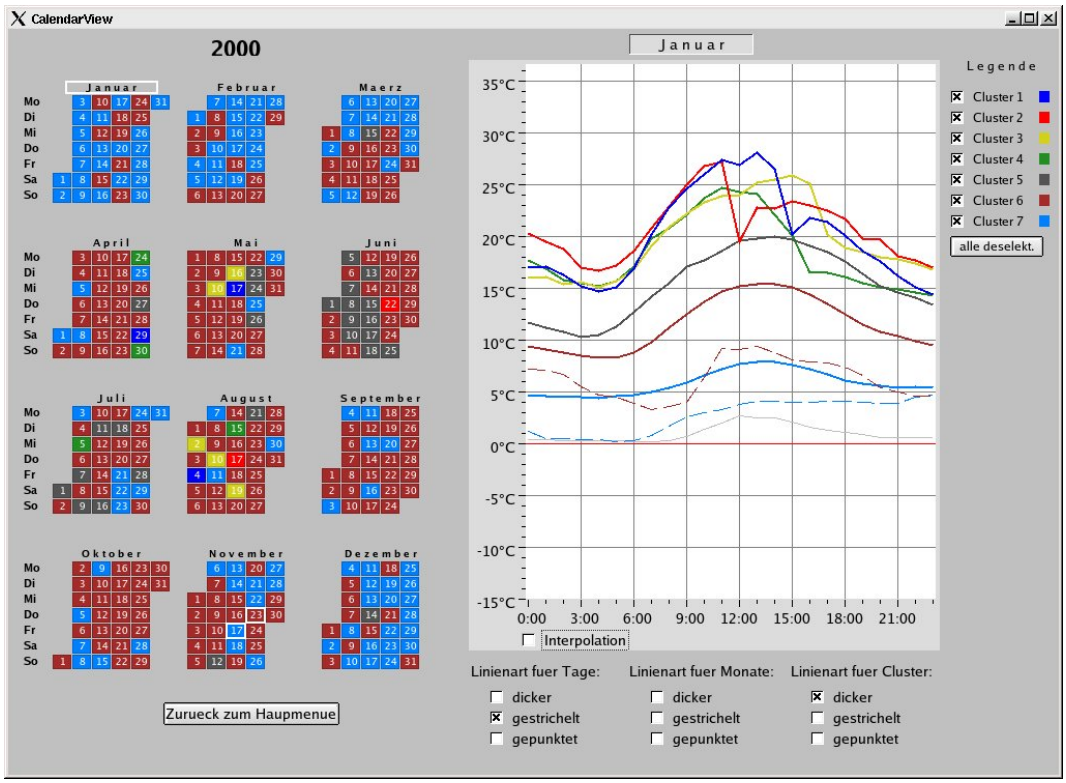


Abbildung 6.13: Kalenderbasierte Clustervisualisierung (basierend auf van Wijk u. van Selow (1999)) der Tagestemperaturverläufe basierend auf stündlichen Messung (Station Potsdam) für das Jahr 2000; Einsatz des Korrelationskoeffizienten als Ähnlichkeitsmaß

Kalenderdarstellung, in der die Clusterzugehörigkeiten jedes Tages farbkodiert werden. Die rechte Seite von Abbildung 6.13 repräsentiert die Tagestemperaturverläufe, welche neben den zentralen Punkten der Cluster auch einzelne Tagesverläufe und gemittelte Monatsverläufe anzeigen kann. Hierbei werden dem Anwender eine Vielzahl von Interaktionen an die Hand gegeben. So kann er für die Tagesverlaufsdarstellung zwischen dem gemittelten Clusterrepräsentanten (Clustermittelwert) und dem für den Cluster typischsten Tagesverlauf (Clustermedian) eines zu ihm gehörenden Tages

[11] Die Umsetzung erfolgte im Rahmen eines betreuten Studentenprojektes (Struck u. Marczok 2003).

auswählen. Zusätzlich können Tage, Monate oder Cluster beliebig an- und ausgeblendet sowie mit verschiedenen Linienstilen hervorgehoben oder zurückgenommen werden. Hierdurch können extreme und mittlere Bedingungen bestimmter Cluster und der zugehörigen Tage effektiv untersucht und Überlappungen vermieden werden.

Damit steht den Klimaforschern ein Werkzeug zur Verfügung, mit dem sie typische und untypische Cluster und deren Beziehungen identifizieren und so Klimazustände besser verstehen können. So repräsentieren in diesem Beispiel (vgl. Abb. 6.13) die Cluster 6 und 7 typische Tagesverläufe. Alle anderen weisen ein eher untypisches Verhalten auf und können als Ausreißer identifiziert werden. Des Weiteren können im farbkodierten Kalender schnelle Änderungen in Clusterabfolgen aufgedeckt werden (z.B. im August in Abb. 6.13, links). Zusammenfassend erlaubt die Technik den Vergleich von Clustern und den zugehörigen Tagesverläufen (Überblick), die Untersuchung einzelner Cluster sowie die Exploration von täglichen und monatlichen Verläufen von Interesse (Details).

Zusätzlich lassen sich bei Bedarf in einer separaten Ansicht auch detaillierte Eigenschaften eines ausgewählten Clusters oder Objektes darstellen (vgl. z.B. Abb. 6.10).

6.1.5 Visueller Vergleich von Clusterungen

Ein spezielles Problem, welches in der Literatur bisher kaum betrachtet wurde, ist der systematische Vergleich von Clusterungen mit räumlichen bzw. zeitlichem Bezug. So wird die Clusterung insbesondere in der Klimaforschung eingesetzt, um verschiedene Modellsimulationen und/oder Messdaten mit einer Vielzahl von Merkmalen in ihrem grundlegenden Verhalten miteinander zu vergleichen und dadurch zu evaluieren (vgl. z.B. Böhm u. a. 2004). Die dabei eingesetzten Standarddarstellungstechniken decken jedoch bei weitem nicht die vielfältigen Fragestellungen ab, die sich bei einem solchen Vergleich ergeben.

Um das breite Spektrum an visuellen Vergleichsmöglichkeiten von Clusterungen zu systematisieren, lassen sich die folgenden Kriterien eingrenzen:

1. Vergleich gleicher bzw. unterschiedlicher Datenquellen (Mess- und/bzw. Simulationsdaten)
2. Vergleich des gleichen Bereiches bzw. unterschiedlicher Bereiche des Beobachtungsraumes (z.B. gleiches Jahr vs. verschiedene Jahre)
3. Vergleich basierend auf gemeinsamer bzw. getrennter Clusterung der zu vergleichenden Daten

Im Falle von getrennter Clusterung kann ferner zwischen dem

4. Vergleich basierend auf gleichen oder unterschiedlichen Clusterverfahren und/oder Proximitätsmaßen

unterschieden werden. Des Weiteren sind für den Entwurf und die Auswahl von Visualisierungstechniken

5. die Art des Beobachtungsraumes, in dem die Cluster vorliegen (räumlich, zeitlich, abstrakt) und
6. die Art der zu vergleichenden Clusterinformationen (Clusterzugehörigkeit und/oder -eigenschaften).

relevant. Für die Effektivität des visuellen Vergleiches spielt weiterhin

7. die Art des bei der Visualisierung gewählten Repräsentationsraumes (ein Bild, mehrere Bilder)

eine wichtige Rolle.

Entsprechend ergeben sich eine Vielzahl von Kombinationsmöglichkeiten, aus denen eine Vielzahl von Problemstellungen für die Visualisierung resultieren. Aufgrund des erhöhten Datenaufkommens bei der gleichzeitigen Untersuchung zweier Datensätze betrifft dies im besonderen die erhöhten

Anforderungen an den **Platzverbrauch** sowie auftretende **Verdeckungen** (vgl. auch Abs. 5.4). Speziell beim Vergleich zweier verschiedener Clusterungen aufgrund der Clusterzugehörigkeit ist die **visuelle Zuordnung ähnlicher Cluster** bisher nicht gelöst. Auch beim visuellen Vergleich von Clusterzugehörigkeiten bedarf es geeigneter Methoden zur **Kodierung von Absolutwerten und Differenzen** gerade beim **Vergleich von Mess- und Vorhersagedaten** in einem Bild (vgl. auch Abs. 5.4).

Im folgenden sollen nun anhand zweier Beispiele für die Bedürfnisse der Klimaforschung zugeschnittene Lösungen zu den genannten Problemstellungen skizziert werden.

Räumlicher Vergleich von geclusterten Daten aus verschiedenen Datenquellen. Typischerweise werden zum Vergleich von Clusterungen im Umfeld der Klimaforschung die Darstellungen nebeneinander gelegt. So zeigt Abbildung 6.14 einen *Bild-zu-Bild-Vergleich* von geclusterten Messdaten und geclusterten Modellresultaten für den brasilianischen Maisdatensatz (vgl. auch Nocke u. a. 2003). Hierfür wird das aggregierte Merkmal „rang“ eingesetzt, welches für jeden Cluster die gemittelten Eigenschaften der zentralen Punkte repräsentiert.

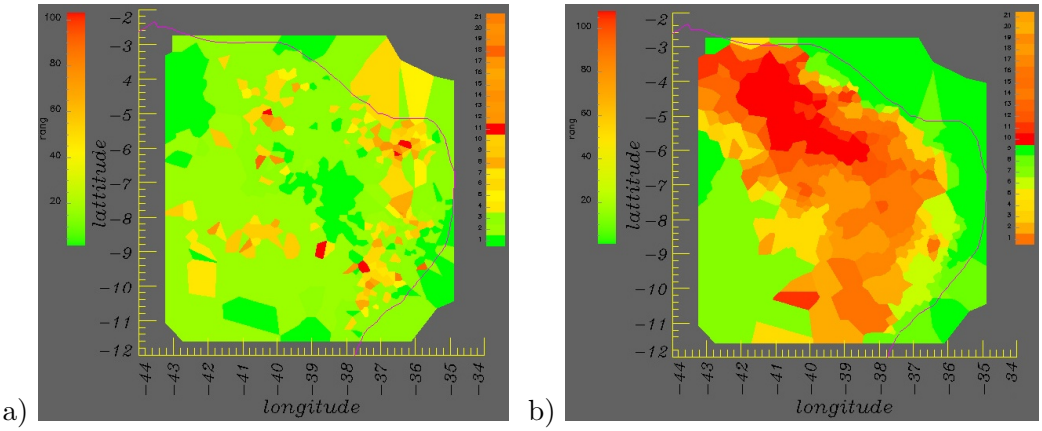


Abbildung 6.14: Vergleichende bildbasierte Clustervisualisierung des Maisdatensatzes (Nordostbrasilien, 1983, Voronoi-tessellierte Gebiete, Farbkodierung nach dem aggregierten Merkmal rang); a) geclusterte Messstationswerte; b) geclusterte Klimamodellszenarienanalyse, welche auf die Positionen der Messstationen abgebildet wurde

Der Vorteil zweier separater Darstellungen ist, dass der Anwender schnell einen guten Überblick über die räumliche Verteilung und die Heterogenität der beiden Clusterungen erhält. So sind in diesem Beispiel die Messdaten wesentlich diffuser, während die Modelldaten wesentlich homogener verteilt sind. Grundsätzlich ist zu erkennen, dass die Gebiete im Landesinneren mit der stärksten Gefahr eines Ernteverlustes und deren Stärke (rot) deutlich zwischen den beiden Clusterungen variieren, während in den Küstenregionen mit besseren Anbaubedingungen (grün) stärkere Übereinstimmungen vorliegen. Dabei zeigt sich, dass beim *image level*-Vergleich bestimmter Regionen oder sogar einzelner Messstationen ein hoher mentaler Aufwand erforderlich ist.

Um auch regionale Details und einzelne Messstationen direkt vergleichen zu können, lassen sich die beiden aggregierten Merkmale alternativ zum image level-Vergleich auch in einem Bild darstellen. Eine erste Lösung hierfür ist die *Konstruktion einfacher, rechteckiger Ikonen*, die an den Positionen der Messstationen angeordnet werden (vgl. Abb. 6.15). Hierbei werden die geclusterten Messdaten auf die untere Hälfte der Ikone und die geclusterten Modellvorhersagen auf die obere Hälfte abgebildet (visuelles Attribut Farbton). Durch die dabei auftretenden Farbdifferenzen können die beiden Datensätze an jedem Datenpunkt miteinander verglichen werden. So ergibt sich ein schneller Eindruck von Regionen guter Vorhersagen (z.B. orange Gebiete bei 40°30’ w.L. und 8°15’ s.B.) und Regionen schlechter Vorhersagen (z.B. rot-grüne Unterschiede im Nordwesten).

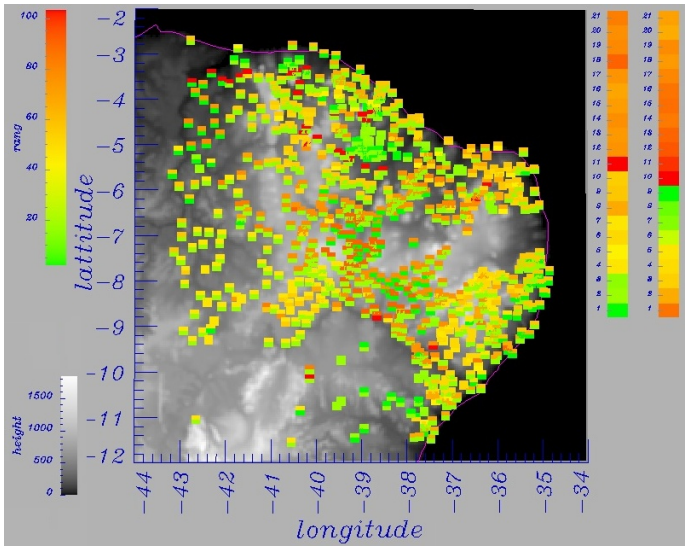


Abbildung 6.15: Vergleichende Clustervisualisierung des Maisdatensatzes mit rechteckigen Ikonen (Nordostbrasilien, 1983, Voronoi-tessellierte Gebiete, Farbkodierung nach dem aggregierten Merkmal rang); linke Clusterlegende sowie unteres Rechteck in der Ikone: geclusterte Messstationswerte; rechte Clusterlegende sowie oberes Rechteck in der Ikone: geclustertes Klimamodellszenarium

Zusätzlich können über zwei nebeneinander gelegte Legenden die Datenwerte des Merkmales „rang“ für die beiden Clusterungen verglichen werden. Hierbei erhält der Anwender einen Eindruck über die jeweilige Anzahlen von Clustern (hier beide gleich (21 Cluster)) und über deren mittlere Eigenschaften. So enthält beispielsweise die Clusterung der Messdaten (rechte Legende) mehr Cluster mit extremen Eigenschaften (dunkelrot und grün) als die Clusterung der Vorhersagedaten (linke Legende).

Bei der verwendeten Farbskala liegt der Fokus des visuellen Vergleiches auf Gebieten mit sehr großer Gefahr des Maisernteverlustes (in rot), da sich diese stark abheben. Um eine einheitlichere Vergleichsbasis für die Bewertung der Modellvorhersagen (auch in den anderen Wertebereichen) zu machen, kann hier eine alternative Farbskala (z.B. isoluminant) verwendet werden, bei welcher Farbunterschiede gleich wahrgenommen werden. Probleme dieser (farb-) ikonenbasierten Darstellung sind, dass die räumliche Verteilung der einzelnen Clusterungen schlechter wahrgenommen werden kann (vgl. Abb. 6.14), dass es beim gestreuten Layout zu Ikonenverdeckungen kommt und dass Differenzen nur abgeschätzt werden können.

Werden alternativ die Differenzen isoliert dargestellt, gehen die Informationen über die Absolutwerte der beiden zu vergleichenden Datensätze verloren (vgl. Abb. 6.16a). Um dies zu vermeiden, werden Darstellungen eingesetzt, die sowohl *Differenzwerte als auch Absolutwerte gemeinsam* repräsentieren. Insbesondere finden sich hierfür geeignete Darstellungen bei der Visualisierung von Unsicherheiten, die neben dem Datenwert auch dessen „Unsicherheit“ in leicht verständlicher Weise kommunizieren. Fast man die Differenz von geclusterten Messdaten und Vorhersagedaten auch als Unsicherheit auf, lassen sich hierfür speziell zugeschnittene Visualisierungstechniken anwenden (vgl. Abb. 6.16b,c). In diesem Beispiel wird eine der beiden Clusterungen auf den Farbton und die Differenz der Clusterungen auf die Transparenz abgebildet werden (vgl. auch Djurcilov u. a. (2001)). Abbildung 6.16b zeigt für das Beispiel des geclusterten Maisdatensatzes, wie die regionalen Unsicherheiten der Modellvorhersage untersucht werden können. So ergeben sich insbesondere in den Gebieten mit Vorhersage extremer Maisanbaubedingungen große Abweichungen zu den geclusterten Vorhersagedaten (abgeschwächte Darstellung der Datenwerte und deutliche Sichtbarkeit des unterliegenden Gitters). Hingegen konnten die Maisanbaubedingungen am Küstenstreifen und im südöstlichen Landesinneren wesentlich genauer vorhergesagt werden (starke Deckkraft der Farben). Alternativ lassen sich, falls die Absolutwerte der Messdaten im Fokus der Betrachtung stehen, auch die Messdaten farbkodieren und deren Differenzen zur Modellvorhersage durch die Transparenz abbilden (vgl. Abb. 6.16c). Diese Kopplung von Absolutwert- und Differenzdarstellung vereinfacht es, regionale Datenverteilungen zu untersuchen und zu vergleichen, und dabei explizit die Differenz (bzw. Modellunsicherheit) mit in die Visualisierung einzubeziehen.

[12] Dieses Merkmal bildet die Summe über alle normierten Elemente des zentralen Punktes, und lässt so Rückschlüsse über die gemittelten Maisanbaubedingungen zu.

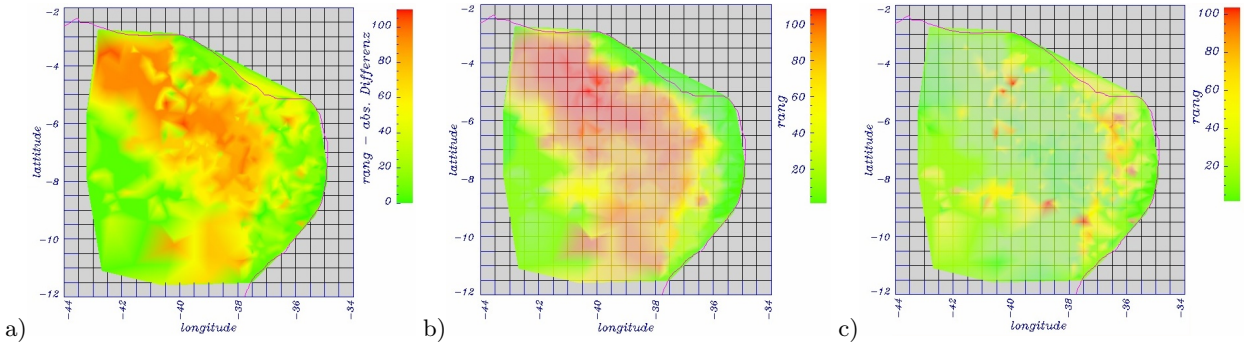


Abbildung 6.16: Vergleichende Clustervisualisierung des Maisdatensatzes (Nordostbrasilien, 1983, Delauney-Triangulation, aggregiertes Merkmal rang[12]); a) Differenzdarstellung; b) geclusterte Klimamodellvorhersagen (Farbton) sowie Differenz zur Messung (Transparenz); c) geclusterte Messdaten (Farbton) sowie Differenz zur Vorhersage (Transparenz)

Zeitlicher Vergleich von geclusterten Daten zwischen unterschiedlichen Bereichen des Beobachtungsraumes. Zum Vergleich von Clustern stellten die bisherigen räumlichen Darstellungen ein aggregiertes Merkmal für jede Clusterung dar. Dabei wird jedoch von der wesentlichen Struktur der Clusterung abstrahiert (den Clusterzugehörigkeiten). Im folgenden soll eine neue Technik[13] (eine Weiterentwicklung der Kalender-Cluster-Visualisierung von van Wijk u. van Selow (1999)) vorgestellt werden, welche es erlaubt, geclusterte Tagestemperaturdaten zweier Jahre unter Einbezug der Clusterzugehörigkeiten sowie der Clustereigenschaften zu vergleichen.

Um die Tagestemperaturverläufe zweier Jahre zu strukturieren, bietet es sich an, die *Tage der beiden Jahre in einer Clusterung zu gruppieren.* Um deren Ergebnisse zu visualisieren, wurde die in Abbildung 6.13 vorgestellte Darstellung der geclusterten Tage eines Jahres so erweitert, dass in der Kalenderdarstellung die einzelnen Monate der zu vergleichenden Jahre direkt untereinander (in zwei aneinander liegenden Spalten) angeordnet wurden (vgl. 6.17). Bei der Darstellung der Tagesverläufe (Abbildung 6.17 rechts) stehen dem Anwender weiterhin alle Möglichkeiten der Ursprungstechnik wie Veränderung der Linienstile und Auswahl von Clustern, Monaten und Tagen zur Verfügung. So wurden in Abbildung 6.17 die Cluster 1 (rot) und 3 (blau) ausgewählt, deren zentrale Punkte

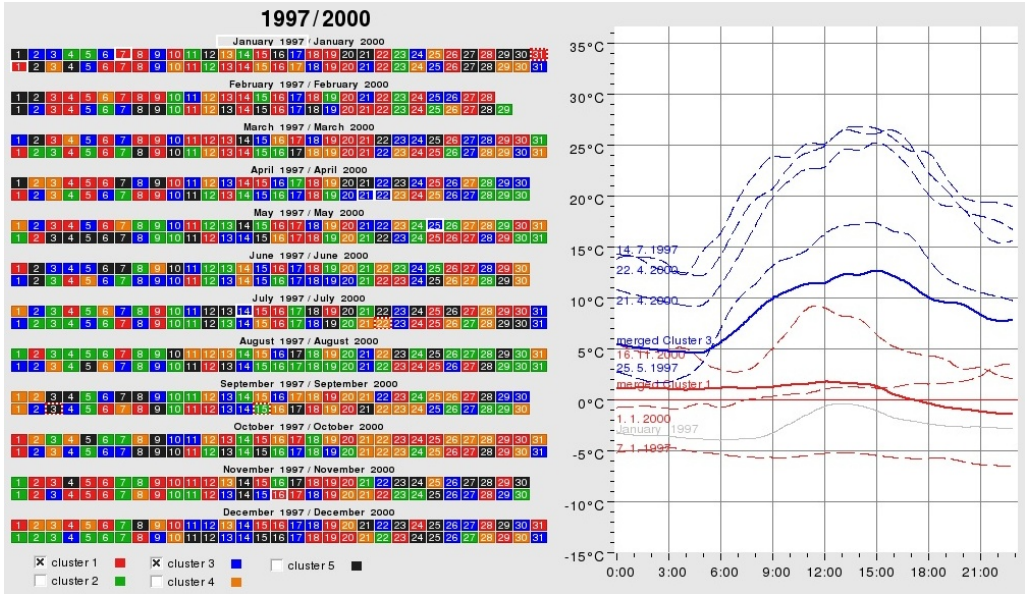


Abbildung 6.17: Vergleichende kalenderbasierte Clustervisualisierung für Tagestemperaturverläufe über zwei Jahre (1997, 2000) basierend auf stündlichen Messungen (Potsdam)

[13]Die Umsetzung erfolgte im Rahmen eines betreuten Studentenprojektes (Kaeding u. Walter 2004).

hervorgehoben und zugehörige Tage in gestrichelten Linien angedeutet, sowie der mittlere Temperaturverlauf eines Monats (Januar) eingezeichnet.

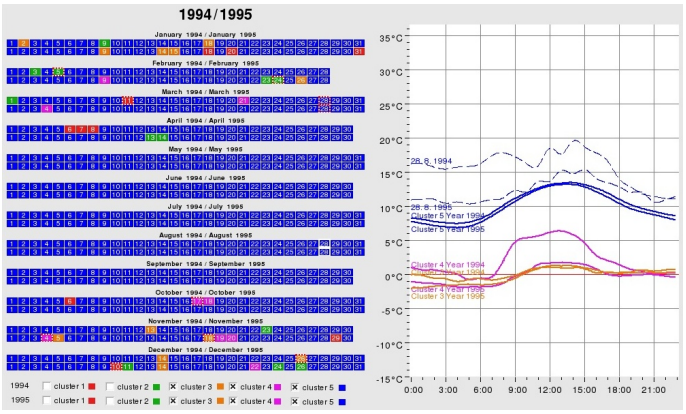


Abbildung 6.18: Vergleichende kalenderbasierte Clustervisualisierung für Tagestemperaturverläufe zweier separat geclusterter Jahre (1994, 1995)

Prinzipiell lassen sich mit der vorgestellten Darstellungstechnik auch *zwei separat geclusterte Jahre vergleichen* (vgl. Abb. 6.18). Hierbei besteht das Problem, dass zwar die durch die Verteilung der Cluster entstehende Struktur repräsentiert wird, aber ähnliche Cluster in beiden Clusterungen farblich abweichend dargestellt werden können. So kann sogar der Eindruck erweckt werden, dass aufgrund der Verwendung gleicher Farben bestimmte Tage ähnliche Eigenschaften haben, während die zugehörigen Cluster stark variieren. Als erste Lösung für dieses Problem kann der Anwender in der umgesetzten Technik interaktiv die einzelnen Cluster separat auswählen und deren zentrale Punkte auch zwischen den beiden Clusterungen vergleichen. So haben in Abbildung 6.18 die Cluster 3 (orange) und 5 (blau) ähnliche zentrale Punkte, während die Cluster 4 (violett) starke Abweichungen aufweisen.

Problem der alleinigen Darstellung der Clusterzugehörigkeiten im Kalender (Abb. 6.17 und 6.18) ist, dass die Ähnlichkeit der Cluster bei der Farbgebung keine Berücksichtigung findet und der Anwender so in den Jahresverläufen auftretende Strukturen nicht identifizieren und vergleichen kann. Deswegen wurde - für den speziellen Fall der Clusterung eines Jahres mit verschiedenen Clusterverfahren - eine *alternative Farbabbildung* entworfen, die es ermöglicht, *ähnliche Cluster ähnlich einzufärben*. Hierbei werden die Farben der Cluster der zweiten Clusterung anhand der Anzahl übereinstimmender Tage an die Farben der ersten Clusterung angepasst, und ggf. dabei auch gemischt (vgl. Abb. 6.19). Dabei werden von der originalen Clustereinfärbung ausgehend (Abb. 6.19 links und Mitte), im rechten, farbangepassten Bild insbesondere die Cluster 5, 6 und 7 der zweiten Clusterung (jeweils untere Zeile) in ihrer Farbgebung deutlich verändert, so dass

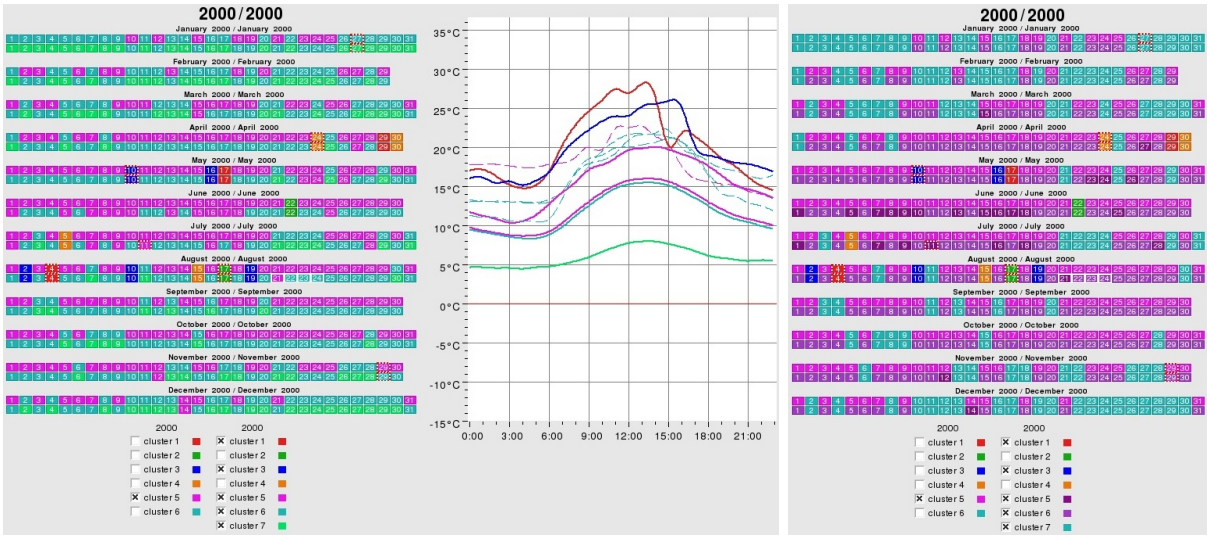


Abbildung 6.19: Vergleichende kalenderbasierte Cluster-Visualisierung für Tagestemperaturverläufe zweier Clusterungen des Jahres 2000 basierend auf stündlichen Messungen (Potsdam); Standardeinfärbung der Cluster (links und Mitte) sowie an die Clusterähnlichkeit angepasste Einfärbung (rechts)

Strukturen beider Clusterungen einander zugeordnet werden können.

Tabelle 6.1 ordnet die vorgestellten vergleichenden Clusterdarstellungen den am Anfang des Abschnitts aufgestellten Kriterien zu. Dadurch wird verdeutlicht, welche Kombinationen sinnvoll miteinander kombinierbar sind und welche nicht. So werden in den durch Fragezeichen gekennzeichneten Bereichen neben den Clusterverfahren auch die zugrunde liegenden Daten variiert, was deren Vergeichbarkeit wesentlich erschwert. Ferner bleibt die Darstellung von geclusterten Daten aus unterschiedlichen Datenquellen noch weitgehend offener Forschungsgegenstand (vgl. Tab. 6.1 rechts unten).

		gleiche Datenquelle		unterschiedliche Datenquellen	
		gleicher Bereich	untersch. Bereiche	gleicher Bereich	untersch. Bereiche
getrennte Clusterung	untersch. Cl.-Verf.	zeitlich, Cl.-zugeh. u. -eigensch., ein Bild (Abb. 6.19)	?	?	?
	gleiches Cl.-Verf.	X	zeitlich, Cl.-zugeh., ein Bild (Abb. 6.18)	räumlich, Cl.-eigensch. (ein Bild in Abb. 6.14, mehrere Bilder in Abb. 6.15 u. 6.16)	
gemeinsame Clusterung		X	zeitlich, Cl.-zugeh., ein Bild (Abb. 6.17)		

Tabelle 6.1: Einordnung der Darstellungen zum Vergleich von Clusterungen (Abkürzungen: *Cl.-Verf.* - Clusterverfahren, *Cl.-zugeh.* - Clusterzugehörigkeit, *Cl.-eigensch.* - Clustereigenschaften)

6.1.6 Clusterung durch spezielle Anordnungen

Klassische Clusterverfahren erzeugen Mengen ähnlicher Objekte, wobei ein Objekt genau einem Cluster zugeordnet wird. Hierbei ist es auf den ersten Blick nicht immer klar, wie „knapp“ diese Entscheidung war und wie homogen oder heterogen die entstandenen Cluster sind. Daneben haben sich Anordungsalgorithmen etabliert, die ähnliche Objekte in räumlicher Nähe anordnen[14]. Hierzu gehören Multidimensional Scaling Verfahren (MDS, vgl. z.B. Kruskal u. Wish 1978), Federkraftmodelle (vgl. z.B. Pfefferer 1996) und neuronale Netze (vgl. z.B. SOMs in Kohonen 1997). Bei der Visualisierung solcher Anordnungen bilden sich Cluster heraus, wobei deren Struktur intuitiv aufgenommen werden kann, ohne dass einzelne Objekte explizit einem Cluster zugeordnet werden müssen.

Typischerweise werden solche Anordnungsalgorithmen eingesetzt, um Datenobjekte (z.B. Dokumente) zu strukturieren und in der Visualisierung auszugeben (vgl. z.B. Wise u. a. 1995). Darüber hinaus gibt es erste Ansätze in der Literatur, auch die Struktur der beteiligten Merkmale bei hochdimensionalen Datensätzen gesondert zu visualisieren. So stellen Yang u. a. (2004) und Yang u. a. (2005) die Merkmalsstruktur mit in der Darstellung angeordneten Rechtecken basierend auf einem MDS-Verfahren dar.

Eine solche Art der Darstellung hat auch Potential bei der Darstellung komplexer Klimasimulationen, da diese häufig mehr als hundert Modellresultate (Merkmale) enthalten. Typischerweise werden hierbei einzelne Merkmale separat (im Beobachtungsraum) dargestellt, wobei ein Überblick über deren Struktur nur sehr begrenzt gegeben ist. Im Rahmen dieser Arbeit wurde als ein erster Ansatz zur Darstellung der Variablenstruktur von Klimamodellen eine Variablenanordnung basierend auf

[14] Dazu bilden sie den n-dimensionalen Informationsraum auf einen zwei- oder dreidimensionalen Raum derart ab, dass die Ähnlichkeit von je zwei Objekten im Informationsraum auch im projizierten Raum weitgehend erhalten bleibt.

selbstorganisierenden Netzen untersucht (vgl. SOM-Packet[15] von Kohonen (1997)). Hierfür wurde eine Darstellungstechnik entworfen und umgesetzt, die Modellresultate eines Klimamodells auf einer Karte anordnet, diese beschriftet sowie die Eigenschaften der Kohonenkarte mit ausgibt (vgl. Abb. 6.20).

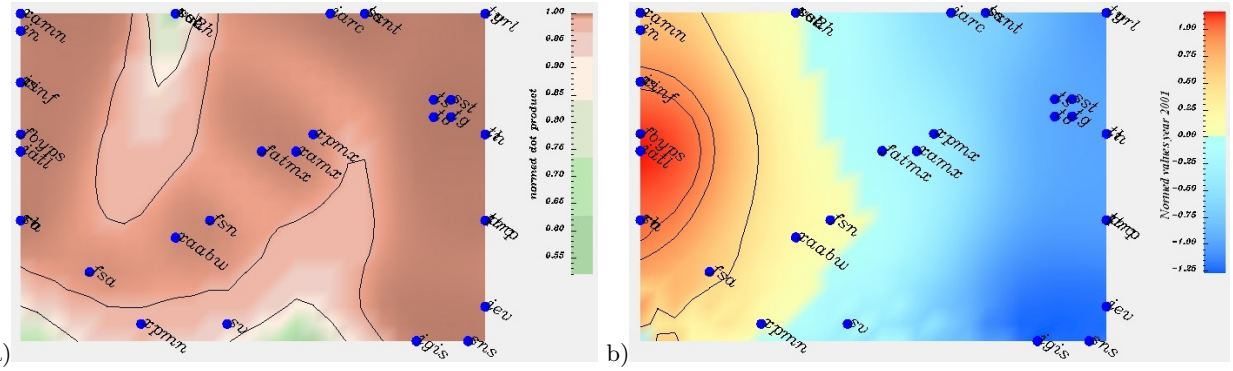


Abbildung 6.20: Strukturierung von Variablen eines Klimamodelllaufes basierend auf selbstorganisieren Netzen (SOMs) basierend auf aggregierten Variablen eines Climber-Modelllaufes; a) Ähnlichkeitsdarstellung benachbarter Neuronen basierend auf dem Skalarprodukt der Vektoren; b) normierte Wertedarstellung für den ersten Zeitschritt (Jahr 2001)

Hierbei wurden 35 aggregierte Merkmale (u.a. globale Oberflächentemperatur (tg) und globaler Niederschlag (prc)) aus einen Modellauf von 200 Zeitschritten normiert (Mittelwert, Standardabweichung), durch das SOM angeordnet und dargestellt. Abbildung 6.20a zeigt die Merkmale (blaue Kreise) auf einer Karte, die nach normierten Skalarprodukten benachbarter Neuronen (bzw. Gitterzellen) eingefärbt wurde. Durch Verwendung einer grün-braun Farbskala (Berg-und-Tal-Metapher) kann der Anwender so leicht homogene Gebiete (braun) von Gebieten mit starker Änderung der Vektoren (grün) des zugrunde liegenden neuronalen Netzwerkes unterscheiden. Dies ermöglicht auftretende Clusterungen von Merkmalen und deren Homogenität einzuschätzen, und dabei die anderen Merkmale im Blick zu behalten.

Um zusätzlich die Eigenschaften der durch die Kohonenkarte erzeugten Clusterung im Detail zu untersuchen, können einzelne Vektorelemente auf einer separaten Karte ausgegeben werden (vgl. Abb. 6.20b). Jede dieser Vektorelemente repräsentiert genau einen Punkt im Beobachtungsraum (in diesem Fall ein Jahr). Unter Verwendung einer Farbskala mit einem Nullpunkt kann der Anwender schnell identifizieren, ob ein bestimmtes Merkmal einen Wert über oder unter seinem Mittelwert aufweist und dabei auftretende Extreme identifizieren. Durch Darstellung der verschiedenen Vektorelemente erhält der Anwender so neben einem Überblick zur Verteilung der Datenwerte der einzelnen Merkmale im Beobachtungsraum (hier über die Zeit) auch ein Gefühl für das Zustandekommen der Kohonenkarte und der Anordnung der Merkmale (vgl. Abb. 6.21). Da bei der Abbildung des 200-dimensionalen auf den 2-dimensionalen Raum die Vektorelemente des neuronalen Netzwerkes an den Positionen der Merkmale nur eine Annäherung der zugrunde liegenden Datenwerte sind, können lediglich allgemeine Aussagen über auftretende Datenwerte getroffen werden. So lässt sich aus Abbildung 6.21 ablesen, dass die Simulation mit Werten nahe dem Mittelwert der einzelnen Merkmale startet (Jahr 2001, Abb. 6.21 links), sich zum Ende der Simulation eine Häufung extremer Werte ergibt (Jahr 2195, Abb. 6.21 Mitte) sowie bezüglich einer Vielzahl von Variablen in den letzten fünf Jahren noch einmal einer starken Schwankung unterworfen ist (Unterschied zwischen Jahr 2195 (Abb. 6.21 Mitte) und 2200 (Abb. 6.21 rechts)).

Abschließend bleibt festzustellen, dass mit den vorgestellten Techniken eine effektive Überblicksdarstellung für die generelle Struktur komplexer Simulationsdaten bereitgestellt werden konnte.

[15]Im Rahmen eines betreuten Studentenprojektes (Scholtz 2006) erfolgte eine Einbindung des SOM-Moduls in das Visualisierungssystem OpenDX.

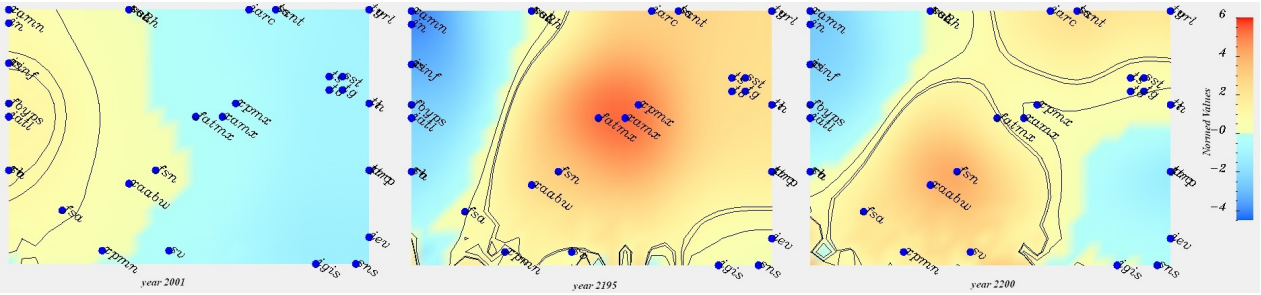


Abbildung 6.21: Strukturierung von Variablen eines Klimamodelllaufes basierend auf Selbstorganisieren Netzen (SOMs) basierend auf aggregierten Variablen eines Climber-Modelllaufes; Vergleich der normierten Wertedarstellungen der Jahre 2001, 2195 und 2200

6.1.7 Einsatz von Clusterungen zur Parametrisierung der Visualisierung

Neben der direkten Darstellung von Clustern und Clustereigenschaften lassen sich Clusterungen auch einsetzen, um Visualisierungstechniken zu parametrisieren. Dabei wird der Fakt ausgenutzt, dass bei der Clusterung eine Abstraktion auf den Daten aufgebaut wird. Zu den Einsatzmöglichkeiten zählen:

1. *Visuelle Hervorhebung, Abschwächung oder Verstecken* von Datenobjekten nach ihrer Clusterzugehörigkeit (s. auch S. 100), z.B. von Ausreißern[16] oder von Datenobjekten in unübersichtlichen Parallele Koordinaten-Darstellungen durch Einfärbung von Linienzügen (vgl. z.B. Johansson u. a. 2004)
2. *Zusammenfassen* von visuellen Primitiven im Darstellungsraum bei gleicher Clusterzugehörigkeit (vgl. Abb. 6.22)
3. *Gruppierung* von Datenobjekten zur Ordnung *von Datenachsen*
4. *Unterstützung von Details-on-Demand* durch flexible Auswahl von Stufen aus hierarchischen Clusterungen (vgl. Abb. 6.19 links und rechts mit abweichenden Hierarchiestufen)

Exemplarisch sollen hier zwei Beispiele für das Zusammenfassen von visuellen Primitiven (Punkt 2) im räumlichen Bezug gegeben werden. So können Ikonen zusammengefasst werden, wenn alle zugehörigen Messstationen in den selben Cluster fallen und entsprechend ähnliche Eigenschaften aufweisen (vgl. Abb. 6.22). Dies ermöglicht, große homogene Regionen schneller zu identifizieren und lokale Datenverteilungen besser zu verstehen.

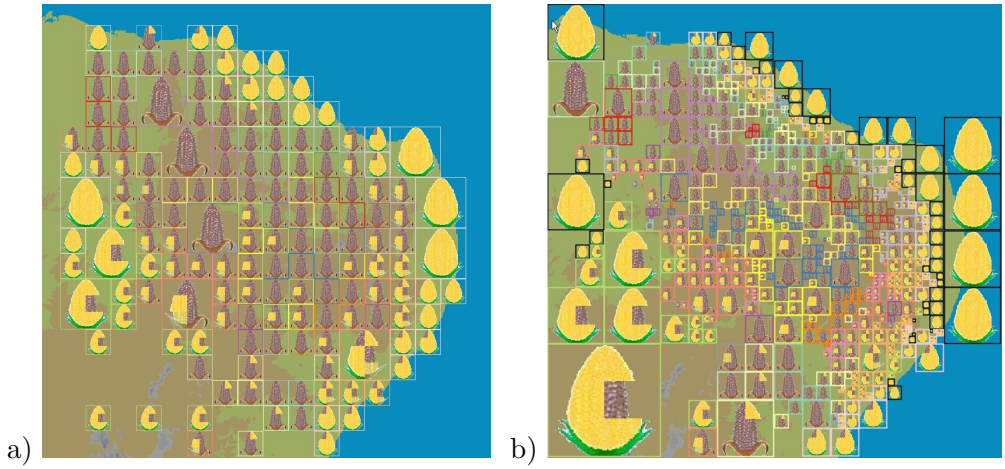


Abbildung 6.22: Metapherbasierte Ikonendarstellungen für gestreute, geclusterte 2D-Klimadaten mit clusterbasierter Ikonenzusammenfassung; a) Reguläres Layout b) Multi-resolution Layout

[16] z.B. über das Single-Linkage-Verfahren bestimmbar

So werden beim regulären Layout benachbarte Gitterzellen - die (überwiegend) Messstationen mit dem selben Cluster enthalten - zu größeren quadratischen Blocks (2×2, 3×3, ...) zusammengefasst, und in diesen Regionen durch eine größere Ikone ersetzt (vgl. Abb. 6.22a). Weiterhin kann - alternativ zur Standard-Multiresolution-Unterteilung, welche erst abbricht, wenn nur noch eine Messstation in einer Gitterzelle liegt (vgl. Abb. 5.11c und 6.12c) - die Unterteilung abgebrochen werden, wenn dort nur noch ein Cluster vorliegt. Entsprechend wird der gesamte Bereich des Quadtree durch eine Ikone ausgefüllt (vgl. Abb. 6.22b).

6.1.8 Diskussion

In diesem Abschnitt wurde die Kombination von Clusteranalyse und Visualisierung systematisch untersucht. Diese enge Kopplung eröffnet neue Möglichkeiten bei der Analyse großer Datenmengen. Insbesondere wurden vielfältige Darstellungen von Clusterungen im räumlichen und zeitlichen Bezug am Beispiel von Klimadatensätzen vorgestellt. Um dabei das Verständnis für die Clusterung und deren Zustandekommen zu verbessern, wurden die Clusterzugehörigkeiten und die Clustereigenschaften insbesondere auch gemeinsam abgebildet.

Besonders hervorzuheben sind die neu entwickelten metapherbasierten Ikonen und deren Layouts sowie die neuen Ansätze zum Vergleich von Clusterungen. Darüber hinaus wurden neue Vorgehensweisen zur Farbkodierung von disjunkten Clusterungen und zur Vereinfachung binärer Dendrogramme vorgestellt.

Neben den klassischen Clusterverfahren und der daran geknüpften Darstellung von deren Ergebnissen ergeben sich spezielle Anforderungen, Clusterungen visuell zu kommunizieren. Hierzu lassen sich insbesondere Anordungstechniken (z.B. SOMs) und deren Visualisierung einsetzen. Jedoch verbleiben hier noch immer Herausforderungen, die Ergebnisse von Clusterfahren und Objektstrukturen leicht verständlich darzustellen. Auch die konsequente Zusammenführung von zum Teil separaten Darstellungen bleibt zukünftigen Arbeiten vorbehalten.

6.2 Visualisierung und Hauptkomponentenanalyse auf Klimadaten

Gerade bei sehr großen, multivariaten Datenmengen, wie sie auch im Umfeld der Klimaforschung auftreten, ist die Untersuchung von Korrelationen, Trends und Ausreißern nicht trivial. Eine etablierte Methode, den Informationsraum so zu transformieren, dass die Beantwortung der genannten Fragestellungen vereinfacht wird, ist die Hauptkomponentenanalyse (kurz PCA, engl.: Principal Component Analysis). Die PCA führt eine Transformation des Koordinatensystems derart durch, dass die Achsen (senkrecht aufeinander stehend) in Richtung der größten Varianzen in den Daten ausgerichtet werden. Die entstehenden Achsen werden auch als Hauptkomponenten bezeichnet. Die Hauptkomponenten geben die grundlegende Struktur zwischen den Variablen wieder, und ermöglichen so eine komprimierte Beschreibung von Korrelationen in den Daten, ein besseres Verständnis der zugrunde liegenden Muster sowie die Extraktion allgemeiner Trends.

Die Kombination von PCA und Visualisierungstechniken wird in vielen Anwendungsdomänen zur Analyse von Datensätzen mit vielen Variablen eingesetzt (vgl. z.B. Landgrebe u. a. 2002; Yang u. a. 2003; Komura u. a. 2004). Allerdings wird die PCA dabei fast ausschließlich als Vorverarbeitung eingesetzt, um signifikante Trends in den Daten zu berechnen und diese dann zu visualisieren. Hierbei ergeben sich die zwei folgenden Probleme:

1. Die PCA erzeugt Hauptkomponenten die teilweise schwer zu interpretieren sind und manchmal auch orthogonal zu dem liegen können, was intuitiv der dominante Trend in den Daten zu sein

scheint. Zur Lösung des letzteren Problems kann entweder eine automatische Rotation der Komponenten vorgenommen (vgl. Joliffe 1986) oder diese interaktiv angepasst werden (vgl. Müller u. Alexa 2004).

2. Die Koordinatentransformation bei der Überführung in den durch die Hauptkomponenten aufgespannten Raum erschwert es für den Anwender die identifizierten Trends mit den originalen Variablen in Beziehung zu setzen.

Im Rahmen der vorliegenden Arbeit wurde deshalb ein Ansatz entwickelt (vgl. Müller u. a. 2006), um diese Probleme durch eine verstärkte Integration der PCA in den Visualisierungsprozess zu reduzieren . Dazu soll im folgenden untersucht werden, wie die PCA die verschiedenen Schritte der Visualisierungspipeline unterstützen kann, um effektivere Darstellungen zu generieren, und dadurch das Verständnis der bei der PCA generierten Werte sowie der originalen Datenwerte zu verbessern.

6.2.1 Hintergrund, Problemstellungen und Lösungsansätze

Grundlagen der Hauptkomponentenanalyse. Formal können die Originaldaten in Matrixnotation als

$$Y = \begin{pmatrix} y_{11} & y_{21} & \cdots & y_{d1} \\ y_{12} & y_{22} & \cdots & y_{d2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1n} & y_{2n} & \cdots & y_{dn} \end{pmatrix} \tag{6.1}$$

beschrieben werden, wobei jede Spalte eine Variable bestimmt, welche jeweils eine Dimension der d-dimensionalen Ursprungsraumes repräsentiert. Die Spalten repräsentieren die einzelnen Datenobjekte (z.B. einen Zeitschritt einer Zeitreihe). Typischerweise müssen die Datenwerte y_{ij} der Matrix vor Ausführung der PCA normalisiert werden.

Bei der PCA werden die Achsen des originalen, von den Variablen aufgespannten Raumes durch eine Rotation so angepasst, dass sie anschließend in Richtung der maximalen Varianzen im Datenraum zeigen. Dabei repräsentiert die erste Achse des resultierenden Hauptkomponentenraumes die meiste Varianz in den Daten, die zweite Variable die meiste verbleibende Varianz und so weiter. Grundsätzlich existieren verschiedene Verfahren zur Bestimmung der Hauptkomponenten. So liegt den folgenden Betrachtungen die Singulärwertzerlegung (SVD, engl.: Singular Value Decomposition) zugrunde. Die Singulärwertzerlegung faktorisiert Y direkt in

$$Y = C\ \Lambda^2\ W \tag{6.2}$$

wobei die Matrix W die *Loadings* repräsentiert, welche aus den Basisvektoren w_i des PCA-Raumes aufgebaut ist. Λ^2 ist die Diagonalmatrix der Eigenwerte, welche die Signifikanzen jeder Hauptkomponente in Bezug auf der durch sie repräsentierten Varianz beschreibt. Die Matrix C enthält die Koordinaten c_i der Datenobjekte im transformierten PCA-Raum. Diese Koordinaten werden auch als *Scores* bezeichnet.

Die bei Ausführung der PCA generierten Informationen bieten eine Vielzahl von Einsatzmöglichkeiten in Kombination mit der Datenvisualisierung. Insbesondere können die Hauptkomponenten w_i direkt als wichtige Trends in den Daten interpretiert und dargestellt werden. Weiterhin beinhaltet die Matrix W Informationen zur Korrelation der verschiedenen Variablen mit diesen Trends. Außerdem können zur Datenreduktion die (gemäß der Matrix Λ^2) weniger signifikanten Achsen bei der weiteren Analyse ausgeschlossen werden.

Probleme und Einsatzmöglichkeiten der PCA in den einzelnen Schritten des Visualisierungsprozesses. Bei der Untersuchung der Kombination von PCA und Visualisierung soll die erweiterte Visualisierungspipeline nach dos Santos u. Brodlie (2004) (Datenanalyse, Filterung,

Mapping und Rendering) zugrunde gelegt werden. Hierbei konzentriert sich die *Datenanalyse* auf automatische Berechnungen, um die Daten zu analysieren oder zu erweitern. Klassischerweise wird die PCA ausschließlich in diesem Schritt betrachtet. Das *Filtering* ist vorwiegend nutzergesteuert und stellt Funktionalität zur Auswahl von Daten von Interesse sowie deren Extraktion bereit. Hierbei können die Datenwerte von Interesse eingeschränkt sowie eine Fokusregion definiert werden (vgl. dos Santos u. Brodlie 2004).

Allerdings muss eine starre Reihenfolge dieser beiden Prozesse - zuerst die *Datenanalyse* und anschließend interaktives, iterativ wiederholbares *Filtering* - wie sie dos Santos u. Brodlie (2004) vorschlagen, nicht immer zweckmäßig sein. So kann es zum Beispiel sinnvoll sein, die PCA nach dem Filtering durchzuführen, um bestimmte Variablen oder Ausreißer auszuschließen. Auch kann, wenn die PCA auf alle Variablen des Datensatzes angewendet wird, eine Interpretation deren Ergebnisse erschwert werden. So werden Raumdimensionen oder die Zeit typischerweise vorher ausgeschlossen, da diese ansonsten auch nicht mehr als Bezugskoordinaten für die Visualisierung zur Verfügung stehen.

Im *Mapping*, dem nächsten Schritt der Visualisierungspipeline, werden die berechneten und/oder selektierten Daten in eine geometrische Repräsentation unter Einbeziehung der definierten Fokusregion überführt. Bisherige Ansätze zum Visualisierungsdesign (z.B. Mackinlay 1986; Roth u. a. 1996) beziehen die PCA nicht für die Definition und Parametrisierung dieses maßgeblich über die Expressivität und Effektivität der Darstellung entscheidenden Schrittes mit ein. Um diese Lücke zu schließen, wird in Abschnitt 6.2.2.3 diskutiert, wie die Ergebnisse der PCA (insbesondere die Loadings w_i und die Eigenwerte Λ^2) für die Erzeugung intuitiver Repräsentationen eingesetzt werden können.

Im *Rendering*, dem letzten Schnitt der Visualisierungspipeline, werden aus den geometrischen Repräsentationen Bilder generiert. Auch hier lohnt es sich, durch eine Kombination aus Daten und PCA-Werten die Darstellung zu verbessern. So erfordert das *Rendering* der transformierten Daten eine mentale Abbildung des Hauptkomponentenraumes auf den originalen Variablenraum. Um dies zu unterstützen ist eine adäquate Beschriftung der resultierenden Hauptkomponenten-Achsen erforderlich. Hierfür muss bestimmt werden, wie stark bestimmte Variablen mit bestimmten Hauptkomponenten korrespondieren. Da bisherige Ansätze hier einen Mangel aufweisen wird dem Beschriftungsproblem in Abschnitt 6.2.2.4 besondere Beachtung geschenkt.

6.2.2 Integration der Hauptkomponentenanalyse in den Visualisierungsprozess

6.2.2.1 Datenanalye und Hauptkomponentenanalyse

Typischerweise wird die PCA im ersten Schritt der Visualisierungspipeline eingesetzt. So kann der untersuchte PCA-Raum durch Reduktion auf wenige signifikante Hauptkomponenten oder auch der Raum aufgespannt aus den Originalvariablen eingeschränkt werden.

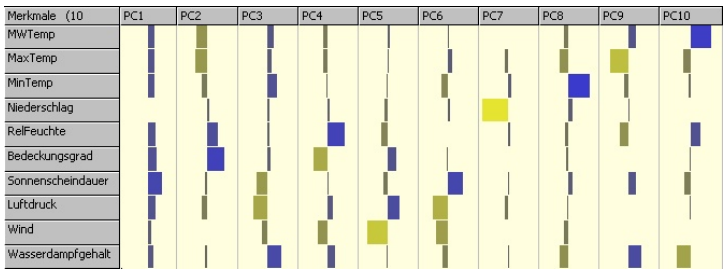


Abbildung 6.23: Visualisierung der *Loadings*-Matrix W für einen Datensatz der Tageswerte aller Maitage von 1893-2003 der Station Potsdam; 10 Merkmale

Darüber hinaus ist es für die weitere Analyse wertvoll, bereits in diesem Schritt das Verständnis der Ergebnisse der PCA zu verbessern, was wiederum tiefere Einsichten in die Muster der Originaldaten ermöglicht. Deshalb soll vorgeschlagen werden, bereits direkt nach der Berechnung der PCA ein erstes Bild zur Darstellung der *Loadings* in der Matrix W bereitgestellt werden. Abbildung 6.23 zeigt eine speziell für die Repräsen-

tation der *Loadings* weiterentwickelte Tabellendarstellung (Data Table View, vgl. Kreuseler u. Schumann 2002a) am Beispiel eines Datensatzes der Tageswerte aller Maitage der Jahre 1893 bis 2003. Hierbei werden *Loadings*-Werte $w_{ij} < 0$ auf gelbe Rechtecke im linken Teil einer Tabellenzelle, *Loadings*-Werte $w_{ij} > 0$ auf blaue Rechtecke im rechten Teil sowie Nullwerte durch das Fehlen eines Rechtecks abgebildet.

Hohe *Loadings*-Werte (größere Rechtecke mit größerer Farbsättigung) in einer Spalte zeigen die Relevanzen der jeweiligen Merkmale zu dem durch die Hauptkomponente repräsentierten Trend, während die Zeilen den Einfluss einzelner Variablen auf die durch die Hauptkomponenten repräsentierten Trends wiedergeben. So zeigt Abbildung 6.23 unter anderem den Haupttrend, welcher durch die erste Hauptkomponente repräsentiert wird (PC1). Hierbei indizieren alle positiven Loadings (in blau) in der Hauptkomponente PC1 eine direkte Proportionalität der Merkmale Temperaturmittelwert, Temperaturmaximum, Temperaturminimum, relative Luftfeuchtigkeit, Bedeckungsgrad, Sonnenscheindauer, Luftdruck, Windstärke und Wasserdampfgehalt an. Der Niederschlag hat keinen Einfluss auf diesen Trend. Der zweite Trend (PC2) wird durch die Merkmale Bedeckungsgrad und relative Luftfeuchte bestimmt, und ist umgekehrt proportional zu den drei Temperaturmerkmalen und dem Luftdruck.

Dieses Beispiel illustriert, wie der Anwender eine schnellen Einblick in wichtige Trends in Datensätzen mit einer Vielzahl von Variablen[17] bekommen kann. Darüber hinaus können jedoch auch versteckte Trends identifiziert werden (so genannte „Ausreißertrends“). Ein Beispiel für einen solchen „Ausreißertrend“ sind die gegensätzlichen Loadings von relativer Luftfeuchtigkeit und dem Bedeckungsgrad in Hauptkomponente PC4.

Das so gewonnene bessere Verständnis über die Ergebnisse der PCA kann benutzt werden, um die folgenden Schritte der Visualisierungspipeline zu steuern und expressive Darstellungen zu generieren. Weiter ist damit eine Basis gelegt um sowohl Überblicksbilder als auch Detaildarstellungen zu erzeugen. So kann der Anwender, um z.B. die wichtigsten Trends näher zu untersuchen, basierend auf der Loadings-Darstellung zu weiteren Darstellungen wie Scatterplotmatrizen wechseln (vgl. Abs. 6.2.2.3).

6.2.2.2 Filterung und Hauptkomponentenanalyse

Im zweiten Schritt der Visualisierungspipeline, dem Filtering, werden die Variablen von Interesse ausgewählt. Dieser Prozess ist hochgradig interaktiv und nutzerzentriert. Unter Berücksichtigung der PCA lassen sich hierbei zwei Strategien unterscheiden:

1. Filtering vor der Ausführung der PCA oder
2. Filtering nach der Ausführung der PCA.

Für die *Filterung vor der Ausführung der PCA* muss entschieden werden, welche Variablen und Datenobjekte (Zeilen bzw. Spalten der Matrix Y) für die anschließende Verarbeitung in der PCA ausgewählt werden sollen. Typischerweise werden zum einen die räumlichen Dimensionen und die Zeit von der PCA ausgenommen um, um die PCA-transformierten Daten in ihrem räumlichen oder zeitlichen Bezug visualisieren zu könne. Zum anderen lassen sich Merkmale ausschließen, die in den Daten vorhandene Trends künstlich verstärken (z.B. beim Vorhandensein mehrerer (stark korrelierter) Temperaturwerte). Zur Identifikation solcher Merkmale aus den Originaldaten kann auch die Visualisierung eingesetzt werden (z.B. Scatterplotmatrizen, Parallele Koordinaten oder Tabellendarstellungen (vgl. auch Müller u. a. 2006)). Um zu verhindern, dass bestimmte durch die PCA berechnete Trends mehr betont werden, als sie im natürlichen Phänomen vorliegen, obliegt es dem Anwender zu entscheiden, ob solche Merkmale aus der PCA ausgeschlossen werden sollen. So können

[17] In diesem Beispiel wurden nur die abhängigen Variablen (die Merkmale) in die PCA einbezogen, während die Zeit als unabhängige Variable herausgenommen wurde.

im Datensatz der Maitagesdaten (aus Abb. 6.23) beispielsweise aus den drei Temperaturmerkmalen eines ausgewählt werden, da diese stark voneinander abhängen.

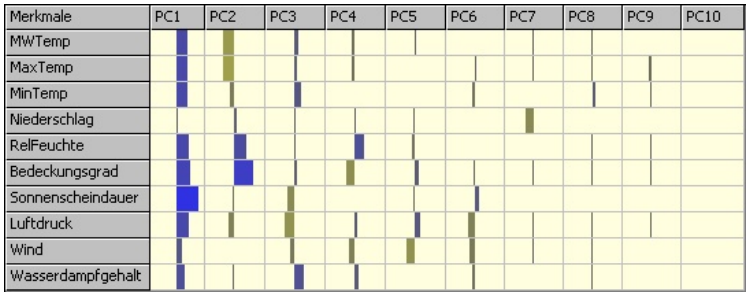


Abbildung 6.24: Visualisierung der durch die Signifikanzen (Matrix λ^2) normierten *Loadings*-Matrix W für einen Datensatz der Tageswerte aller Maitage von 1893-2003 der Station Potsdam; 10 Merkmale

Die *Filterung nach der Ausführung der PCA* hat ein großes Potential, die Vielzahl vorliegender Trends im Sinne des Überblick und Detail zu explorieren. So wird durch Einsatz einer tabellarischen *Loadings*-Darstellung (vgl. Abb. 6.23) einen Überblick über die in den Daten vorliegenden Trends und die zugehörigen Variablen gegeben. Basierend darauf kann er Hauptkomponenten, Originalvariablen oder Kombinationen von beiden auswählen. So können dann bestimmte Trends und der Beitrag bestimmter Variablen im Detail untersucht werden (vgl. Abs. 6.2.2.3).

Des Weiteren können auch die durch die PCA generierten Signifikanzinformationen (Matrix λ^2) zur Sortierung[18] und quantitativen Bewertung der Trends herangezogen werden. Hierdurch kann die Aufmerksamkeit auf die wichtigsten Trends gelenkt werden (vgl. Abbildung 6.24 mit einer Tabellendarstellung der normalisierten *Loadings*). Der erste Trend (PC1) hat eine hohe Signifikanz $(99,9)$ und betrifft nahezu alle Variablen. Auch der zweite Trend (PC2) hat noch eine hohe Signifikanz $(20,4)$. Diese beiden repräsentieren die meiste Varianz des Datensatzes, während die normalisierten Loading-Werte für die nächsten relevanten Hauptkomponenten bereits stark abfallen. Dieses Wissen ermöglicht, einen 2-dimensionalen Unterraum des Originalraumes auszuwählen und die Untersuchung auf die wichtigsten Trends des Datensatzes konzentrieren. So kann z.B. ein einfacher Zeitgraph (vgl. Abb. 5.12) die den beiden Hauptkomponenten zugehörigen *Scores* im Detail darstellen.

6.2.2.3 Mapping und Hauptkomponentenanalyse

Im *Mapping*-Schritt werden die Daten auf die visuellen Attribute (z.B. Position, Farbe oder Größe) abgebildet, wobei maßgeblich über die Eignung der resultierenden Darstellung entschieden wird. Die aus einer Hauptkomponentenanalyse gewonnenen Informationen können genutzt werden zu entscheiden, welche Variablen auf welche visuellen Attribute abgebildet werden sollen (z.B. PCA-*Scores* vs. Originaldaten). So können zum Beispiel die zu einem Trend gehörende Variable auf korrespondierende visuelle Attribute abgebildet, während Variable, die zu unterschiedlichen Trends beitragen, auf verschiedene visuelle Attribute gemappt werden (z.B. Position vs. Farbe, vgl. Mackinlay 1986).

Durch die Fokussierung auf bestimmte Trends oder Hauptkomponenten im *Filtering* ist es auch möglich, im *Mapping* kompaktere Bilder zu generieren, wodurch komplexe Darstellungen mit wiederkehrenden Abhängigkeiten vermieden werden können (vgl. Abb. A.11 im Anhang, wo die Merkmale mit Beitrag zur zweiten Hauptkomponente in einem Scatterplot dargestellt werden).

Neben der Unterstützung der PCA beim Visualisierungsdesign der Originaldaten, können auch die in den Hauptkomponentenraum überführten Daten direkt dargestellt werden (insb. die Scores C). So können zum Beispiel für den Datensatz der Mai-Tagesdaten eine Scatterplot-Matrix der ersten vier Hauptkomponenten mit den wichtigsten Trends in den Daten direkt wiedergegeben werden (vgl. Abb. 6.25). Interessanterweise, ergibt sich in diesem Beispiel eine starke Abhängigkeit zwischen den Hauptkomponenten P1 und P3.

[18] PC1 hat die höchste und PCN die geringste Signifikanz.

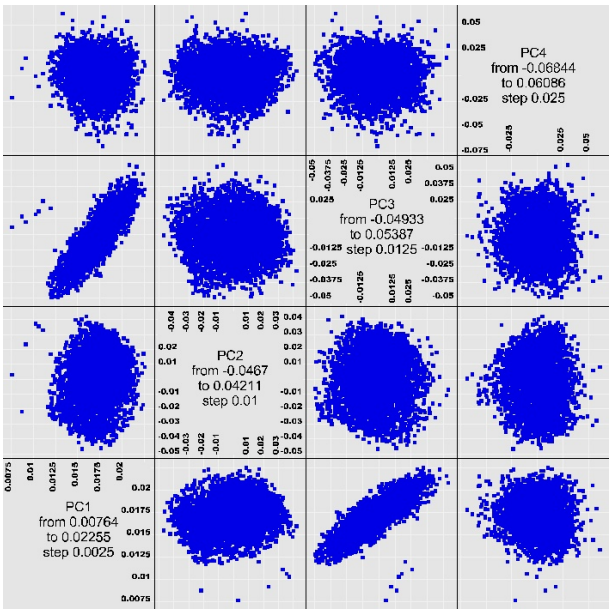


Abbildung 6.25: Scatterplotmatrix-Darstellung der ersten vier Hauptkomponenten (*Scores*-Matrix S) für den Datensatz der Tageswerte aller Maitage (im System InfoVis3D erzeugt)

Dies ist typischerweise ein Zeichen für eine komplexere Abhängigkeit zwischen den zugehörigen Merkmalen, die nicht vollständig durch ein lineares Varianzmodell beschrieben werden kann.

Als Alternative zu diesen Standarddarstellungen lässt sich die Tabellendarstellung (vgl. Kreuseler u. Schumann 2002a) auch einsetzen, um die *Scores* analog zu den *Loadings* darzustellen. Abbildung 6.26 zeigt die *Scores* in der Tabelle geordnet nach der ersten Hauptkomponente. Der Vorteil dieser Art der Darstellung ist, dass einzelne Datenobjekte mit bestimmten *Scores* leichter identifizierbar sind, und so einen tieferen Einblick in die Bedeutung und die Stärke der Trends ermöglichen. So repräsentieren die letzten fünf Zeilen den durch die erste Hauptkomponente repräsentierten Trend am stärksten (u.a. die Tage 13.05.1969, 27.05.2001 und 27.5.1963) sowie die ersten fünf Zeilen diesen Trend am schwächsten (Maitage 1945).

Bei der Darstellung der Hauptkomponenten für sich wird von den Originalvariablen abstrahiert. Damit ist es für den Anwender schwer, die dargestellten Trends und die darin vorliegenden Muster mit den Originalvariablen in Beziehung zu setzen. Um die Interpretierbarkeit der Achsen zu verbessern, können neben einer geeigneten Beschriftung (vgl. Abs. 6.2.2.4) auch PCA-transformierte Daten und Originaldaten in einer Darstellung kombiniert werden. So kann in dem Beispiel der Mai-Tagesdaten ein Scatterplot eingesetzt werden, um die Beziehung zwischen den Merkmalen *relative Luftfeuchte* und *Bedeckungsgrad* sowie der zweiten Hauptkomponente zu veranschaulichen (vgl. Abb. 6.27a). Hierbei zeigt sich eine starke Korrelation zwischen diesen drei Variablen, wobei die individuellen Unterschiede in den einzelnen Scatterplots deutlich sichtbar werden.

Neben der gemeinsamen Darstellung von abhängigen Variablen (Merkmalen) und zugehörigen Hauptkomponenten können auch unabhängige Variable (Dimensionen des Beobachtungsraumes) - welche typischerweise von der PCA ausgenommen werden - und Hauptkomponenten gemeinsam dargestellt werden. So können für den Datensatz der Maitage die Hauptkomponenten zum Beispiel mit einer pixelbasierten Darstellungstechnik in ihrem zeitlichen Bezug veranschaulicht werden (vgl. Abb. 6.27b).

Alternativ dazu können anstelle der *Score*-Werte auch die *Loadings* zusammen mit den Originalwerten dargestellt werden. Diese repräsentieren die Richtung der Hauptkomponenten und können z.B. in einer Scatterplot-Matrix direkt als Gerade in die einzelnen Plots eingezeichnet werden (vgl. Müller u. Alexa 2004).

6.2.2.4 Rendering und Hauptkomponentenanalyse

Bisher wurde die direkte Visualisierung von Originaldaten und/oder von durch die Hauptkomponentenanalyse generierten Daten diskutiert. Darüber hinaus bleibt zu untersuchen, inwieweit sich die explizit gemachten Trend-Informationen auch zur Unterstützung des Rendering-Prozesses einsetzen lassen, wo bei sie nicht direkt dargestellt werden.

Fokussierung und Anordnung. Im speziellen können die PCA-Scores beim Brushing & Lin-

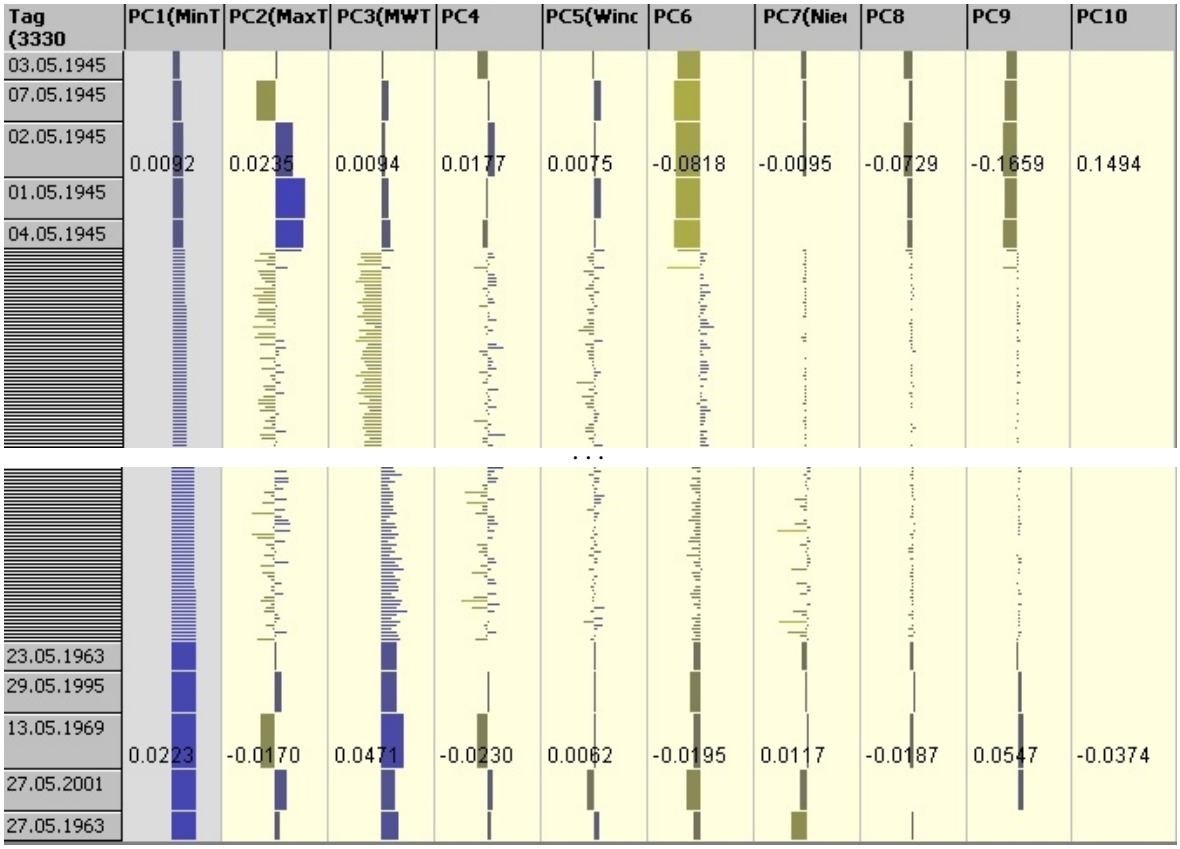


Abbildung 6.26: Tabellendarstellung der in den Hauptkomponentenraum transponierten Daten (*Scores*-Matrix S) für den Datensatz der Tageswerte aller Maitage; sortiert nach erster Hauptkomponente (PC1); zwei Foki; Rechtecke normiert nach Signifikanzen, Beschriftung mit unnormierten *Score*-Werten

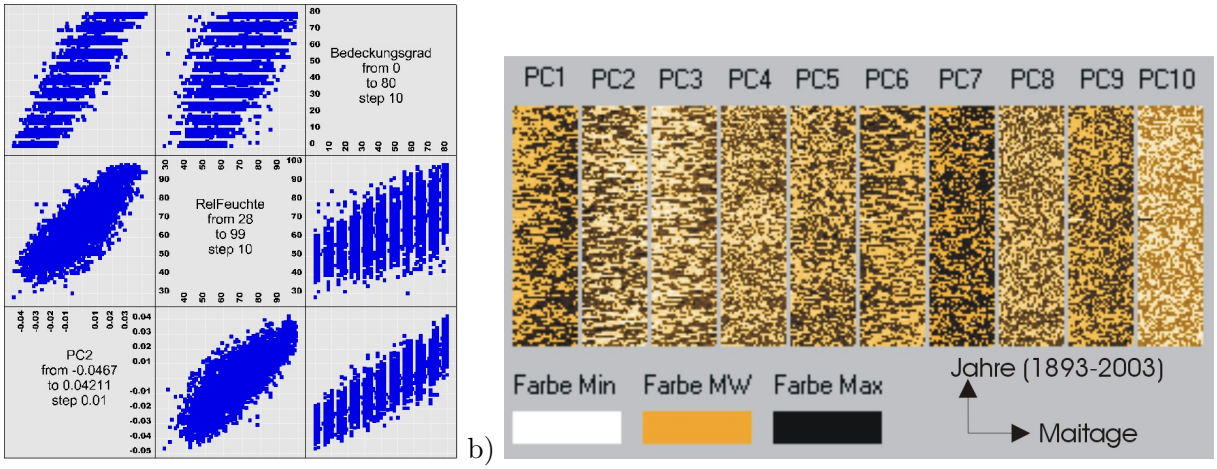


Abbildung 6.27: Darstellungen von Hauptkomponenten und Merkmalen: a) Gemischte Scatterplotmatrix-Darstellung mit Hauptkomponente PC2 und zwei Merkmalen, b) Pixelorientierte Darstellung aller Hauptkomponenten

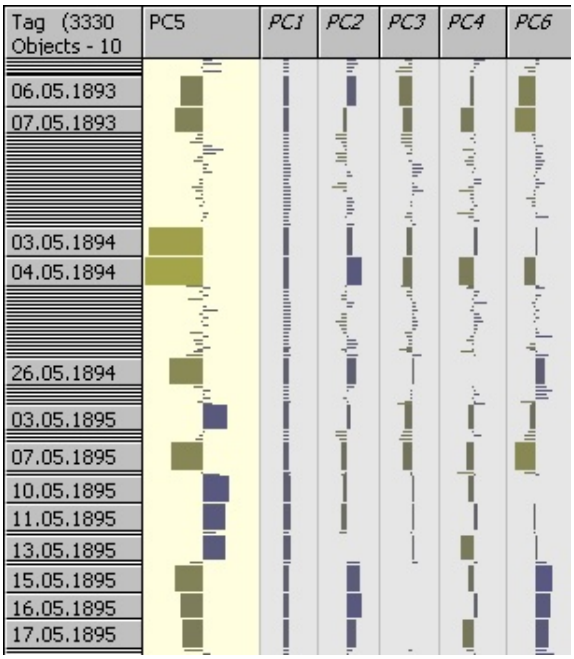


Abbildung 6.28: Tabellendarstellung der Scores (PC1-PC6) mit semantischer Linse (große Absolutwerte von PC5); Maitagesdatensatz

king eingesetzt werden. Dabei können z.B. Datenobjekte mit hohen *Score*-Werten in einer Darstellung der Hauptkomponenten selektiert, und in einer Darstellung der Originaldaten hervorgehoben werden (vgl. Tabellendarstellung und Scatterplot in Müller u. a. (2006)). Neben einer einfachen Hervorhebung können auch weitere Informationen für die so selektierten Objekte präsentiert werden (z.B. Beschriftungen oder semantisches Zooming). So können beispielsweise auch Datenobjekte in einer Tabellendarstellung unter Einsatz einer Linsenfunktion fokussiert werden (vgl. Abb. 6.28). Hierbei werden Datenobjekte mit hohen absoluten Trendwerten in der Hauptkomponente PC5 mit mehr Details dargestellt (durch größere Zeilenbreite).

Diese Art des Einsatzes von PCA-*Scores* kann zur Steuerung des Renderingprozesses sowohl manuell als auch automatisch erfolgen. So können wichtige Trends, Datenobjekte oder Ausreißer in der Startdarstellung automatisch fokussiert und diese Fokussierung dann interaktiv verändert werden.

Neben dem Fokussieren und Hervorheben von Objekten kann die PCA auch eingesetzt werden, um eine Sortierung der originalen Variablen oder Datenobjekte durchzuführen. So können zum Beispiel die Datenobjekte in der Tabellendarstellung der Originaldaten nach der ersten Hauptkomponente PC1 umsortiert werden (vgl. Abb. A.12 im Anhang).

Beschriftung. Wie bereits diskutiert, ist ein generelles Problem, die Hauptkomponenten mit den Originaldaten in Beziehung zu setzen (vgl. auch Abs. 6.2.2.3). Als Lösungen für dieses Problem wurden kombinierte Darstellungen aus originalen Variablen und Hauptkomponenten (Abb. 6.27a) sowie die indirekte Darstellung von Hauptkomponenten im originalen Datenraum verwendet (Abb. 6.28). Hierdurch erhält der Anwender einen interaktiven Zugang zur Bedeutung der durch die PCA erzeugten Trends.

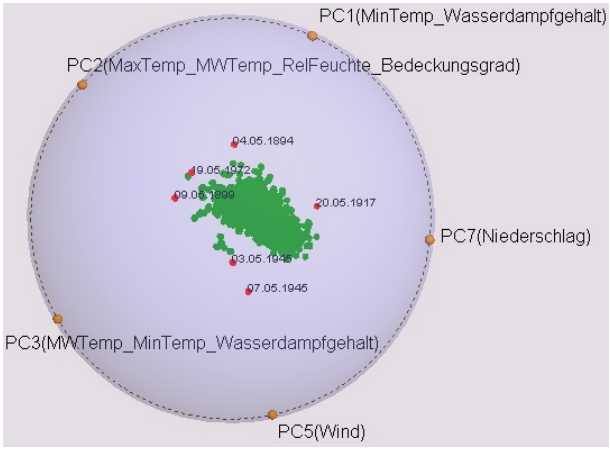


Abbildung 6.29: Beschriftung von Hauptkomponenten mit assoziierten Variablen (Technik ShapeVis (vgl. Theisel u. Kreuseler 1998) aus dem System InfoVis3D)

Als eine weitere Lösung dieses Problems kann auch die Information aus den *Eigenwerten* (Matrix λ^2) und den *Loadings* (Matrix W) benutzt werden, um die Bedeutung der Hauptkomponenten explizit in die Visualisierung einzubeziehen. Herausforderung in diesem Zusammenhang ist es, die Bedeutung der Hauptkomponenten durch kompakte Achsenbeschriftungen für den Anwender verständlich zu machen. Abbildung 6.29 stellt eine erste Lösung dieses Problems dar, indem die Beschriftung der Hauptkomponentenachsen um die Namen der Originalvariablen ergänzt werden, die sie überwiegend konstituieren (über einen Schwellwert auf den normalisierten *Loadings*-Werten, 70% Zugehörigkeit nach Joliffe (1986)). Zusätzlich werden diese nach ihrer *Signifikanz* sortiert. Ein Problem besteht hierbei darin, dass nicht immer klar zu entscheiden ist, ob

eine Variable zu einem Trend gehört. Ein zweites Problem ergibt sich aufgrund dessen, dass einige Trends häufig eine größere Menge von Variablen beinhalten, was zu langen, die Darstellung dominierenden (PC2 und PC3 in Abb. 6.29) bzw. zu verdeckten oder abgeschnittenen Beschriftungen (Abb. 6.26) führt. Hierfür ist es sinnvoll, Beschriftungen „on demand“ darzustellen (z.B. durch Variation von Länge und Ausprägung der Beschriftungen).

6.2.3 Diskussion

In diesem Abschnitt wurde ein erster Ansatz zur systematischen Kombination von Hauptkomponentenanalyse und Visualisierung in den verschiedenen Schritten des Visualisierungsprozesses untersucht sowie deren Eignung am Beispiel eines Klimadatensatzes demonstriert. Es lässt sich konstatieren, dass die Anreicherung der Originaldaten mit den durch die PCA generierten Matrizen neue Möglichkeiten bei der Untersuchung von Trends eröffnet. Hierzu wurden interaktive Standardtechniken zur direkten oder indirekten Darstellung von PCA-Werten (*Scores*, *Loadings* und *Eigenwerte*) eingesetzt und zum Teil an die speziellen Erfordernisse angepasst. Durch die präsentierten Techniken wird neben neuen Möglichkeiten zur Untersuchung von Trends insbesondere auch das Verständnis über die Bedeutung der Hauptkomponenten verbessert.

Die hier vorgeschlagene Verknüpfung von Visualisierung und PCA ermöglicht es, den gesamten interaktiven, iterativen Analyseprozess zu unterstützen (vgl. das Mantra von Shneiderman (1996) „Overview first, Zoom and Filter, then Detail on Demand“). Sie stellt Hauptkomponenten sortiert durch ihre Signifikanz bereit, die eine gute Basis für eine Dimensionsreduktion liefern und damit die Erzeugung von *Überblicksbildern* ermöglichen. Des Weiteren, kann das *Zooming* durch die PCA-Werte gesteuert werden, z.B. um den Fokus auf interessante Trends oder Ausreißer zu setzen. Analog können Ausreißer oder Trends für bestimmte Aufgaben auch *herausgefiltert* werden. Letztendlich können nützliche *Details*, zum Beispiel adäquate Beschriftungen, hinzugefügt werden.

Tabelle 6.2 fasst die mit den vorgestellten Darstellungen durchführbaren Aufgaben auf Basis der verwendeten PCA-Daten auch in Kombination mit den Originaldaten zusammen. Auf der Diagonale stehen dabei Darstellungen, die genau die Daten der zugehörigen Spalte/Zeile darstellen und keine anderen.

Als Herausforderung für zukünftige Arbeiten verbleibt, das Potential des neuen Ansatzes auszubauen, zu testen und die eingesetzten Techniken zu verbessern (z.B. die Beschriftung der Achsen, wobei insbesondere auch die *Score*-Werte und deren Beziehung zu den Originalwerten einbezogen werden müssen).

6.3 Visuelles Data Mining zur Klimamodellbildung, -simulation und -evaluation

Bisher wurde eine Vielzahl von Visualisierungstechniken vorgestellt. Diese wurden entweder separat nach ihrem Bezug (räumlich, zeitlich, Merkmalsraum), nach speziellen Zielen (Vergleich) oder nach ihrer Kombinierbarkeit mit automatischen Methoden betrachtet. Damit ist eine gute Basis gegeben, um einen intuitiven, interaktiven Zugang zu komplexen Klimadatensätzen und neue Einsichten in die Daten zu erhalten und daraus Schlussfolgerungen über Klimaprozesse abzuleiten. Somit können gut abgegrenzte Fragestellungen der Anwender beantwortet werden.

Betrachtet man nun aber den komplexen Prozess der Modellierung und Simulation, so ergibt sich ein breites Spektrum an Aufgabenstellungen, welche in einem hohen Maße miteinander vernetzt sind. Dieser Prozess beinhaltet die Bildung erster Hypothesen, die Spezifikation eines Modells sowie dessen systematische Simulation und Evaluation. Häufige Vorgehensweise zur Bewältigung der dabei

	PC-*Scores*	**PC-*Loadings***	**PC-*Eigenwerte***	**Originaldaten**
PC-*Scores*	Analyse von Trends im PC-Raum (Abb. 6.25, 6.26, 6.28, 6.29) Interaktive Umsortierung von PCs (Abb. 6.26) Fokussieren auf Obj., die stark zu Trend beitragen (Abb. 6.26, 6.28, 6.29)	Beschriftung von PC-Achsen (Abb. 6.26, 6.29)	Beschriftung von PC-Achsen (Abb. 6.26, 6.29) Fokussieren auf wichtige Trends (Abb. 6.26)	Tieferes Verständnis für die Bedeutung der PC-Achsen erhalten (Abb. 6.27a,b) Trends im Beobachtungsraum untersuchen (Abb. 6.27b) Fokussieren und ordnen auf/von Objekten im Originalraum (Abb. A.12)
PC-*Loadings*	/	Kompakter Überblick über Stärke und Beziehung der Trends (Abb. 6.23, 6.24)	Fokussieren auf wichtige Trends (Abb. 6.24)	Tieferes Verständnis für die Bedeutung der PC-Achsen erhalten (Abb. 6.23, 6.24) Fokussieren und ordnen der Originalvariablen
PC-*Eigenwerte*	/	/	**?**	**?**
Originaldaten	/	/	/	Typische Visualisierungsaufgaben (Abb. 6.27a, A.11, A.12)

Tabelle 6.2: Kombinationen von generierten PCA-Daten und Originaldaten für die Visualisierung und zugehörige Ausgaben; jede Zelle repräsentiert eine Kombination von bestimmten Arten von Eingabedaten für eine direkte oder indirekte Visualisierung; die untere Dreiecksmatrix bleibt aus aus Symmetriegründen leer

auftretenden Aufgaben ist es, separate Module umzusetzen und die Modellierung in einer Standardprogrammiersprache durchzuführen. Dies hat neben dem Nachteil einer begrenzten Wiederverwendbarkeit auch eine eingeschränkte Transparenz in die internen Prozesse zur Folge. So beschränkt sich der Einsatz von Visualisierungstechniken zumeist auf die visuelle Ausgabe der simulierten Daten. Eine systematische Unterstützung des gesamten Modellierungs- und Simulationsprozesses durch Visualisierungsmethoden, aber auch durch geeignete grafische Nutzerschnittstellen (GUI) und durch automatische Verfahren, wurde in der gängigen Literatur vernachlässigt.

In diesem Spektrum präsentiert die vorliegende Arbeit einen allgemeinen Entwurf, und demonstriert erste Ansätze zu dessen Umsetzung. Ansatz dabei ist, isoliert vorliegende VDM-Techniken systematisch in den Modellierungsprozess einzubeziehen (vgl. hierzu auch Nocke u. a. 2003). Dazu muss untersucht werden, wie separat vorliegende VDM-Methoden gebündelt werden können, um diesen Prozess durchgängig mit ihnen sowie mit geeigneten Nutzerschnittstellen begleiten zu können. Eine solche starke Verknüpfung von VDM-Methoden und Methoden zur Modellbildung, -analyse und -evaluation ist ein bisher weitgehend offenes Problem. Im Ergebnis zielt es darauf ab, dem Modellierer einen Werkzeugkasten an die Hand zu geben, mit dem viele der bisherigen Aufgaben wesentlich erleichtert werden zu können.

Insbesondere soll hierzu der Modellierungsprozess von Klimamodellen verbessert werden. Eine spezielle Herausforderung aus Sicht der Anwendung ist es, den Modellierer bei der Identifikation von reduzierten (Klima-)Modellen zu unterstützen. Solche reduzierten Modelle ermöglichen es, spezifische Klimaprozesse zu untersuchen, die das Verhalten der klimatischen Systems unter bestimmten Bedingungen dominieren (vgl. Pedlosky 1987). Die Anwendbarkeit der Methodik soll anhand eines Beispiels aus der Klimaforschung - einem reduzierten Atlantikzirkulationsmodell - demonstriert werden.

Im folgenden wird ein Ansatz entworfen, welcher die grundsätzlichen Schritte bei der Unterstützung des Modellierungsprozesses durch das Visuelle Data Mining spezifiziert. Um dessen Potential zu verdeutlichen, werden anschließend die einzelnen Schritte am Beispiel einer Klimamodellierung de-

monstriert. Hierbei handelt es sich um bisher isoliert vorliegende Techniken, deren Integration zukünftigen Arbeiten vorbehalten bleibt.

6.3.1 Lösungsansatz, Herausforderungen und Anwendungshintergrund

Erst durch die Kombination von VDM-Methoden sowohl in der explorativen als auch in der konfirmativen Analyse ist die Basis gelegt, um den gesamten Modellbildungs-, Simulations- und Evaluationsprozess zu unterstützen. So kann z.B. eine Kombination aus Visualisierungtechniken und statistischen Methoden eingesetzt werden, um ein **Apriori-Wissen** über die vorhandenen Daten zu gewinnen und darauf basierend erste **Hypothesen** über Muster und Trends abzuleiten (hier: qualitative Identifikation von Klimaphänomenen). Darauf aufbauend kann die Stärke dieser Muster - durch die Wahl geeigneter statistischer Methoden - quantifiziert werden.

Auf diesem Wissen aufbauend kann dann ein geeignetes Modell entworfen werden. So können z.B. im Klimaumfeld aufgrund der visuellen Analyse von Modellresultaten und/oder Messdaten wichtige Systemvariablen für einen interessierenden Prozess identifiziert und basierend darauf die Freiheitsgrade eines komplexen physikalischen Modells reduziert werden. Dies bildet dann die Basis dafür, ein neues (reduziertes) Modell zu entwerfen. Allgemein kann der Modellierer bei der **Modellierung** neben der Bereitstellung geeigneter GUIs durch VDM-Methoden dabei unterstützt werden, ein neues Modell zu spezifizieren (z.B. Gleichungen für wichtige Systemvariable zu definieren), einen Überblick über die Struktur des Modells zu erhalten bzw. zu behalten sowie sein Modell basierend auf vorhandenen Daten geeignet zu parametrisieren.

Des Weiteren haben VDM-Methoden auch ein Potential, um die **Analyse eines gegebenen Modells** zu verbessern. So können sie überprüfen, ob das zugrunde liegende mathematische Modell korrekt in ein numerisches Modell überführt wurde, z.B. beim Einsatz von Differentialgleichungssystemen (DGLS) durch Diskretisierung der Modellgleichungen und die Anwendung von geeigneten Lösungsverfahren. Dies betrifft zum einen die Darstellung von ersten Ergebnissen von Modellsimulationen, um einen qualitativen Eindruck über die Güte der Modellierung zu bekommen und ggf. wieder zum Modellierungsschritt zurückzukehren. Zum anderen lässt sich mit Hilfe von VDM-Verfahren eine systematische Überprüfung der Modelleigenschaften bei Anwendung spezieller Simulationsverfahren (z.B. durch Multi-run-Simulationen) unterstützen.

Der letzte Schritt ist der Einsatz von VDM-Methoden zur **Modellevaluation**. Hier unterstützen sie zu überprüfen, ob wichtige Systemprozesse durch das Modell korrekt wiedergegeben wurden, was insbesondere einschließt, auftretende Fehler zu identifizieren, zu lokalisieren und zu quantifizieren. Dies bezieht insbesondere eine systematische Validierung der simulierten Modelldaten mit Referenzdaten oder Daten aus anderen Modellen ein.

Basierend auf den beschriebenen Schritten lässt sich der Einsatz von VDM-Verfahren für den Prozess der Modellierung und Simulation wie folgt zusammenfassen:

1. Datenanalyse im Vorfeld der Modellierung:
   - Exploration von Mess- oder simulierten Daten
   - Hypothesenbildung
2. Modellspezifikation:
   - Interaktiv und visuell gestützte Modellierung
   - Visuelle gestützte Parametrisierung von Modellen
   - Analyse des Modells/der Modellstruktur
3. Modellanalyse:
   - Analyse von Eingabedaten
   - Analyse der Ausgabedaten
4. Modellevaluierung:

- Modellausgaben mit simulierten Daten anderer Modelle validieren
- Modellausgaben mit Messdaten validieren
- Überprüfung von Hypothesen

Um den Modellierungsprozess optimal zu unterstützen, muss er **interaktiv** und **iterativ** gestaltet werden. Dies schließt insbesondere ein, dass der Modellierer möglichst direkt in die einzelnen Schritte eingreifen und bei Bedarf direkt zu vorangegangen Schritten zurückkehren kann. Herausforderung ist hierbei, die einzelnen Schritte geeignet miteinander sowie mit den VDM-Methoden in einem Framework zu verknüpfen. Um eine hohe Interaktivität und Nutzertoleranz zu gewährleisten, bedarf es ferner geeigneter Strategien, um **langandauernde Simulationsprozesse** und automatische Datenanalyseverfahren geeignet von schnellen Verfahren zu separieren. Weiterhin müssen **große, heterogene Datenmengen** verarbeitet werden können (Zustandsraum, Parameterraum, räumlicher und zeitlicher Bezug). Dies erfordert einen hohen Grad an Nutzerunterstützung sowie die Anwendbarkeit und Wiederverwendbarkeit für verschiedene Szenarien und Modelltypen.

6.3.2 Diskussion des Ansatzes am Beispiel eines Atlantikmodells

Natürlich kann die diskutierte Methodenverknüpfung nicht allgemeingültig gelöst werden. Vielmehr bedarf es einer klaren Einschränkung auf eine Anwendungsdomäne mit klar umrissenen Zielstellungen. Für die folgenden Betrachtungen soll hierfür die Reduktion und Evaluation von Klimamodellen zugrunde gelegt werden. Im speziellen ist zur Unterstützung der Modellierung reduzierter Modelle durch das VDM die folgenden Funktionalität erforderlich:

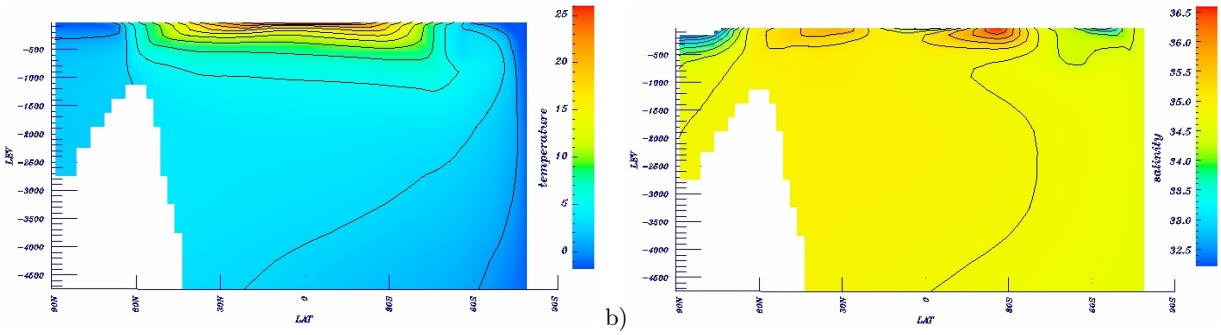
- Visuelle Exploration von Messdaten und von Simulationen komplexer Modelle um interne Prozesse und Muster zu identifizieren: hier lassen sich Feature-Extraktionsverfahren und Verfahren zur Feature-Visualisierung einsetzen (z.B. durch Einsatz von Clusterverfahren in Kombination mit Clustervisualisierung (vgl. auch Abs. 6.1)).
- Ableitung, Modifikation und effektive Berechnung von reduzierten Modellen: Auch hierbei lässt sich visuelle Unterstützung bereitstellen. Hierzu gehört die interaktive Definition von gewöhnlichen Differentialgleichungssystemen (gDGLS). Mit Hilfe der Visualisierung von Referenzdaten können ferner Parameter und Startwerte in ein solches gDGLS justiert werden.
- Analyse des Modellverhaltens: Hierbei können Verfahren eingesetzt werden, um die Sensitivität des reduzierten Modells auf die Veränderung des Integrationsverfahrens oder auf die Adaption von Startwerten einzuschätzen.
- Vergleich und Evaluierung des reduzierten Modells: Hierzu können VDM-Methoden zur gemeinsamen Darstellung von Parametern des neu spezifizierten, reduzierten Modelles mit den Parametern eines komplexen Modells oder mit Messdaten verglichen werden.

Diese Funktionalität wurde im Laufe der Arbeit in Form einzelner Module exemplarisch für das Beispiel eines Atlantikzirkulationsmodells umgesetzt. Anhand der dabei erzielten Ergebnisse soll im folgenden die Sinnhaftigkeit des vorgestellten Entwurfes illustriert werden. Eine allgemeine Lösung der aufgezählten Problemstellungen und deren systematische Integration ist jedoch aktueller Forschungsgegenstand.

6.3.2.1 Datenanalyse im Vorfeld der Modellierung

Im ersten Schritt erfolgt eine Suche in vorhandenen Datenbeständen, um neues Wissen über Zusammenhänge im realen System aufzudecken und so die Basis für die Definition eines neuen Modells zu legen. Hierzu lassen sich ein Großteil der bisher vorgestellten Techniken einsetzen, insbesondere auch eine Kombination aus Clusteranalyse, PCA und Visualisierung, um die auftretenden großen Datenmengen handhabbar zu machen.

Exploration von Mess- oder simulierten Daten. Ziel hierbei ist es, einen Überblick über wichtige Eigenschaften und Abhängigkeiten von beteiligten Variablen zu gewinnen (hier für ein komplexes Atlantikmodell). So illustriert Abbildung 6.30 die räumliche Verteilung der Merkmale Temperatur und Salzgehalt zum Anfang der Simulation (2D-Schnitt im Atlantik). Insbesondere



b)

Abbildung 6.30: Exploration eines komplexen Atlantikmodells (vertikaler Schnitt im Ozean); a) Temperaturverteilung; b) Verteilung des Salzgehaltes

lassen sich so Bereiche mit starken Temperatur- und Salzgehaltdifferenzen an der Oberfläche identifizieren und lokalisieren. Um ferner auch die Veränderung der Daten über die Zeit zu untersuchen, lassen sich auch 3D-Darstellungen mit Ersetzung einer Achse durch die Zeit einsetzen. Speziell für das Atlantikmodell (Climber-II) wurden solche Darstellungen im vorangegangenen Kapitel diskutiert (vgl. Darstellung von Temperatur & Salzgehalt mit DVR (Abb. 5.13e), mit Isoflächen (Abb. 5.13f) sowie mit der Differenzmethode (Abb. 5.14b-d).

Hypothesenbildung. Im Klimaumfeld lassen sich multivariate statistische Methoden mit Fehlermaßen kombinieren, um wichtige Variable und deren Einfluss auf komplexe Prozesse zu identifizieren und zu quantifizieren (Böhm 1999; Kücken u. a. 2002). Eine Kombination solcher Methoden mit Visualisierungstechniken bietet eine breite Basis zur Findung und Testung erster Hypothesen. So können für das Atlantikmodell basierend auf den Erkenntnissen (zu bestehenden Potentialunterschieden) erste Hypothesen über bestehende Muster in den Daten aufgestellt werden. Weiterhin ermöglicht das Merkmal „overturning stream function“ eine Zirkulation zu identifizieren und zu lokalisieren. So zeigt Abbildung 6.31 diese Zirkulation reichend von 60°n.B. bis 30°s.B. in variierenden Tiefenschichten.

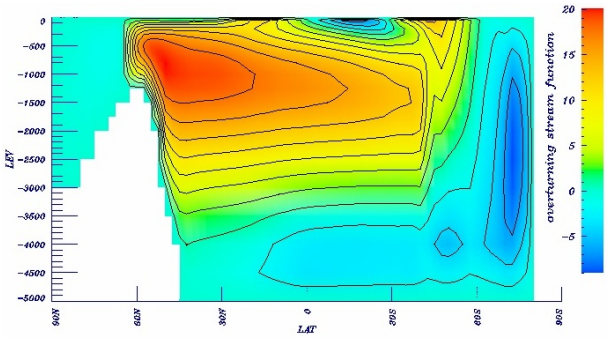


Abbildung 6.31: Hypothesenbildung im Atlantikmodell basierend auf Stromfunktion (*overturning stream function*))

Basierend auf der Hypothese einer Zirkulation in diesem Bereich können nun wichtige antreibende Modellparameter identifiziert und genauer untersucht werden. In diesem Fall sind dies die bereits bei der Exploration untersuchten Merkmale Temperatur und Salzgehalt, die wesentlich zum Antrieb dieser Strömung beitragen. Um die Potentialunterschiede in diesen Merkmalen zu verdeutlichen, wurden für diese beiden zusätzlich die Gradienten berechnet und als Pfeile in einem Bereich von Interesse dargestellt (vgl. Abb. 6.32).

Zusätzlich lässt sich auch der gemeinsame Einfluss der beiden Merkmale auf eine Zirkulation untersuchen, indem statt der beiden separaten Darstellungen eine Linearkombination der normierten Merkmale dargestellt wird. Hierbei kann der Anwender je nach Wahl der Gewichtung den Salzgehalt (Gewicht 0, Abb. 6.32a), die Temperatur (Gewicht 1, Abb. 6.32b) oder eine beliebigen Kombination dieser beiden erhalten (z.B. Gewicht 0.5 in Abb. 6.32c).

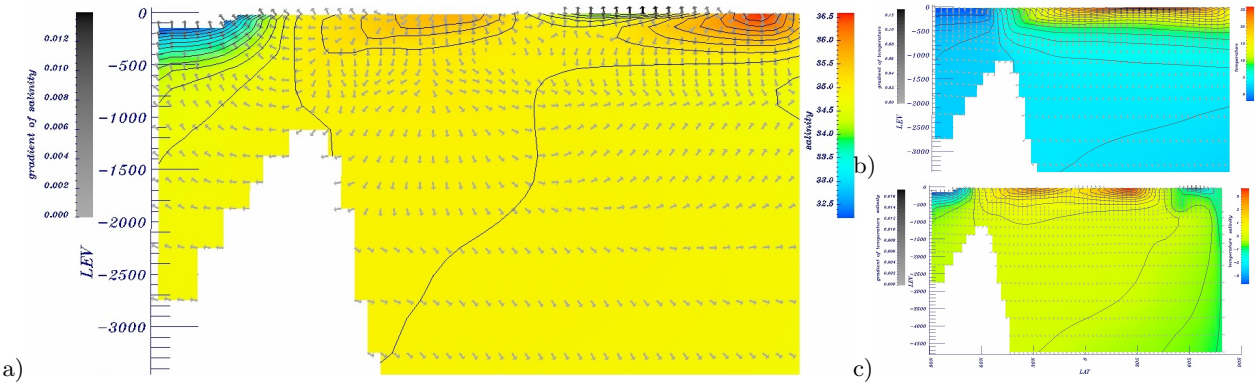


Abbildung 6.32: Hypothesenbildung basierend auf den Gradienten (Pfeildarstellung) der beteiligten Merkmale; a) Salzgehalt; b) Temperatur; c) Linearkombination (0.5) aus normierter Temperatur und normiertem Salzgehalt

6.3.2.2 Modellspezifikation

Im nächsten Schritt wird basierend auf den aufgestellten Hypothesen ein neues Modell entworfen. Im Beispiel des Atlantikmodells heißt dies, dass ein reduziertes Modell gefunden werden muss, welches eine Zirkulation beschreibt, die durch die Parameter Temperatur und Salzgehalt angetrieben wird. Abbildung 6.33 zeigt das zugrunde liegende mentale Modell (vgl. Rahmstorf u. Ganopolski 1999), welches basierend auf den Hypothesen entworfen wurde. Dieses reduzierte Modell enthält vier rechteckige Bereiche (vier Regionen im Atlantik repräsentierend), wobei jeweils benachbarte Bereiche basierend auf ihrer Salzgehalt- und Temperaturdifferenz einen Massestrom antreiben.

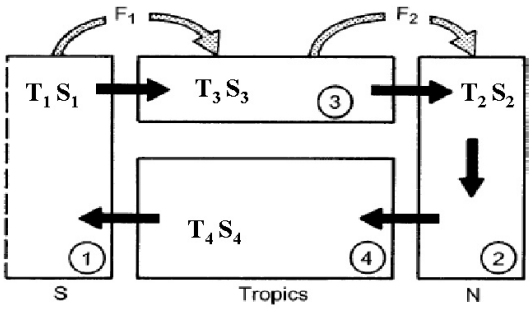


Abbildung 6.33: Illustration des Box-Atlantikmodells mit vier rechteckigen Bereichen, den Parametern Temperatur T_i und Salzgehalt S_i für jeden Bereich sowie dem Frischwasserzufluss F_1 und F_2

Im folgenden soll aufgezeigt werden, wie die Anwender bei der Spezifikation eines solchen vereinfachten Modelles durch GUI- und VDM-Methoden unterstützt werden kann.

Interaktiv und visuell gestützte Modellierung. Um die Modellierung von vereinfachten (Klima-)Modellen visuell zu unterstützen, wurde ein interaktives Werkzeug (*DesModelEditor*) zur Spezifikation und Simulation von diskretisierten gewöhnlichen Differentialgleichungen (gDGLS) entworfen[19] (vgl. Abb. 6.34 mit spezifiziertem Atlantikmodell).

Mit diesem Werkzeug kann eine beliebige Anzahl von diskretisierten Differentialgleichungen (Abb. 6.34 oben), Initial- und Randbedingungen (Abb. 6.34 Mitte) sowie Nebenbedingungen und Konstanten (Abb. 6.34 unten) spezifiziert werden. Neben der Modellierungsfunktionalität können auch experimentspezifische Metadaten wie Schrittweiten und Anzahl an Integrationschritten festgelegt werden (oben rechts). Verschiedene Lösungsverfahren wurden integriert, die auch eine Fehlerberechnung und eine adaptive Schrittweitenanpassung erlauben (Mitte rechts). Darüber hinaus wurde eine spezielle Visualisierungsschnittstelle entworfen, die es erlaubt, den Modellierungsprozess eng mit dem VDM zu koppeln. Neben einer Vielzahl von Techniken zur Darstellung von Experimentdaten innerhalb des Werkzeuges (vgl. Abs. 6.3.2.3) können über eine externe Schnittstelle auch direkt Daten an das Visualisierungssystem OpenDX übermittelt und dort dargestellt sowie Daten von dort empfangen werden.

[19] Die Umsetzung erfolgte im Rahmen einer betreuten Studienarbeit (Holst 2003) und wurde im Rahmen einer Diplomarbeit weiterentwickelt (Zornow 2006).

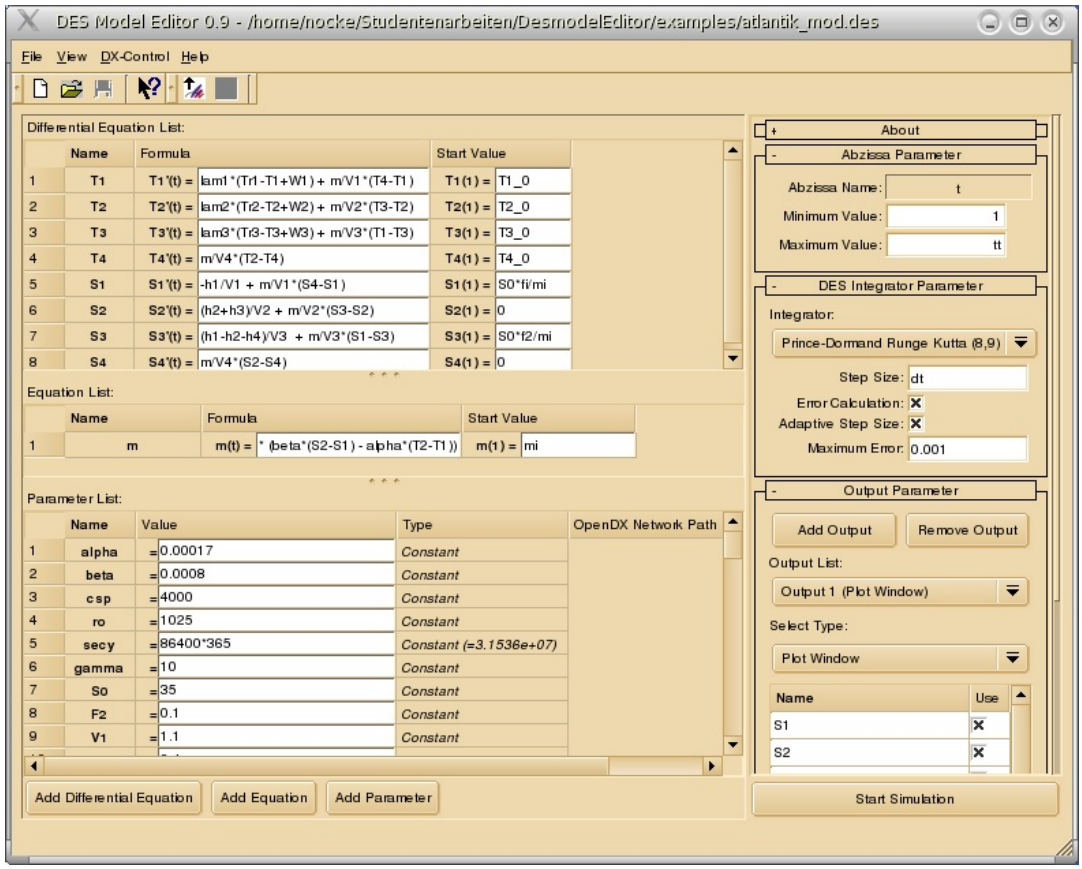


Abbildung 6.34: Modellspezifikation mit dem *DES Model Editor*: Interaktive Definition von gDGLS und deren Integration; vereinfachtes Atlantikmodell mit 9 Variablen (Mittelwerte der 4 Atlantikregionen in Temperatur ($T_1 \dots T_4$) und Salzgehalt ($S_1 \dots S_4$); Massestrom m)

Mit dem *DesModelEditor* steht dem Anwender so ein interaktives Werkzeug zur Verfügung, welches (auch im Vergleich zu gängigen Mathematik-Softwaresystemen) leicht bedienbar ist, eine Vielzahl von Lösungsverfahren bereithält und eine flexible Kopplung mit interaktiven Visualisierungstechniken erlaubt.

Parametrisierung von Modellen. Auch zur Parametrisierung von Modellen kann die Visualisierung eingesetzt werden. Insbesondere betrifft dies die Unterstützung des Nutzers bei der Spezifikation der Startwerte des entworfenen Modells. Bei diesem (zum Teil zeitaufwendigen) Prozess kann der Anwender unterstützt werden, das spezifizierte Modell mit vergleichbaren Anfangswerten wie ein anderes Modell oder basierend auf Messwerten zu parametrisieren. Abbildung 6.35 zeigt, wie der Anwender in einer Darstellung des Merkmals „overturning stream function“ vier rechteckige Bereiche auswählt, welche jede eine Region mit ähnlichen Eigenschaften der antreibenden Variablen für die atlantische Zirkulation repräsentiert. Aufgrund dieser Bereiche können gemittelte initiale Bedingungen für das vereinfachte Modell berechnet sowie später ein Modellvergleich durchgeführt werden (vgl. Abs. 6.3.2.4).

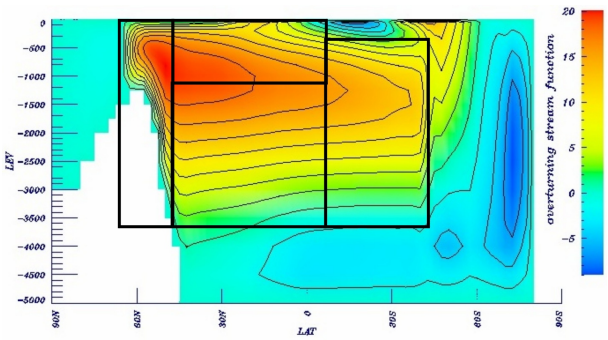


Abbildung 6.35: Parametrisierung des reduzierten Atlantikmodells basierend auf der Stromfunktion (*overturning stream function*) des komplexen Atlantikmodells)

Analyse des Modells/der Modellstruktur. Bei komplexen Modellen mit einer Vielzahl von Zustandsgrößen und Parametern, die sogar in ver-

schiedenen, miteinander verknüpften Teilmodellen vorliegen können, kann die Struktur des Modells schnell unübersichtlich werden. Hierzu lassen sich Visualisierungstechniken zur Darstellung von Strukturen einsetzen (vgl. hierzu auch Schulz u. a. 2006b). Abbildung 6.36 zeigt eine Strukturdarstellung des Einflusses der verschiedenen Differentialgleichungen im Atlantikmodell aufeinander (System Colossus, vgl. Schulz (2005) und Schulz u. a. (2006a)).

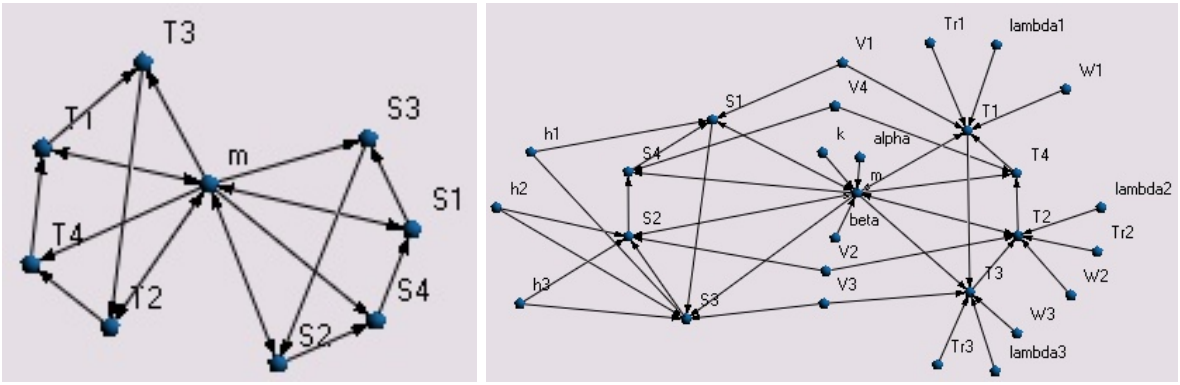


Abbildung 6.36: Darstellung der Struktur des Atlantikmodells (System Colossus, vgl. Schulz (2005)); a) Reduzierte Darstellung der Zustandsgrößen T_i und S_i sowie des Masseflusses m; b) Darstellung aller Parameter und Zustandsgrößen

Für dieses relativ kleine gDGLS ergibt bereits ein relativ komplexer gerichteter Graph (Abb. 6.36b). Bei der Darstellung noch komplexerer Systeme - wie sie in der Klimaforschung üblich sind - besteht die Herausforderung, Strukturdarstellungstechniken durch Einsatz von Interaktionstechniken (z.B. *Information Hiding* oder *Fokus & Kontext*) geeignet anzupassen, um verschiedene Aspekte des eines solchen Differentialgleichungssystems bei Bedarf im Detail untersuchen zu können. So zeigt Abbildung 6.36a eine Darstellung, in der auf die Zustandsgrößen sowie den Massestrom fokussiert wird.

Ferner lassen sich aus einer solchen Darstellung Aussagen über wichtige Modellzustandsgrößen und -parameter ableiten sowie die Modellstruktur kommunizieren. So kann aufgrund der Nachbarschaft eines Knotens schnell auf den Einfluss einer Größe auf das Modell abgeschätzt werden (auf welche Größen sie direkt oder indirekt einwirkt). Weiterhin wird deutlich, ob wichtige Kausalzusammenhänge durch das Modell korrekt wiedergegeben wurden (z.B. die Zyklen der Salzgehalte und Temperaturen in Abb. 6.36).

6.3.2.3 Modellanalyse

Nach der Modellspezifikation ist der nächste Schritt, das Modellverhalten genauer zu untersuchen (Stabilität, Störungen im Modell, ...). Hierbei ergeben sich insbesondere im Klimaumfeld besondere Herausforderungen, wenn es sich bei den simulierten Daten um 3D-zeitabhängige Daten mit mehr als 100 Merkmalen handelt. Dieser Schritt kann stark von existierenden VDM-Methoden profitieren. Speziell lassen sich hier alle Visualisierungtechniken für die Darstellung der Daten im räumlichen und zeitlichen Bezug sowie im Merkmalsraum einsetzen (vgl. Abs. 5.1, 5.2 und 5.3). Darüber hinaus - um sich einen Überblick über die Abhängigkeiten und Trends in den Merkmalen zu erhalten - sind VDM-Techniken besonders geeignet, welche die Beziehungen der Variablen untereinander darstellen (SOM-Karte in Abs. 6.1.6 sowie PCA-Darstellungen in Abs. 6.2). Weiterhin ergeben sich spezielle Anforderungen bei der Einbeziehung von Fehlern und Schrittweiten in die visuelle Analyse solcher Daten sowie an die Darstellung von Multirun-Simulationsdaten.

VDM gestützte Modellsimulation und Analyse der Ausgabedaten. Eine direkte, visuelle Rückkopplung bei der Durchführung von Simulationen kann den Prozess der Modellbildung wesentlich unterstützen. So kann der Anwender den Einfluss bestimmter Simulationsverfahren und

-parameter oder von Startwerten direkt beurteilen. Insbesondere bei gDGLS mit relativ geringen Berechnungszeiten lässt sich der Simulationsprozess direkt mit der Visualisierung koppeln. Hierzu wurde in den DesModelEditor eine spezielle Komponente zur gekoppelten Visualisierung von Modellvariablen, Schrittweiten und Verfahrensfehlern eingebunden[20] (vgl. Abb. 6.37).

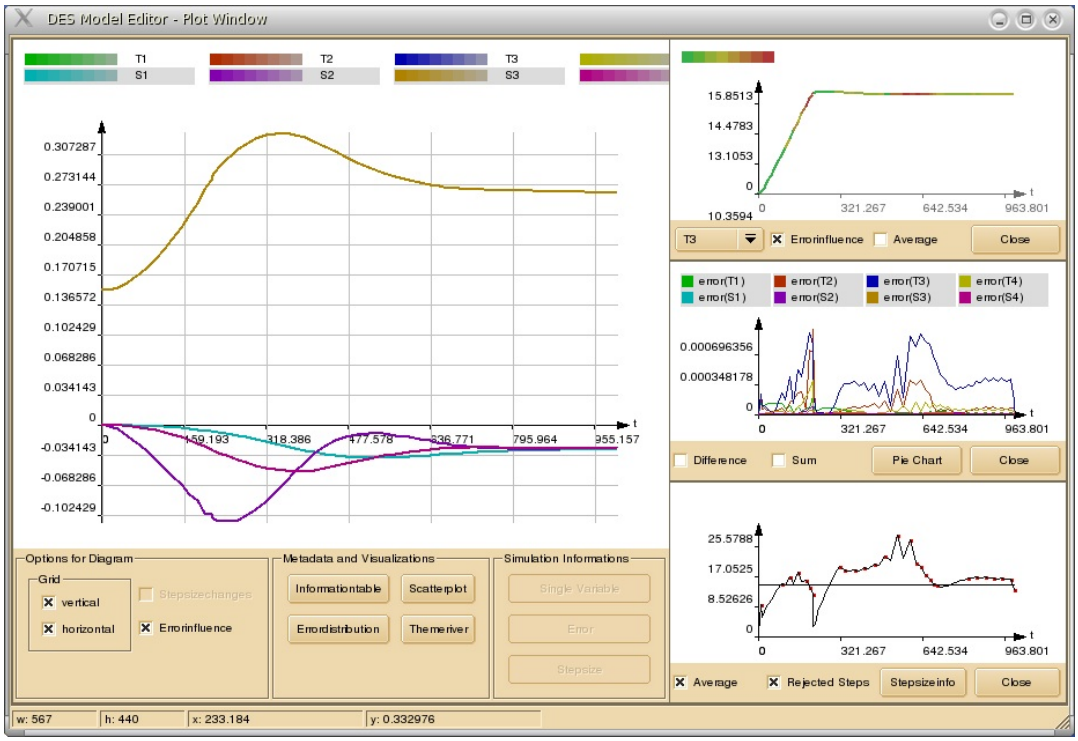


Abbildung 6.37: Visualisierung eines gDGLS mit dem *DES Model Editor*; Darstellung von (relativen) Salzgehalten $S_1 \ldots S_4$ (links), einer Temperaturfunktion (rechts oben), sowie den zugehörigen Fehlerabschätzungen (rechts Mitte) sowie den Schrittweitenanpassungen des Verfahrens (rechts unten)

Die Idee dieser Komponente ist es, statische Darstellungen, wie sie sich in Mathematikbüchern finden, dem Anwender interaktiv verfügbar zu machen. Herausforderung bei der Darstellung von Modellen aus gDGLS ist es dabei, eine eine hohe Anzahl von Zustandsgrößen und anderen Parametern mit unterschiedlichen Wertebereichen zusammen mit den Fehlern effektiv untersuchen zu können. Erst durch eine direkte Interaktion mit den Modellausgaben können bei größeren Modellen effektive Visualisierungen erzeugt werden. So lassen sich selektierte Modellausgabevariablen auch mit verschiedenen Wertebereichen in mehreren, nebeneinander liegenden Darstellungen gekoppelt werden (vgl. Abb. 6.37, wo separate Plots für Salzgehalte, deren Fehlerabschätzungen sowie die durch das Verfahren vorgenommenen Schrittweitenanpassungen dargestellt werden). Hierbei lässt sich die Fehlerbehaftung im Sinne einer Unsicherheit direkt für die Darstellung der Kurve einsetzen. Je nach Fragestellung können fehlerbehaftete Zeitbereiche abgeschwächt (Abbildung auf Transparenz in Abb. 6.37 rechts) oder verstärkt (Abbildung auf rot-grün-Skala) dargestellt werden.

Zusätzlich werden dem Anwender verschiedene Interaktionsmöglichkeiten bereitgestellt, z.B. das Ein- und Ausblenden einzelner Graphen, die textuelle Darstellung einzelner Werte und die textuelle und visuelle Einblendung von statistischen Kenngrößen der Kurven.

Allerdings sind die Möglichkeiten, versteckte Abhängigkeiten aufzudecken, mit solchen Zeitgraphen begrenzt. Hierfür bieten sich Darstellungstechniken aus der Informationsvisualisierung an, welche für solche Simulationsdaten bisher kaum eingesetzt werden. Um z.B. auch Abhängigkeiten höherer

[20] vgl. hierzu betreute Diplomarbeit Zornow (2006).

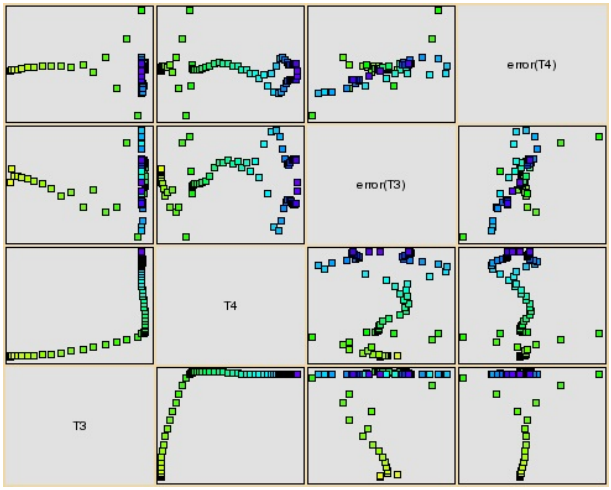


Abbildung 6.38: Scatterplot-Darstellung der Temperaturen (T_2 und T_3) deren abgeschätzten Fehlern ($error(T_2)$ und $error(T_3)$)

Ordnung von Modellgrößen untersuchen zu können, lassen sich in diesem Kontext Scatterplotmatrizen eingesetzen. So zeigt Abbildung 6.38 die gemeinsame Scatterplot-Matrix-Darstellung zweier Temperaturen (T_3 und T_3) in Kombination mit ihren Fehlern ($error(T_2)$ und $error(T_3)$) sowie durch Farbkodierung deren zeitlichen Verlauf. Diese Darstellung erlaubt unter anderem einzuschätzen, in welchen Wertebereichen einer Zustandsgröße die meisten und größten Fehler auftreten.

Eine zweite Darstellungstechnik, die es erlaubt, auch eine größere Anzahl von Variablen in ihrem zeitlichen Verlauf übersichtlich darzustellen, ist die Technik Themenfluss. So zeigt Abbildung 6.39 erste Darstellungen der acht Modellzustandsgrößen des Atlantikmodells (Abb. 6.39 links) sowie die zugehörigen Fehler (Abb. 6.39 rechts). Vorteil dieser Technik ist die Vermeidung von Verdeckungen sowie eine verbesserte Vergleichbarkeit auch von Größen unterschiedlicher Wertebereiche durch die bei dieser Technik vorgenommene Normierung. Inwieweit unterschiedliche Größen (hier: Temperaturen und Salzgehalte) in einer Darstellung zusammen dargestellt werden dürfen und wie diese geeignet angeordnet und farbkodiert werden, ist stark kontextabhängig. So ergibt sich z.B. bei der Fehlerdarstellung für das Atlantikmodell der visuelle Effekt, dass sich eine starke Schwankung einer oder mehrerer Fehler (vgl. Fehlerdarstellung in Abb. 6.39b) im „Inneren des Flusses“ auch auf Fehler am „Rande des Flusses“ fortpflanzt, obwohl diese selbst dieser Oszillation nicht unterworfen sein müssen.

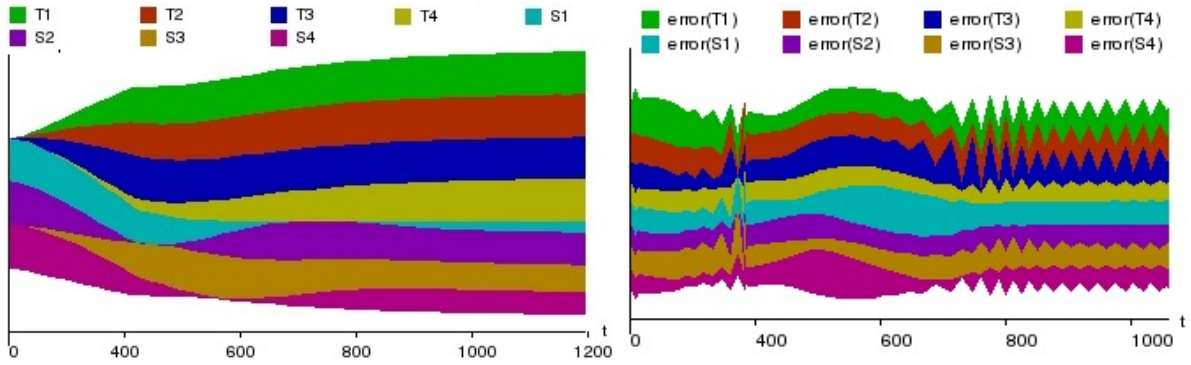


Abbildung 6.39: Themenflussdarstellungen des Atlantikmodells im *DES Model Editor*; Kombinierte Darstellung von normierten Temperaturen $T_1 \ldots T_4$ und Salzgehalten $S_1 \ldots S_4$ (links), sowie den zugehörigen Fehlerabschätzungen $error(T_1) \ldots error(T_4)$ sowie $error(S_1) \ldots error(S_4)$ (rechts)

Neben der Untersuchung einzelner Simulationen, ist gerade auch im Klimaumfeld eine wichtige Aufgabe, mehrere Modellsimulationen durchzuführen und gemeinsam auszuwerten, um einen allgemeineren Eindruck über die Modelleigenschaften zu erhalten. Ein erster Ansatz zur Darstellung von Ergebnissen solcher Multi-Run-Simulationen (im räumlichen Bezug) ist die Abbildung der verschiedenen Läufe auf verschiedene Attribute einer Ikone (vgl. Abb. 5.3). Um mehrere Simulationen eines gDGLS untersuchen zu können, bieten sich insbesondere multivariate Zeitdarstellungstechniken an.

Eine erste Möglichkeit zum Vergleich zweier Simulationen des Atlantikmodells ist deren kombinierte Darstellung in einem Zeitgraphen, wobei die einzelnen Modellläufe farblich voneinander separiert werden (vgl. Abb. 6.40).

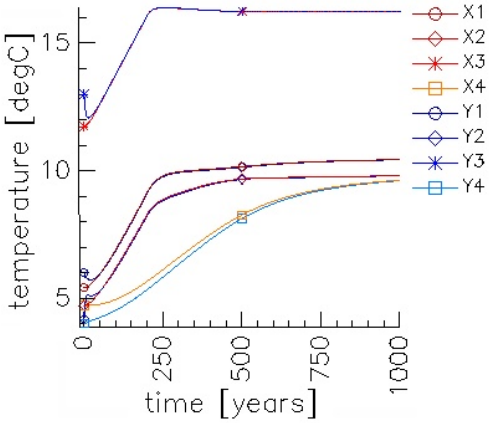


Abbildung 6.40: Zeitgraph zweier Modellläufe (Lauf X: rot, Lauf Y: blau) des vereinfachten Atlantikmodells mit je 4 Temperaturvariablen

Die rote Kurvenschar repräsentiert die vier Atlantikregionen für den Referenzlauf X, während die blaue Kurvenschar einen Lauf Y darstellt, bei dem interaktiv die Startwerte angepasst wurden. Die Kurven der einzelnen Regionen (1 bis 4) können separat identifiziert werden (durch kleine Glyphen und verschiedene Helligkeiten). Dies ermöglicht Stabilität und Konvergenz des Modells einzuschätzen. So konvergieren zum Beispiel die Temperaturkurven zu den Atlantikregionen 1, 2 und 3 (X1 und Y1, ...) in beiden Läufen in weniger als 100 Integrationsschritten, während dies bei den Variablen X4 und Y4 erst nach mehr als 500 Integrationsschritten geschieht.

6.3.2.4 Modellevaluierung

Die Modellevaluierung ist von besonderer Wichtigkeit, um die Leistungsfähigkeit des Modells qualitativ und quantitativ einzuschätzen. Dies beinhaltet den Vergleich des Modells basierend auf dessen Simulationen mit anderen Modellen und mit Messdaten. Herausforderungen hierbei sind der Vergleich von Modelldaten auf abweichenden Gittern (vgl. Abs. 5.4). Insbesondere lassen sich hierzu statistische Verfahren und Visualisierungstechniken miteinander koppeln (vgl. gemeinsames Paper mit den Klimaforschern Böhm u. a. 2004). Beispielsweise existiert eine große Anzahl an statistischen Maßen, um Modellfehler in Klimamodellen zu quantifizieren (vgl. z.B. Jones u. a. 1995).

Modellausgaben miteinander validieren. Im Falle eines komplexen Modells mit einem vereinfachten Modell (gDGLS) können die Vorhersagen dieser beiden Modelle miteinander verglichen werden. So lässt sich beim Beispiel zweier Atlantikmodelle deren Vorhersagen im räumlichen und zeitlichen Bezug vergleichen. Für den zeitlichen Vergleich müssen beim komplexen Modell Mittelwerte über die Regionen gebildet werden und diese über die Zeit aufgesammelt werden. Diese können dann z.B. in einem einfachen Zeitgraphen (analog zu Abb. 6.40), verglichen werden. Für den räumlichen Vergleich lassen sich die Daten des vereinfachten Modells für jeden Zeitschritt gleichmäßig auf die im komplexen Modell zugehörige Region auftragen. Dann können Darstellungstechniken zur vergleichenden Visualisierung angewendet werden (vgl. Abs. 5.4 sowie Abb. 6.16), wobei insbesondere klar wird, welche Unsicherheiten sich durch die Modellvereinfachung in welchen Regionen ergeben.

Modellausgaben mit Messdaten validieren. Um die Aussagekraft eines Modells bewerten zu können, wird dieses mit realen Messdaten verglichen. Abschnitt 6.1.5 gibt hierzu verschiedene Beispiele, wie Clusteranalyse und Visualisierung am Beispiel der Vorhersage des Maisernteausfalles in Nordostbrasilien zur Modellevaluation eingesetzt werden können.

Überprüfung von Hypothesen. Basierend auf den Modellevaluation können dann insbesondere durch Anwendung von Fehlermaßen, aber auch durch Visualisierungsmethoden, quantifiziert werden, wie gut das neue Modell die Prozesse in den Ausgangsdaten wiedergeben kann.

6.3.2.5 Diskussion

In diesem Abschnitt wurde ein neues Vorgehensmodell zur Kopplung von Modellierungsfunktionalität mit visuellen Data Mining-Techniken entworfen. Damit ist ein erster Ansatz gegeben, um den gesamten Modellierungsprozess interaktiv zu unterstützen, womit der hier vorgestellte Entwurf deutlich über die bisherige Modellierungspraxis hinausgeht. Er bildet die Basis, für das prototypische Framework VisAna.

Abbildung 6.41 zeigt dessen Architektur. Basierend auf dieser Architektur können im Framework Visana Graphen von VDM-Techniken (bzw. Operatoren) verwaltet, mit ihnen graphisch interagiert und diese ausgeführt werden (vgl. hierzu Nocke u. a. 2003). Spezielle Features des Frameworks sind dynamische Parameterdialoge in Abhängigkeit des VDM-Operators von Interesse [21] und die Verwaltung von „erfolgreichen“ Graphen und deren Parameter.

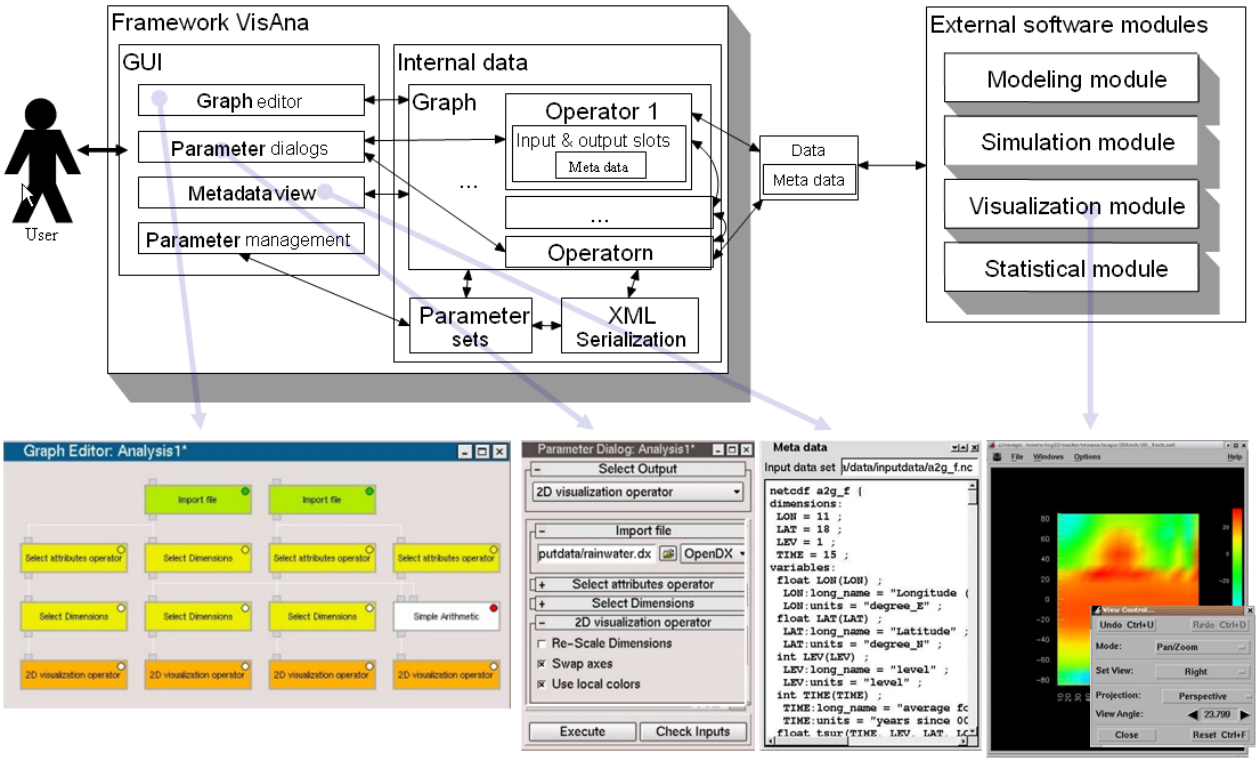


Abbildung 6.41: Architektur-Schema des Framework VisAna: Kombination interner Methoden (Daten- und Metadatenmanagement, Verwaltung von Graphen von VDM-Operatoren), einer graphischer Nutzerschnittstelle (Zugriff auf Operatorgraphen und deren Parameter sowie auf Metadaten) und einer externen Datenschnittstelle zur Ansteuerung einer Vielfalt an Softwaremodulen (Modellierungs- und Simulationssoftware und VDM-Techniken)

Um die Anwendungsmöglichkeiten eines solchen Frameworks beispielhaft zu demonstrieren, wurden im Sinne eines Baukastenssystems verschiedene Schritte in diesem Prozess umgesetzt. Dabei zeigt sich, dass gerade im Bereich der Klimamodellierung vielfältige Fragestellungen durch das VDM unterstützt werden können. Neue Techniken stellt diese Arbeit hierbei insbesondere bei der interaktiv gesteuerten Modellspezifikation (interaktive Parametrisierung von Modellen z.B. über interaktive Spezifikation von Anfangswerten und bei der expliziten Darstellung der Modellstruktur) sowie bei der Modellanalyse (interaktive Exploration von Verfahrensfehlern für gDGLS) vor.

Um den hier vorgestellten Entwurf in seiner Breite umzusetzen, ergeben sich eine Vielzahl von Problemstellungen, die in dieser Arbeit nur in Ansätzen gelöst werden konnten. So unterstützen gängige Systeme diesen Prozess nur zum Teil: es fehlen in Visualisierungssystemen und in Data Mining Systemen die notwendigen Simulationsfunktionalitäten, während Modellierungs- und Simulationsumgebungen zwar zum Teil statistische Methoden beinhalten und Visualisierung an die Simulationsausgabe ankoppeln, jedoch nur begrenzt Visualisierungstechniken direkt in den gesamten Prozess einbeziehen. Weiterhin setzen die in diesem Kontext auftretenden Datenmengen sowie die Zeitanforderungen, die bei der Simulation typisch sind, für eine interaktive Unterstützung Grenzen und erfordern neue Vorgehensweisen.

[21] die alle auf ihn wirkenden Parameter anderer Operatoren darstellen

Auch wenn erste Experimente mit den vorgestellten Techniken das Potential des beschriebenen Ansatzes verdeutlichen konnten, müssen diese weiterentwickelt und deren Integration im Framework VisAna vorangetrieben werden. So ist zu untersuchen, inwieweit sich interaktive VDM-Methoden auch bei der Modelling und Simulation von partiellen Differentialgleichungssystemen, welche typischerweise die Basis von Klimamodellen bilden, einsetzen lassen.

Auch wenn das visuelle Data Mining vor allem auf die Datenexploration fokussiert, kann es auch für die konfirmative Analyse angewendet werden (Shneiderman 2002), gerade um den Modellbildungs- und Simulationsprozess zu unterstützen. Eine systematische Untersuchung von Einsatzmöglichkeiten des VDM in der konfirmativen Analyse steht bisher allerdings noch aus. So wurden erste Ansätze, das VDM auch für die Datenkonfirmation einzusetzen, in den beiden vorangegangen Abschnitten vorgestellt (vgl. z.B. Abs. 6.1.5 zum Vergleich von Clusterungen). Die schließt u.a. ein, Techniken zur Darstellung von Unsicherheiten in diesem Prozess konsequent einzusetzen sowie neue Techniken zur Darstellung von Multi-Run-Simulationsexperimenten zu entwerfen (für erste Ansätze in dieser Richtung vgl. auch Nocke u. a. 2007).

6.4 Zusammenfassung

Aufbauend auf den im vorangegangenen Kapitel vorgestellten Visualisierungstechniken wurden in diesem Kapitel deren enge Kopplung mit automatischen Analyseverfahren untersucht. Herausforderung hierbei ist es, die Visualisierung nicht als statische Ausgabe der Ergebnisse der automatischen Verfahren einzusetzen, sondern die verschiedenen Verfahren in einem interaktiven Prozess miteinander zu koppeln. Das Kapitel hat gezeigt, wie durch diese enge Verknüpfung neue Erkenntnisse über die Resultate automatischer Verfahren, aber auch über deren Zustandekommen gewonnen werden können. Dies verbessert wesentlich das Verständnis der Anwender in die Verfahren und stärkt damit ihr Vertrauen.

Insbesondere wurden hierfür allgemeine Konzepte zur Kopplung von Clusteranalyse und Visualisierung sowie Hauptkomponentenanalyse und Visualisierung entworfen. Hierfür wurde eine Vielzahl von Visualisierungstechniken angepasst und speziell auf die Anforderungen bei der Analyse von Klimadaten zugeschnitten. Bei der visuell gestützten Clusteranalyse wurden neue Methoden zur Darstellung von Clusterzugehörigkeiten und -eigenschaften mit dem Fokus auf deren Darstellung im räumlichen und zeitlichen Bezug vorgestellt. Bei der visuell gestützten Hauptkomponentenanalyse lag der Fokus auf der konsequenten Einbeziehung der Hauptkomponenten in die Schritte des Visualisierungsprozesses und den sich daraus ergebenden Möglichkeiten zur Identifikation und zum Verstehen von Trends und Abhängigkeiten in den Daten.

Weiterhin wurde untersucht, wie der gesamte Prozess der Modellbildung, -simulation und -evaluation durch VDM-Verfahren unterstützt werden kann und hierfür ein allgemeines Vorgehen abgeleitet. Am Beispiel des Entwurfes eines reduzierten Atlantikmodells konnten das Potential der Unterstützung des VDM in diesem Prozess aufgezeigt werden. Neben dem Einsatz von VDM-Methoden zur Datenexploration bezieht dies insbesondere auch die Datenkonfirmation ein, für welche diese enge Kopplung von visuellen und automatischen Verfahren noch wenig erforscht ist.

Darüber hinaus kann die Analyse von Klimadaten und die Unterstützung bei Modellierung von Klimamodellen auch von weiteren automatischen Methoden profitieren. Über die Extraktion von Clustern und Hauptkomponenten hinaus schließt dies jede Art von Berechnung von Metadaten über die Daten mit ein, die dann zur Unterstützung des Visualisierungsprozesses eingesetzt werden können (vgl. Abs. 7.1).

Weiterhin können Darstellungen von Klimadaten durch Anreicherung mit zusätzlichen Daten profitieren. So kann das Verständnis über die Ursachen klimatischer Phänomene in räumlichen Klimadarstellungen durch deren Anreicherung mit geographischen Informationen verbessert werden. Problem

gerade bei Höhenkarten ist es, für einen bestehenden Datensatz eine passende Höhenkarte mit einer geeigneten Auflösung zu erhalten. Hierfür wurde das Werkzeug WorldMapTool[22] entwickelt, welches es erlaubt, für eine beliebige, rechteckige geographische Region eine Höhenkarte mit geeigneter Auflösung zu extrahieren[23]. Durch ein visuelles Interface wird der Anwender bei der Interaktion mit der Höhenkarte unterstützt (vgl. Abb. 6.42). Diesem Werkzeug liegt eine Multiresolution-

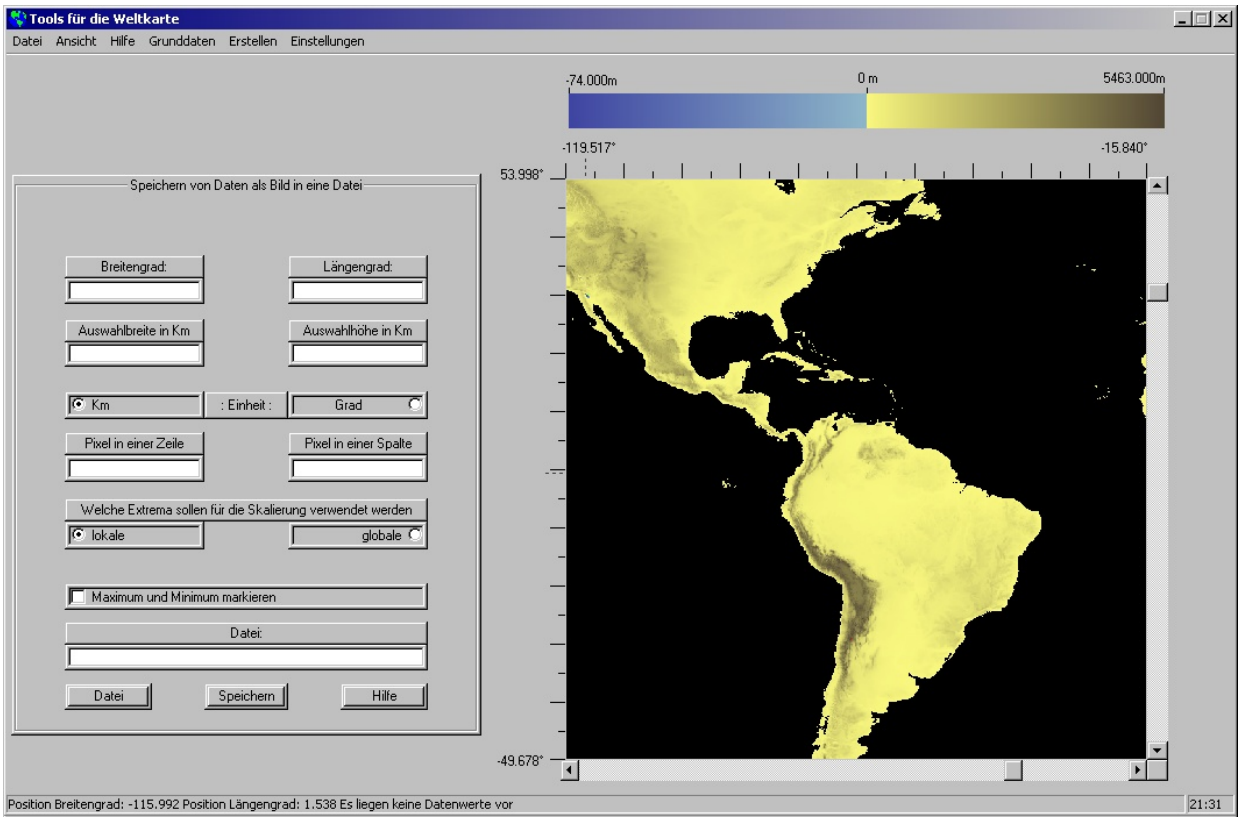


Abbildung 6.42: Auswahl eines Höhenkartenabschnittes mit dem WorldMapTool

Hierarchie[24] zugrunde, in der durch trilineares Filtering beliebige Höhenkartenauflösungen für das gewählte Gebiet berechnet werden können. Um flexibel auf Änderungen der Höhenkarte reagieren und bei Bedarf höher aufgelöste Karten einbeziehen zu können, kann der Anwender auch neue Kartensegmente in die Hierarchie einfügen, welche dann dynamisch aktualisiert und ggf. um neue Hierarchiestufen erweitert wird.

Probleme ergeben sich angesichts der Heterogenität der vorgestellten VDM-Methoden. Die Integration von zum Teil in unterschiedlichen Systemen oder Programmiersprachen umgesetzten Modulen ist eine besondere Herausforderung für ein allgemeines, anwenderfreundliches Framework. Weiterhin ergibt sich insbesondere das Problem der Vielfalt an Methoden. Um den Anwender bei deren Auswahl und Parametrisierung in Abhängigkeit von seinen Aufgabenstellungen und der Datencharakteristika zu unterstützen, wird im folgenden Kapitel eine neuartige Methodik entworfen.

[22] Die Umsetzung erfolgte im Rahmen eines betreuten Studentenprojektes (Thiede u. Krüger 2004).

[23] basierend auf dem GTOPO30-Datensatz mit einer Auflösung von ca. 1km × 1km bei einer Gittergröße von 43200 x 22800; http://edc.usgs.gov/products/elevation/gtopo30/gtopo30.html

[24] Um eine hohe Speichereffizienz zu erreichen, werden die einzelnen Hierarchiestufen separat komprimiert.

Kapitel 7

Visualisierungsdesign

Die zunehmende Komplexität und Größe von erhobenen Datensätze auf der einen Seite und die Vielzahl hierfür entworfener (visueller) Analysetechniken auf der anderen Seite erfordert neue Methoden der Anwenderunterstützung bei der Auswahl und Parametrisierung dieser Techniken. North u. a. (2002) stellen fest, dass moderne relationale Datenbanken bereits über ein fortgeschrittenes Datenmanagement verfügen, während heutige Visualisierungstechniken häufig von nur einem einzigen Programmierer und für ein spezifisches Problem entworfen werden. Auch wenn diese Aussage sich nicht auf innovative, flexible und anpassbare Visualisierungsumgebungen bezieht, trifft es zu, dass viele Visualisierungstechniken den Kontext der Untersuchung nicht adäquat einbeziehen. Dies führt dazu, dass Anwender Informationen dargestellt bekommen, die nicht ihrem aktuellen Interesse entsprechen. Insbesondere lassen sich in diesem Zusammenhang die Datencharakteristika und die aktuellen Analyseziele unter Einbeziehung der Spezifika der Anwendung einsetzen, um geeignete Darstellungen zu generieren. Dies ermöglicht, die Lücke zwischen interner Technikimplementation, dem Visualisierungsexpertenwissen und dem Kontextwissen des Anwenders zu schließen.

Basis bei der automatisch unterstützten Generierung von Darstellungen ist es, dass die hierfür relevanten Einflussfaktoren beschrieben und erhoben werden. Insbesondere werden in der Literatur eine Vielzahl von für den Visualisierungsprozess relevante Datencharakteristika isoliert betrachtet. Eine einheitliche Beschreibung und eine systematische Untersuchung von Methoden zur Erhebung solcher Metadaten liegt bisher nicht vor. Analog liegen eine Vielzahl von Spezifikationen von Analysezielen und -aufgaben vor. Auch wenn Ansätze zu deren Integration existieren, so ist die anwendungsspezifische Formulierung von Zielen und deren Abbildung auf anwendungsunabhängige Ziele ein noch immer offenes Problem.

Entsprechend wird in diesem Kapitel eine allgemeine Beschreibung von Metadaten vorgestellt sowie Ansätze zu deren Erhebung diskutiert (Abs. 7.1). Im Anschluss daran wird ein neuer Ansatz zur Beschreibung von Analysezielen vorgestellt (Abs. 7.2). Basierend auf diesen wesentlichen Einflussfaktoren wird dann ein pragmatischer Ansatz zur halbautomatischen Auswahl und Parametrisierung von Visualisierungstechniken vorgestellt, welcher sich an den speziellen Anforderungen der Klimafolgenforschung orientiert (Abs. 7.3).

7.1 Metadaten für das visuelle Data Mining

Im Datenbankumfeld haben sich einheitliche Standards für die Datenhaltung und für Schnittstellen zum Zugriff auf die Daten durchgesetzt. Im Unterschied dazu finden sich im Visualisierungsumfeld eine Vielzahl von zum Teil proprietären Formaten. Dies hat zur Folge, dass die Anwender von Visualisierungssoftware den Im- und Export ihrer Daten je nach System neu umsetzen müssen. Zwar

bieten gängige Visualisierungssysteme eine Unterstützung bei der Beschreibung des Datenformates eines Datensatzes (z.B. der Data Prompter in OpenDX) und liefern auch Einleseroutinen für Standardformate (z.B. NetCDF), jedoch sind die ersteren zum Teil relativ kompliziert zu erlernen und die zweiteren häufig nicht für die speziellen Bedürfnisse der Anwendung ausreichend.

Die Einbindbarkeit der eigenen Daten stellt somit einen begrenzenden Faktor für den Einsatz von Visualisierungssystemen dar, was z.B. dazu führt, dass im Klimaumfeld neben der Benutzung von Standardtools (z.B. GMT, Grads, OpenDX sowie GIS-Systemen) gerade auch „General Purpose"-Systeme mit beschränkter Visualisierungs- und Interaktionsfunktionalität wie Matlab und Excel vielfach verwendet werden.

Weiterhin ist es für Visualisierungslaien häufig nicht leicht, sich in der heterogenen Terminologie in der Visualisierungsliteratur zurechtzufinden. So werden z.B. je nach Umfeld die Begriffe Variable, Dimension und Merkmal sowie die von ihnen aufgespannten Räume synonym verwendet. Zwar existieren Notationen (z.B. Brodlie 1992; Wong u. Bergeron 1997), welche grundlegende Dateneigenschaften beschreiben, jedoch sind diese Notationen relativ abstrakt, was ihre Verbreitung in der Praxis eingeschränkt hat. Um dem Problem mehrdeutiger Terminologie zu begegnen, wurden erste Anstrengungen unternommen, um die Visualisierungsterminologie zu vereinheitlichen (vgl. Brodlie u. a. 2004).

Darüber hinaus stellen die Daten eines Datensatzes nur einen Ausschnitt aus dem zu untersuchenden Sachverhalt dar. Lux (1998) spricht in diesem Zusammenhang von einem Eisberg, bei dem die Daten den sichtbaren Teil ausmachen, und sich eine Vielzahl von Ebenen darunter befinden (Formeln, örtlicher und zeitlicher Kontext, Beziehungen und Abhängigkeiten, abgeleitete Werte, Hierarchien auf den Daten, Abstraktionen, Simulationsebene). Das Einbeziehen dieser Ebenen in den Analyseprozess ist häufig essentiell für dessen erfolgreiche Bearbeitung (vgl. hierzu auch Westphal u. Blaxton 1998).

In diesem Umfeld setzt die hier vorliegende Arbeit an, um bei Bedarf geeignete Daten zusätzlich zu den vorliegenden Daten (so genannte Metadaten) bereitzustellen, um die Datenanalyse und das Datenmanagement auf den verschiedenen Ebenen zu unterstützen (vgl. hierzu auch Nocke u. Schumann 2002). Eine geschlossene, allgemeine Metadatenbeschreibung ist im Visualisierungsumfeld bisher noch weitgehend offener Forschungsgegenstand, da Metadaten in der Literatur überwiegend isoliert für bestimmte Datenklassen betrachtet werden (z.B. GIS, Strömungsdaten,...). Dazu stellt diese Arbeit - über die genannten Ansätze im Umfeld der Visualisierung hinaus - eine allgemeine, erweiterbare Beschreibung von Metadaten bereit, welche die vielfältigen Entscheidungen im Rahmen des Designs und des Managements des gesamten VDM-Prozesses unterstützen (vgl. Abs. 7.1.2).

Insbesondere hat eine geschlossene Beschreibung von Metadaten und deren Einsatz im Umfeld des VDM ein großes Potential bei der Schaffung von Standards für den Datenaustausch, für das Softwaredesign von Visualisierungs- und VDM-Systemen, für die prinzipielle Einschätzung den Anwendbarkeit von VDM-Techniken sowie deren Kopplung und für vielfältige Arten der Anwenderunterstützung (Auswahl und Parametrisierung von VDM-Techniken, explizite Darstellung von Metadaten, effektive Speicherung und Wiederherstellung von Mining-Prozessen, anwendungsabhängige Spezifikation und Bezeichnung von Dateneigenschaften).

Dabei fokussiert diese Arbeit nicht auf die Definition eines neuen Datenformates, sondern auf die Spezifikation von relevanten Metadaten für das visuelle Data Mining und Wege zu deren Erhebung. Um die Notwendigkeit des Einsatzes von Metadaten zu verdeutlichen, sei hier auf das Beispiel der Modellausgaben eines Klimamodells verwiesen (vgl. Petoukhov u. a. 2000). So können die simulierten Daten eines solchen Modells in unterschiedlichen räumlichen Dimensionalitäten (1D, 2D, 3D) mit unterschiedlichen Gitterstrukturen die Dynamik von Ozeanen, Atmosphäre, Vereisung u. a. ausdrücken. Die Daten sind hierbei Zeitreihen, welche auch mit unterschiedlichen Granularitäten auftreten können. Dabei können pro Beobachtungsfall bis zu 40 Merkmale auftreten (vgl. z.B.

Petoukhov u. a. 2000). Dieses Beispiel demonstriert, dass eine umfassende und systematische Behandlung von Metadaten wünschenswert ist (um z.B. in diesem Fall verschiedenen Teilmodelle visuell gemäß ihrer räumlichen und zeitlichen Skalen verknüpfen zu können).

Über die einheitliche Beschreibung von Metadaten hinaus ist im speziellen die effektive Erhebung von Metadaten ein Problem. Hierbei muss der Fragestellung nachgegangen werden, welche Metadaten zu welchem Zeitpunkt des VDM-Prozesses benötigt werden und wie die Erhebung dieser Metadaten effektiv durchgeführt werden kann. Neben der Beantwortung der Frage, ob ein Metadatum prinzipiell (effektiv) erhoben werden kann, ist dafür zu untersuchen, inwieweit Metadaten in einer geeigneten Reihenfolge unter der Einbeziehung des Nutzer in diesen Erhebungsprozess erfolgen kann. Hierzu wurde ein Framework zur (semi-)automatischen Erhebung und Speicherung von Metadaten entworfen und umgesetzt (vgl. Abs. 7.1.3). Der Ansatz hierbei ist es, Metadaten in einer interaktiven Abfolge von Schritten zu verfeinern. Dabei werden in einer ersten Darstellung bestimmte Eigenschaften der Daten präsentiert. Basierend auf diesen Metadaten und dem Wissen des Anwenders lassen sich dann bei Bedarf Metadaten erheben bzw. verfeinern. Hierdurch wird graduell neues Wissen über die Daten generiert. Insbesondere erfordern hierbei verschiedene Komplexitäten von Metadatenerhebungsschritten die Einbeziehung von verschiedenen Automatisierungsgraden sowie einer Zeitkontrolle für langwierige Erhebungsalgorithmen.

7.1.1 Stand der Forschung

Allgemein lassen sich Metadaten nach Inhalt, Kontext und Struktur einteilen (vgl. Gilliland-Swetland 2000). Robertson u. Hutchins (1997) unterteilen speziell für das Umfeld der Visualisierung in *beschreibende*, *abgeleitete* und *historische* Metadaten. *Beschreibende* Metadaten, spezifizieren die zugrunde liegenden Eigenschaften der Daten, welche deren Speicherstruktur und den Zugriff darauf bestimmen. Metadaten, die in irgend einer Art aus den Daten extrahiert wurden, werden als *abgeleitete* Metadaten bezeichnet. *Historische* Metadaten geben Aufschluss über den Ursprung der Daten, wozu z.B. Fehler und Unsicherheiten bei ihrer Erhebung gehören.

Bisher wurden im Visualisierungsumfeld überwiegend *beschreibende* Metadaten eingesetzt. Einige Arbeiten beschäftigen sich jedoch auch mit der Erhebung und dem Einsatz von wichtigen Eigenschaften der Daten zu deren kompakter Beschreibung. Hierzu zählen Arbeiten aus dem Gebiet der Feature Visualisierung (vgl. z.B. Silver 1997; Reinders u. a. 2001; Weinkauf u. a. 2004). Die Idee hierbei ist es, die relevanten Muster eines Datensatzes anstelle oder in Kombination mit den Originaldaten darzustellen. Bei diesen Mustern handelt es sich um *abgeleitete* Metadaten.

Notationen und Taxonomien. Darüber hinaus wurden verschiedene Notationen und Taxonomien zur Beschreibung der Dateneigenschaften vorgestellt (vgl. Schumann u. Müller 2000, für eine Übersicht). Hierzu zählen die Notationen von Bergeron u. Grinstein (1989) (*Anzahl der Merkmale, Dimensionalität des Gitters, Verbund der Beobachtungspunkte*), Brodlie (1992) (*Anzahl der Merkmale, Dimensionalität des Gitters, Datentyp, Skalentyp, Wirkungskreis*) und Wong u. Bergeron (1997) (*Anzahl der Merkmale, Dimensionalität des Beobachtungsraumes*). Graw u. a. (1997) erweitern diese Notationen um die *Anzahl der Beobachtungsfälle, die Qualität der Datenmenge, den Umfang des Wertebereiches* und *um zeitliche Eigenschaften der Daten.*

Im Unterschied zu diesen formalen Taxonomien dient die Task-by-Data-Type-Taxonomy von Shneiderman (1996) zur Einteilung in der Praxis auftretender grundlegender Datenklassen (*zeitlich, 1D, 2D, 3D, Multi-D, Baum, Netzwerk, Arbeitsraum*). Hierbei werden insbesondere *Eigenschaften des Beobachtungsraumes, des Merkmalsraumes* und *Beziehungen zwischen den Datenobjekten* einbezogen.

Die bisher beschriebenen Notationen sind zum Teil relativ abstrakt und auf wichtige Datencharakteristika fokussiert. Sie erlauben es, relevante Dateneigenschaften kompakt zu beschreiben und legen

die Basis für die Definition bestimmter Datenklassen. Zur konkreten Beschreibung einer breiten, erweiterbaren Palette zur Unterstützung und Steuerung des VDM-Prozesses sind diese jedoch nur begrenzt geeignet. Im folgenden sollen deswegen allgemeine Taxonomien skizziert werden.

Zhou u. Feiner (1996) definieren eine (objektorientierte) Taxonomie von Datencharakteristika mit dem Ziel, auch sehr heterogene Datenmengen beschreiben zu können. Dabei unterscheiden Sie die folgenden sechs Dimensionen: Datentyp, Datendomäne, Datenattribute, Datenrelationen, Datenrolle und Datenbedeutung. Mit dieser Taxonomie werden neben der Datenbeschreibung (Datentyp, Datenattribute), auch Beziehungen zwischen den Daten (Datenrelationen) sowie semantische Informationen über Erhebung und Bedeutung der Daten mit einbezogen (Datendomäne). Besonders an dieser Taxonomie ist, dass eng mit den Daten gekoppelt Analyseziele und Nutzerpräferenzen einbezogen werden (Datenrolle und Datenbedeutung in einem Anwendungszusammenhang). Auf diese Art können allgemeingültig Daten beschrieben werden (z.B. zeitabhängige, gemessene, kontinuierliche „formlose“ Wasserbewegung als Teil des Ozeans).

Darüber hinaus finden sich in der angrenzenden Literatur verschiedene Taxonomien (z.B. Arens u. a. 1993; Steinacker u. a. 2001).

Metadaten in Visualisierungssystemen. In modernen Visualisierungssystemen lassen sich Daten verschiedener Charakteristika einlesen und verarbeiten. Dabei konzentrieren sich die dabei verwendeten Datenstrukturen und Datenformate im wesentlichen auf beschreibende Metadaten. So werden im IBM DataExplorer Daten und Metadaten in einem Dateiformat gespeichert, wobei flexibel verschiedene Daten- und Gittertypen verwaltet und für die Visualisierung erforderliche Metadaten hinzugefügt werden (vgl. Abram u. Treinish 1995). Ein weiteres Beispiel für die gemeinsame Verwaltung von Daten und Metadaten findet sich im System InfoVis (vgl. Fequete 2004). Hier werden unter anderem spezielle Charakteristika von Bäumen und Graphen in Kombination mit den Daten abgespeichert.

Metadaten im Datenbankumfeld. Im Umfeld von Datenbanken ist der Einsatz von Metadaten weit verbreitet. Hierbei werden neben den (Typ-)Eigenschaften von einzelnen Datenbankspalten auch Beziehungen zwischen Datenbanktabellen beschrieben. Neben der direkten Speicherung auch abgeleiteter Metadaten in der Datenbank (z.B. Mittel- oder Extremwerte von Zeitreihen in Wrobel 2004) hat sich als Austauschformat in diesem Umfeld das XML-Format etabliert, welches es ermöglicht, flexibel Metadaten im Ascii-Format zu beschreiben. Viele solcher Metadaten lassen sich auch für die Steuerung des VDM einsetzen, allerdings decken sie die erforderlichen Informationen für ein gestütztes Mining nur zum Teil ab, da sie insbesondere nicht speziell auf die Datenexploration hin ausgerichtet sind.

Metadaten zum Einsatz im Visualisierungsdesign. Die Spezifikation von Metadaten hat viele Vorteile für das Datenmanagement in Visualisierungs- und VDM-Systemen. Darüber hinaus finden sich in der Literatur eine Vielzahl von Arbeiten, die darauf abzielen, das Wissen über die Datencharakteristik zur Steuerung des Visualisierungsprozesses mit dem Ziel geeigneter Darstellungen einzusetzen.

Eine der ersten Arbeiten in diesem Umfeld findet sich bei Mackinlay (1986). Dieser Ansatz konzentriert sich vor allem auf Daten mit *quantitativem Skalentyp*. Roth u. Mattis (1990) erweiterten diesen Ansatz unter Einbeziehung weiterer, auch komplexerer Dateneigenschaften (*Datentyp, Wertebereichseigenschaften, Fehler, Beziehungen zwischen den Relationen*). Bei diesen Ansätzen liegt der Fokus bei der Datenbeschreibung von relationalen Daten zur Findung geeigneter 2D-Darstellungen für diese Daten.

Eine allgemeinere Einteilung von Datencharakteristika zur automatischen Findung geeigneter Darstellungen findet sich bei Wehrend u. Lewis (1990) (z.B. *Form, Ort oder Struktur*). Besonders an diesem Ansatz ist, dass hierbei die Eigenschaften der Daten aus Sicht der Analyseziele behandelt werden, zum Beispiel, ob der Anwender bestimmte, sich durch die *Verteilung der Daten* herausbil-

dende *Formen* finden oder versteckte *Strukturen* in den Daten aufdecken will (vgl. hierzu auch Abs. 7.2).

Eine allgemeine Beschreibung von Metadaten mit dem Ziele der Steuerung der Visualisierung liefert die Arbeit von Lange (2006), die sich vor allem auf Metadaten für multivariate Daten konzentriert. Neben der *Anzahl der Variablen* und *der Beobachtungsfälle* sowie der Beschreibung der *Wertebereichseigenschaften (Skalentyp, Umfang)* werden auch die *Qualität der Daten* und die *Struktur der Datenmenge* mit in die Beschreibung der Datencharakteristik einbezogen. Insbesondere finden sich bei Lange (2006) eine Vielzahl von Metadaten für die *Struktur der Datenmenge*, Metadaten zur Beschreibung des *Beobachtungsraumes* (z.B. *Gittereigenschaften* wie *Dimensionalität, Verbund* und *Wirkungsbereich der Gitterpunkte*) und *funktionale Abhängigkeiten der Variablen.*

Neben diesen allgemeinen Ansätzen wurden eine Vielzahl von Metadaten beschrieben, um spezielle Probleme beim Visualisierungsdesign zu unterstützen. So beschreiben Andrienko u. Andrienko (1999) zur Unterstützung der visuellen Analyse geographischer Daten ein objektorientiertes Metadatenmodell. Bergman u. a. (1995) benutzen *Frequenz* und *Datentyp* zur Findung einer geeigneten Farbabbildung. Um ein Ranking von Scatterplots durchzuführen, werden von Shneiderman (2002) *Korrelationen zwischen den Variablen* benutzt. Um Beziehungen von Datensätzen einer Datenbank expressiv darzustellen, definieren Golovchinsky u. a. (1995) essentielle Metadaten für Graphen und Bäume (u. a. *Eigenschaften des Wertebereiches* von Merkmalen, die den Graph bilden und den Eigenschaften des Graphen (u. a. *gerichtet, symmetrisch, zyklenfrei*)) und benutzen diese für ein konstruktives Visualisierungsdesign.

Metadaten für spezielle Datenklassen. Die Arbeit von Golovchinsky u. a. (1995) ist ein erstes Beispiel für eine Vielzahl von Verfahren, welche sich auf die Extraktion und den Einsatz der Datencharakteristika spezieller Datenklassen in der Visualisierung konzentrieren. Gerade im Bereich der Strömungsvisualisierung ist eine typische Vorgehensweise, Charakteristika von Strömungfeldern (*kritische Punkte, separierende Linien und Flächen, Wirbel, Schockwellen, Trennflächen zwischen Flüssigkeiten*) zu extrahieren und darzustellen (vgl. z.B. Pagendarm u. Seitz 1993; Silver 1997; Jiang u. a. 2002; Mann u. Rockwood 2002; Bennet u. a. 2003; Weinkauf u. a. 2004). Auch in der Volumenvisualisierung werden die *Oberflächen von charakteristischen Körpern* in den Daten segmentiert und verschiedenartig in der Visualisierung dargestellt (vgl. z.B. Kanitsar u. a. 2001; Wang u. Mueller 2004; Bade u. a. 2006; Iserhardt-Bauer u. a. 2006).

Metadaten zur Beschreibung der Datenqualität. Die Einbeziehung der Datenqualität als entscheidendes Datencharakteristikum in den Analyseprozess ist ein erst in Ansätzen erforschtes Gebiet. Viele Anwendungen (z.B. Wetter- und Klimavorhersagen) stellen noch immer Mess- und Simulationsergebnisse dar, ohne die Datenqualität explizit in die Visualisierung einzubeziehen. Auch wenn auf diesem Gebiet noch Forschungsbedarf besteht, zeigen erste Ansätze, wie fehlerhafte Daten und Unsicherheiten in den VDM-Prozess einbezogen werden können (vgl. Lodha u. a. 1996; Pang u. a. 1997; Cedilnik u. Rheingans 2000; Djurcilov u. a. 2001; Griethe u. Schumann 2005).

Metadaten für Wetter- und Klimadaten. Darüber hinaus haben die gemessenen bzw. simulierten Daten spezieller Anwendungen zugeschnittene Methoden zu deren Beschreibung. So ist neben der Art, dem Typ und dem Ursprung des Modells und dessen Version auch die Art der Modellanregung für die Bewertung und Reproduzierbarkeit einer Klimasimulation von großer Bedeutung. Analoges gilt auch für die Umstände einer Wetterbeobachtung.

Ein Beispiel für die Speicherung verschiedenartiger Metadaten in einem Data-Ware-House-artigen System zum Management klimatischer Daten ist die Arbeit von Wrobel (vgl. Wrobel 2004, S. 200ff). Zu den dort beschriebenen Zeitreihenmetadaten gehören neben allgemeinen Information zur Erhebung der Daten auch für die Visualisierung interessante Metadaten, wie z.B. zeitliche Auflösung und räumliche Verteilung der Messdaten, Einheiten von Merkmalen oder Zuordnung der Daten zu Kontinenten oder Flusseinzugsgebieten (vgl. Wrobel 2004, S. 205ff). Solche Metadaten eignen sich

insbesondere, um eine flexible, effektive Suche auf großen Klimadatenbeständen durchführen zu können.

Daneben existieren eine Vielzahl von Arbeiten, die interessante Muster und Ereignisse aus Wetter- und Klimadaten ableiten, wie sie z.B. für Seefahrtsämter und Luftfahrtgesellschaften von Interesse sind (vgl. Schröder 1997, S.18). So sollen Vereisungsgefahr, Luftturbulenzen oder Stürme aus Wettersimulationen abgeleitet werden. Um die Vielzahl an interessanten oder warnwürdigen Ereignissen zu finden, wurden verschiedene automatische Analyse- und Filtermodule entwickelt. So benutzen Ribarsky u. a. (2002a) eine semantisch angereicherte Hierarchie zur Speicherung von zeitveränderlichen 2D- und 3D-Wetterdaten aus verschiedenen Quellen (Doppler-Radar-Messung, Satellitenbilder, ...). Zusätzlich wird diese mit gewissen Wetterereignissen angereichert, die automatisch bei der Erfassung der Daten in die Hierarchie eingefügt werden. Hierzu gehören die Überschreitung gewisser Windgeschwindigkeiten oder Niederschlagswerten in Zyklonen, aber auch Form, Ausdehnung und Trajektorien von Sturmfronten über die Zeit sowie auf diesen Mustern abgeleitete Kandidaten für Tornados (vgl. auch Eilt u. a. 1995). Durch die damit verbundene Anreicherung der Daten können verschiedene Datenquellen (Doppler-Radar-Messung, Satellitenbilder, ...) zu einem einzigen Wetterereignis gruppiert werden. Extreme Ereignisse werden längerfristig gespeichert. Durch Umsetzung der Hierarchie als komprimierter Wald von verschachtelten Quadtrees und Octrees, können eine Vielzahl von flexiblen Anfragen effizient durchgeführt werden.

Ein anderes Beispiel für abgeleitete Metadaten im Klimaumfeld sind aus einem atmosphärischen Strömungsfeld extrahierte Windrichtungen, die von Macêdo u. a. (2000) in einem Scatterplot dargestellt werden. Wong u. a. (2000) extrahieren mit Hilfe eines geschwindigkeitsbasierten Filters aus Wetterdaten kritische Punkte, in deren Umgebungen besondere Eigenschaften wie Scherkräfte und Zirkulationen ein Zeichen für potentielle Wetterinstabilitäten sind. Moorhead u. Zhu (1993) identifizieren und verfolgen Wirbel und Wetterfronten durch einen 3D-Kanten-Operator. Ma u. Smith (1993) stellen ein Verfahren zur Verfolgung von Wolken unter Einsatz der Lagrange-Methode vor.

Erhebung von Metadaten. In der Literatur findet sich eine Vielzahl von Verfahren zur Extraktion bestimmter Features (s.o.). Diese werden jedoch zumeist isoliert betrachtet (vgl. Absatz *Metadaten für spezielle Datenklassen*). Ein allgemeines Vorgehen zur datenklassenunabhängigen Extraktion von Metadaten im Umfeld des VDM gibt es bisher nicht. Ein erster systematischer Ansatz hierfür wurde in einer parallel laufenden Dissertation von Lange (2006) beschrieben. Hier werden drei grundlegende Problemkreise diskutiert, welche die „Unvollständigkeit einer Datenbeschreibung“ verursachen können (Lange 2006, S. 86):

- „*Erhebung der Metadaten:* Die Dateneigenschaften sind nicht bekannt, nicht erfasst, schwierig zu ermitteln oder selbst erst Ziel der visuellen Analyse.
- *Formulierung der Metadaten:* Der Nutzer versteht die Bedeutung (Sinn, Zusammenhänge, Auswirkungen) gewisser Dateneigenschaften und ihrer Beschreibungsformen nicht oder nicht richtig. Oder Dateneigenschaften, die gewisse Interpretationen voraussetzen [...] sind nicht eindeutig beschreibbar, weil das zugrundeliegende Phänomen oder Modell nicht vollständig bekannt oder nicht verständlich ist.
- *Speicherung und Verwaltung der Metadaten:* Die verwendeten Datenformate und Datenverwaltungswerkzeuge lassen eine Speicherung von Metadaten nicht oder nicht in ausreichendem Maße zu. Oder sie erfordern eine Strukturierung der Daten, die deren ursprünglicher, inhärenter Struktur nicht ausreichend gerecht wird und folglich die Ableitung von (strukturbeschreibenden) Metadaten erschwert.“

Aus diesen Problemen leitet sich die Notwendigkeit ab, flexibel Metadaten zu erheben und zu speichern, sowie verschiedene Methoden und Datenformate in einem Werkzeug verfügbar zu machen (vgl. Lange 2006). Ferner identifiziert Lange (2006) als wichtige Quellen für die Erhebung von

Metadaten *die Daten* selbst, *das Datenformat* mit ggf. vorliegenden Metadaten und den *Nutzer* als Kenner bzw. Erzeuger der Daten. In Anlehnung und konsequente Weiterführung an die Ideen dieser Arbeit werden in Abschnitt 7.1.3 Methoden zur Metadatenerhebung und deren Integration in ein Werkzeug zur Metadatenerhebung systematisch vorgestellt.

Visualisierung und Interaktion von/mit Metadaten. Neben dem Datenmanagement und der Steuerung des VDM-Prozesses lassen sich Metadaten auch direkt darstellen und interaktiv explorieren. So führen dos Santos u. Brodlie (2004) zwei neue visuelle Werkzeuge ein, welche es neben einer schnellen Übersicht über wichtige Metadaten auch erlauben, dem Nutzer das Daten-Filtering zu erleichtern. Dafür werden Metadaten zu den Variablen in zwei interaktiven Visualisierungen dargestellt. Im „Interaction Graph“ werden alle Variablen in einem vollständig vernetzten Graph (Clique) dargestellt, und der Nutzer kann mittels Selektion von Variablen-Knoten Teilräume des Datenraumes auswählen. Mit dem „n-dimensional Window“ werden alle Variablen und ihre Wertebereiche in einer Art von Sternikone dargestellt. Mittels Interaktion mit Kontrollpunkten auf den sternförmigen Achsen kann er dann sowohl die darzustellenden Wertebereiche der Variablen als auch eine Region von Interesse auswählen. Die Methode kann sowohl zur Filterung von Dimensionen (Reduktion des Beobachtungsraumes), als auch von Merkmalen (Reduktion des Merkmalsraumes) verwendet werden. Vorteil dieser Art von Datenfilterung ist deren hohe Transparenz, die eine mentale Vorstellung der Variablen des Datensatzes sowie ihrer Ausprägungen vermittelt, und es ihm zum anderen ermöglicht, “smooth“ durch höherdimensionale Datenräume zu navigieren und dabei die Beziehung von Räumen und Unterräumen zu erhalten.

Weitere Beispiele zur expliziten Darstellung von Datencharakteristika (neben den Ansätzen zur Feature-Visualisierung) finden sich bei Taylor u. Bendford (1998) (Veranschaulichung von *Beziehungen in einem Datenbanksystem*), Yang u. a. (2005) (*Anordnung der unabhängigen Variablen zur Exploration von Abhängigkeiten*) sowie im System GeoVISTA (2007) (explizite *Entropie- und Korrelationsvisualisierung*). Eine systematische Untersuchung, welche Metadaten sich durch ihre Visualisierung im Analyseprozess direkt einsetzen lassen, ist bisher noch weitgehend offener Forschungsgegenstand.

Resümierend lässt sich feststellen, dass eine Vielzahl von Metadaten in verschiedenen Bereichen des VDM eingesetzt werden. Insbesondere wurde die Erhebung und der Einsatz von bestimmten Datencharakteristika zur Beschreibung von Datenformaten, für spezielle Datenklassen und Anwendungen untersucht. Eine geschlossene Beschreibung, oder sogar eine Standardisierung von Metadaten für Visualisierungs- und Miningzwecke - wie im Datenbankumfeld etabliert - existiert jedoch nicht. Dies resultiert vor allem aus der Heterogenität der Daten aus verschiedenartigen Anwendungsgebieten.

In diesem Problemfeld setzt die hier vorliegende Arbeit an, um Ansätze zur Metadatenbeschreibung und -erhebung aus existierenden Ansätzen mit dem Fokus auf speziellen Aspekten zu integrieren. Das Ziel der im folgenden zu entwerfenden Metadatenbeschreibung ist es, allgemeine, grundlegende Metadaten zusammenzufassen und diese (exemplarisch) um spezielle Metadaten für bestimmte Datenklassen anzureichern, die flexibel erweitert werden können. Dabei sollen neben der gängigen Einbeziehung von beschreibenden Metadaten insbesondere abgeleitete Metadaten systematisch eingeordnet werden, um eine praktikable Grundlage zum Management des VDM-Prozesses zu legen.

7.1.2 Eine allgemeine Spezifikation von Metadaten für das Visuelle Data Mining

Für eine allgemeine Spezifikation von Metadaten speziell für das VDM wurden die folgende Konzepte identifiziert:

- **Separierung:** Trennung von Metadaten, die für beliebige Datensätzen gültig sind von Metadaten, die nur für spezifische Datenklassen relevant sind

- **Integration:** Einbeziehung aller Arten von Metadaten (*beschreibende*, *abgeleitete* und *historische* Metadaten)
- **Strukturierung:** Einsatz von Metadaten für die Beschreibung der internen Struktur des Datensatzes
- **Validität und Unsicherheit in den Daten:** Einbeziehung von Metadaten zur Beschreibung von Fehlern und von Unsicherheiten (Datenqualität)
  - der Daten
  - der Analysetechniken und der *abgeleiteten* Metadaten
- **Erweiterbarkeit:** Möglichkeit, Metadaten zu erweitern, z.B durch
  - Metadaten für gekoppelte Datenmengen
  - Metadaten für neue Datenklassen
- **Nutzerorientierung:** Einbeziehung von Metadaten für verschiedener Arten von Nutzerinteraktion:
  - Einbeziehung von Metadaten, welche sowohl abhängig als auch unabhängig von speziellen Anwenderprofilen und Analysezielstellung sind
  - Einbeziehung von Metadaten unter Einbringung von Anwenderwissen

7.1.2.1 Beschreibung von grundlegenden Metadaten

Basierend auf diesen Kriterien soll im folgenden ein neuer Ansatz zu einer allgemeinen Spezifikation grundlegender, weitgehend von speziellen Anwendungen unabhängiger Metadaten beschrieben werden. Hierbei werden neben den typischerweise im Vordergrund stehenden *beschreibenden* Metadaten auch eine Vielzahl in der Literatur isoliert betrachteten *abgeleiteten* Metadaten sowie wichtige *historische* Metadaten einbezogen.

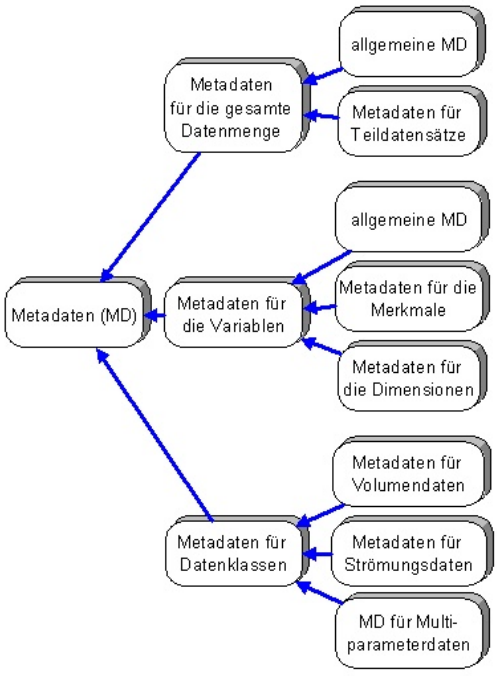


Abbildung 7.1: Hierarchie der Metadaten

Die Spezifikation der *beschreibenden* Metadaten orientiert sich hierbei an den Metadaten etablierter Notationen (vgl. Bergeron u. Grinstein 1989; Brodlie 1992; Wong u. Bergeron 1997; Graw u. a. 1997) und den Prinzipien der Metadatenbeschreibung im NetCDF-Format (vgl. Rew u. a. 1993). Davon ausgehend sollen Metadaten - auch aus praktischen Erwägungen geleitet - wie folgt systematisiert werden (vgl. Abb. 7.1): Ein Datensatz besteht aus Werten verschiedener Variable. Es werden Metadaten benötigt, um diese Variablen, grundlegendende Eigenschaften des gesamten Datensatzes (Beziehungen zwischen Variablen und Werten) und die speziellen Eigenschaften von in den Daten enthaltenen Datenklassen zu beschreiben. Entsprechend werden drei Hauptklassen unterschieden: *Metadaten für die Variablen*, *Metadaten für den gesamten Datensatz* und *Metadaten für Datenklassen* (vgl. Abb. 7.1).

Metadaten für die Variablen spezifizieren die Charakteristika der Variablen im allgemeinen als auch die speziellen Eigenschaften einzelner Variable. Wegen der Unterschiede zwischen abhängigen und unabhängigen Variablen soll im folgenden zwischen Metadaten für die Dimensionen und Metadaten für die Merkmale (abhängige Variable) unterschieden werden.

Die **Metadaten für die gesamte Datenmenge** spezifizieren die Eigenschaften des gesamten Datensatzes. Dabei lässt sich zwischen allgemeinen Metadaten, die sich auf die grundlegenden Eigenschaften der Datenmenge beziehen (z.B. die allgemeinen Beziehungen der Variablen untereinander), und Metadaten für relevante Teildatensätze, welche Teilmengen von Interesse und ihre Eigenschaften beschreiben, unterscheiden.

Um die spezifischen Charakteristika spezifischer Datenklassen einzubeziehen, sollen weiterhin **Metadaten für Datenklassen** beschrieben werden. Hier sollen insbesondere Metadaten für Volumendaten, für Strömungsdaten und für Multiparameterdaten einbezogen werden.

Die verschiedenen in der Literatur behandelten Metadaten (vgl. Abs. 7.1.1) lassen sich nun in dieses Schema einsortieren. In Anhang B.1 werden eine Vielzahl von Metadaten in kompakter Art aufgelistet und in das Schema integriert.

Die Flexibilität dieser Metadatenbeschreibung erlaubt Erweiterungen auf allen Ebenen, zum Beispiel lassen sich weitere Datenklassen wie GIS[1]-Daten oder gestreute Daten und ihre Spezifika einbeziehen. Allgemeine Metadaten hierfür finden sich bereits in den Metadaten für die Dimensionen und für den gesamten Datensatz.

Je nach Anwendungshintergrund können nun Metadaten gezielt verfeinert werden. Ein Beispiel hierfür ist eine *Raumsegmentierung* basierend auf den Eigenschaften der Raumgebiete und/oder extrahierten *Regionen von Interesse* (Stichwort „Feature Extraction“). Zum Beispiel kann dies für Strömungsdaten durchgeführt werden, indem kritische Punkte und separierende Linien/Ebenen zwischen diesen kritischen Punkten bestimmt werden. Diese topologischen Charakteristika sind von besonderer Bedeutung für die Analyse von Strömungsdaten und können genutzt werden, um Regionen von Interesse von Regionen mit geringerem Interesse zu separieren. Um dies zu verdeutlichen, sollen zwei Visualisierungen des elektrostatischen Feldes eines Wassermoleküls verglichen werden (vgl. Abb. 7.2). Eine konventionelle Stromliniendarstellung zeigt eine hervorstechende Region in der Mitte des Datensatzes mit einer Häufung kritischer Punkte. Allerdings lässt sich in diesem Gebiet das Verhalten der Strömung nur im Detail und die Anzahl kritischer Punkte nur begrenzt identifizieren (vgl. Abb. 7.2a).

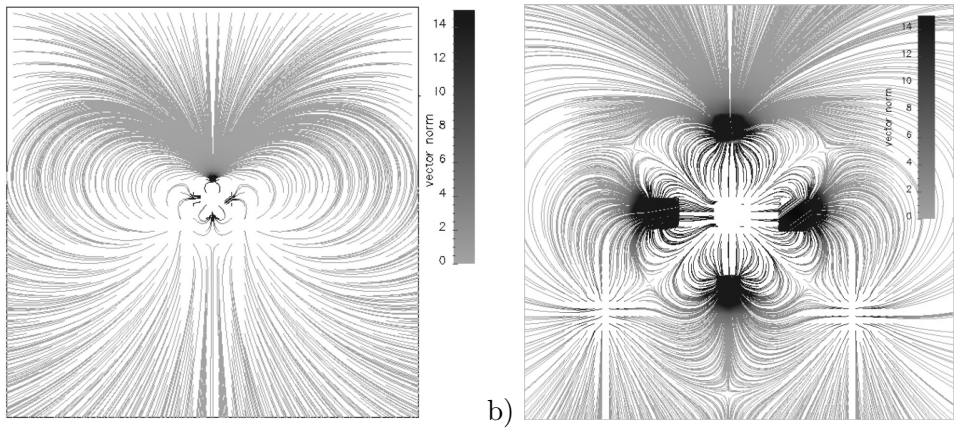


Abbildung 7.2: Stromliniendarstellungen des elektrostatischen Feldes eines Wassermoleküls (der Betrag der Vektoren wird auf die Farben der Stromlinien abgebildet); gemessen durch Konrad-Zuse-Institut Berlin; a) Überblicksdarstellung; b) Darstellung der Region von Interesse in der Mitte des Vektorfeldes, unter Einsatz einer erhöhten Anzahl von Stromlinien in dieser Region

Durch Klassifikation des Bereiches von Interesse im Zentrum des Vektorfeldes (basierend auf der Anzahl der dort auftretenden kritischen Punkte) lässt sich nun im Sinne des Visualisierungsdesigns die Darstellung dynamisch anpassen. Um wesentliche Details des Vektorfeldes besser identifizieren zu

[1] Geographische Informationssysteme

können, kann die entsprechende Region vergrößert und dort die Anzahl an ausgesetzten Stromlinien erhöht werden (vgl. Abb. 7.2b).

Ein weiteres Beispiel für die Verfeinerung und den Einsatz der vorgestellten Metadatenbeschreibung ist die Reduktion der Datenmenge für eine bestimmte Analyseaufgabe unter Einbeziehung des Nutzerwissens und automatischer Clusterverfahren (vgl. hierzu auch Kreuseler u. a. 2003). Diese Kombination ermöglicht eine flexible, einfach anzuwendende *Extraktion von Strukturen und Teilmengen von Interesse*. So lassen sich die Ergebnisse einer Clusterung (z.B. ein Baum bei einem hierarchischem Clusterverfahren) als Metadaten auffassen, und hieraus weitere Datencharakteristika wie Ausreißerdatensätze, gleiche Datensätze oder die Eigenschaften einzelner Cluster ableiten. Diese Informationen sind für verschiedene Aufgaben bei der Visualisierung einsetzbar. Zum Beispiel lassen sich Ausreißer visuell hervorgehoben oder die Komplexität der Darstellung durch das Verstecken von Clustern reduzieren (vgl. z.B. Herman u. a. 1998)).

7.1.2.2 Spezifikation von Metadaten für Klimadaten

Diese allgemeine, systematische Metadatenbeschreibung bildet die Grundlage, Metadaten für spezielle Anwendungen zu verwenden. Aufgrund ihrer Flexibilität lassen sich anwendungsspezifische Datencharakteristika leicht in die bestehende Klassifikation einordnen. Bei dem im Rahmen dieser Arbeit entworfenen Systemen VisAna und SimEnvVis erleichterte dies wesentlich das Softwaredesign und sicherte eine leichte Erweiterbarkeit an neue Anforderungen aus der Anwendung.

Um dies zu verdeutlichen, werden im folgenden verfeinerte anwendungsspezifische Metadaten speziell für Klimadaten vorgestellt, ohne dabei einen Anspruch auf Vollständigkeit für dieses Gebiet zu beanspruchen. Diese basieren auf den tatsächlichen Anforderungen aus der Anwendung.

- Allgemeine Metadaten für die gesamte Datenmenge / Ursprung einer Datenmenge:
  - Name des simulierten Klimamodells$^{\circ}$
  - Experimenttyp$^{\circ\triangleright}$(Verhaltensanalyse, Monte-Carlo)
  - Anzahl an Simulationsläufen$^{\circ\triangleright}$
- Metadaten für relevante Datenteilmengen:
  - Kopplung mehrerer Teilmodelle/Teildatensätze mit unterschiedlichen Unterräumen des gesamten Beobachtungsraumes$^{\circ}$(insbesondere aufgespannt durch geogr. Länge, geogr. Breite und Höhe, Zeit, aber auch Experimentdimensionen); Beschreibung der beteiligten Dimensionen und Gitter
- Allgemeine Metadaten für die Variablen / weitere semantische und historische Informationen:
  - physikalische Größe$^{\circ}$, z.T. unter Einbeziehung von deren Wirkungsbereich (z.B. Temperatur in 2m über dem Boden, Meeresspiegeldruck, Nettoprimärproduktion an Biomasse, ...)
  - Einheit der Variable$^{\circ}$
  - Datenerzeugung durch Anwendung spezieller Operatoren$^{\triangleright}$(z.B. Mittelwert, Abweichung vom Mittelwert des Startzeitpunktes, ...)
- Metadaten für die Dimensionen:
  - Lage der Beobachtungspunkte$^{\circ}$an den Ecken oder in Mitte der Gitterzellen (vgl. auch Wirkungskreis)
  - Art der Längen/Breitenzählung$^{\circ}$(Länge von 0° bis 360° vs. von -180° bis +180°)
  - Land- oder Seemasken$^{\circ\star}$(Definition der Daten lediglich auf dem Land oder im Meer)

Legende: $\circ$ BESCHREIBENDE $\star$ ABGELEITETE $\triangleright$ HISTORISCHE Metadaten

Diese Metadaten lassen sich auf verschiedene Arten für das VDM einsetzen (vgl. auch Abs. 7.3). Dies reicht vom Einsatz bei der Beschriftung (z.B. Einheiten oder Informationen zum Ursprung und zur Datenerzeugung) über die konkrete Abbildung der Variable (z.B. Auswahl einer geeigneten Farbskala) oder der korrekten Darstellung einer Weltkarte (Art der Längen/Breitenzählung) bis zur Bestimmung der Eignung einzelner Techniken (z.B. inwieweit sie den geographischen Bezug der Daten geeignet wiedergeben).

7.1.3 Erhebung von Metadaten

Prinzipiell ist es möglich, eine Vielzahl von Metadaten zu erheben und zu speichern. Gängige Datenspeicherungsformate wie das NetCDF-Format liefern, auch in Abhängigkeit des bei der Datenspeicherung betriebenen Aufwands, eine Fülle von vor allem *beschreibenden* Metadaten. Auf der anderen Seite ermöglichen eine Vielzahl automatischer (Mining-)Verfahren, eine große Menge an *abgeleiteten* Metadaten zu erheben. Weil die Bestimmung potentiell nützlicher Metadaten deswegen mit hohem Zeit- und Resourcenaufwand verbunden sein kann, ist es dringend erforderlich, effektive, praktikable Wege zur Metadatenerhebung zu finden.

Dies kann durch die folgenden Strategien erreicht werden:

- Festlegung einer **geeigneten Ordnung** der Metadatenerhebungsprozesse, unter Vermeidung wiederkehrender Berechnungen und unter gegenseitiger Unterstützung der Ergebnisse einzelner Prozesse
- **Unterteilung** des Erhebungsprozesses in automatische und halbautomatische Berechnungen sowie Nutzereingaben mit verschiedenen Stufen von Unterstützung. In Abhängigkeit des Anwenderprofils können die **Interaktionsstufen** zwischen einem hohen Grad an Interaktion und Parametereingabe bis zu einem hohen Grad der Benutzung von Standardwerten reichen.
- **Verfeinerung bzw. Nachberechnung** von Metadaten bei Bedarf in einem iterativen Explorationsprozess
- Benutzung von **Zeitrahmen** für die Metadatenextraktion (z.B. schnelle Analyse, ..., keine Zeitbeschränkungen für die Analyse) für die geeignete Parametrisierung oder sogar Vermeidung von zeitintensiven Prozessen

In Abhängigkeit der mit den Daten bereits gespeicherten Metadaten muss je nach Datenquelle/Datenformat der Metadatenerhebungsprozess entsprechend angepasst werden. Im Rahmen dieser Arbeit soll dies am Beispiel zweier Typen von Datenquellen illustriert werden: textuelle Daten aus einfachen Ascii-Tabellen (Abs. 7.1.3.1) und im NetCDF-Format gespeicherte Daten (Abs. 7.1.3.2).

7.1.3.1 Erhebung von Metadaten aus Texttabellen

Textdateien in Tabellenform, in denen eine Vielzahl von Daten (z.B. im Internet) vorliegen, enthalten typischerweise neben den Namen der Variablen und ggf. den Namen der Beobachtungsfälle („data records“) keine weiteren Metadaten. Werkzeuge, welche den Anwender beim Einlesen solcher Daten unterstützen (z.B. der Data Prompter des OpenDX), sind zum Teil schwer zu bedienen, und geben nur im begrenztem Maße Unterstützung bei der Metadatenerhebung. Deswegen wurde in Nocke (2000) eine speziell für solche Datensätze zugeschnittene Metadatenerhebung konzipiert und im Framework Metadatum umgesetzt.

Abbildung 7.3a zeigt das Hauptablaufschema zur Erhebung von Metadaten. Nach dem Einlesen des Datensatzes werden allgemeine *Metadaten für die Variablen* und die *spezifischen Metadaten für die Merkmale* bestimmt. Dies legt die Grundlage für die folgenden Prozesse und deren Zugriff auf die Datenwerte, indem Skalen- und Datentypen der Variablen festgelegt werden. Des Weiteren erfolgt in diesem Schritt die Bestimmung der Verteilungseigenschaften der Daten, welches die folgenden

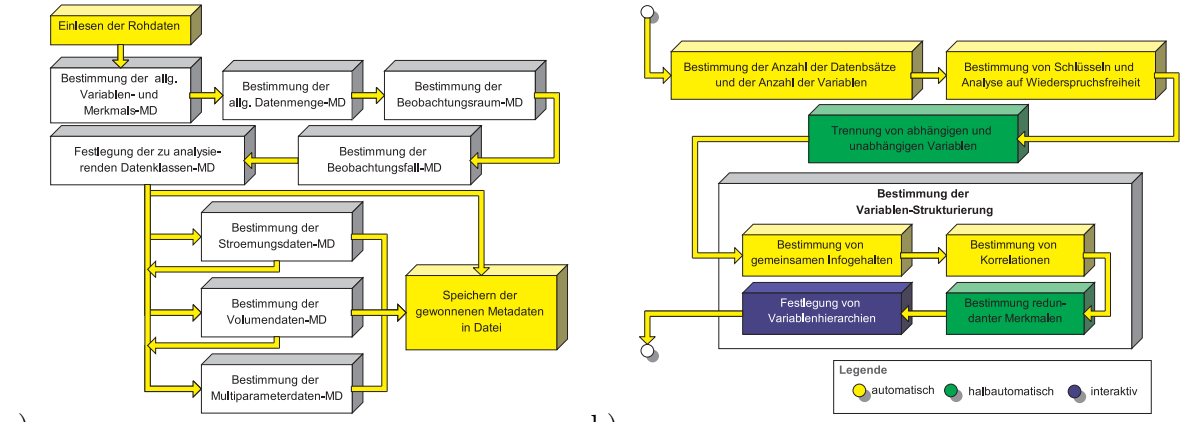


Abbildung 7.3: Ablaufschemata zur Erhebung von Metadaten (MD); a) Allgemeines Ablaufschema; b) Erhebung von allgemeinen Metadaten für die gesamte Datenmenge

Schritte beschleunigt. Im Anschluss daran erfolgt die Extraktion der allgemeinen Metadaten für die gesamte Datenmenge. Dies schließt statistische Berechnungen von Korrelationen und die Klassifikation der Variablen in abhängige und unabhängige Variable ein. Darauf folgt die Erhebung der Metadaten für die Dimensionen, und abschließend werden spezifische Metadaten für die einzelnen Datenklassen erhoben.

Zur Illustration eines solchen Metadatenerhebungsschrittes sei hier die Erhebung der allgemeinen Metadaten für die gesamte Datenmenge ausgeführt (vgl. Abb. 7.3b). Nach der Bestimmung der Anzahl der Beobachtungsfälle und der Anzahl der Variablen (basierend auf der Anzahl an Spalten und Zeilen in der Textdatei) erfolgt eine Schlüsselanalyse[2], um die abhängigen von den unabhängigen Variablen zu separieren. Durch Auswahl des kürzesten Schlüssels kann nun die Separation automatisch oder unter Anpassung des Anwenders erfolgen. Entspricht der vorgeschlagene Schlüssel nicht der Erwartung des Anwenders, kann dies ein Indiz für Inkonsistenzen in den Daten sein. In diesem Falle kann der Datensatz angepasst und die Metadatenerhebung erneut ausgeführt werden. Im Anschluss daran erfolgen vier Schritte zur Spezifikation gesonderter Strukturinformationen zwischen den Variablen: Bestimmung von (1) (gemeinsamen) Informationsgehalten und (2) Korrelationen zwischen Variablen, worauf basierend dann (3) Variablen bei Bedarf halbautomatisch aus dem weiteren Analyseprozess ausgeschlossen werden können und (4) eine Definition von Variablenabhängigkeiten (z.B. zwischen Monat und Tag oder basierend auf einer Hauptkomponentenanalyse) erfolgt.

Diese Methodiken mündeten in das prototypische Framework Metadatum, welches am Beispiel von Texttabellen illustriert, wie automatische Verfahren mit Nutzerinteraktionen kombiniert werden können, um die Nutzer eines VDM-Systems im Vorfeld der eigentlichen Analyse wesentlich dabei zu unterstützen, die für eine anschließende (visuelle) Analyse erforderlichen Eigenschaften ihrer Daten zu erheben (vgl. Nocke 2000; Nocke u. Schumann 2002). Neben allgemeinen Metadaten werden in diesem Framework auch Metadaten für verschiedene Datenklassen einbezogen. Auch lassen sich Tabellen im Sinne einer Fremdschlüsselbeziehung miteinander verknüpfen. Je nach Erfahrung und Zeitrahmen der Anwender können Metadatenerhebungsschritte durch Standardbelegungen belegt werden und langandauernde Analyseverfahren abgekürzt oder ganz vermieden werden. Abbildung A.15 im Anhang zeigt einen Screenshot des entworfenen Frameworks.

Eine solche, halbautomatische Metadatenerhebung für textuelle Tabellendaten und deren Erhebung gibt es bisher nicht. Gerade bei der konsequenten Einbeziehung einer Vielzahl von Metadaten

[2]Schlüssel sind Kombinationen von Variablen, deren Tupel eine eindeutige Zuordnung jedes Beobachtungsfalles ermöglicht.

und deren Erhebung mit verschiedenen Mechanismen zur Nutzerunterstützung bei deren Erhebung beschreitet diese Arbeit Neuland.

7.1.3.2 Erhebung von Metadaten aus NetCDF-Daten

Im Unterschied zur Metadatenextraktion aus Textdateien, lassen sich aus Dateien im NetCDF-Format bereits wesentlich mehr *beschreibende*, aber auch *abgeleitete* und *historische* Metadaten extrahieren. Da diese zum einen häufig nicht vollständig und zum anderen auch inkonsistent sein können, ist auch hierfür ein Konzept zur Erhebung von Metadaten erforderlich. Die Separierung von abhängigen und unabhängigen Variablen und Metadaten zu Minima, Maxima und Fehlwerten der einzelnen Merkmale liegen typischerweise bereits vor. Probleme (z.B. Instabilitäten von Visualisierungstechniken) können jedoch auftreten, wenn der Datenerzeuger z.B. inkonsistente Metadaten spezifiziert hat (z.B. falsche Minima oder Maxima), oder wenn für eine spezielle VDM-Technik erforderliche Metadaten nicht oder unter einem anderen Namen abgespeichert wurden (z.B. Einheit oder Skalentyp). Des Weiteren erlaubt das NetCDF-Format nicht explizit, gestreute Daten zu verwalten. Erfolgt dies über Definition eines separaten Index für jede Variable, der jeden Beobachtungspunkt eindeutig identifiziert, wird auch hier eine Separation von abhängigen und unabhängigen Variablen erforderlich. Darüber hinaus stößt eine Speicherung von komplexeren Metadaten wie Hierarchien im NetCDF-Format an Grenzen.

Dementsprechend ergibt sich die Notwendigkeit, die bereits vorliegenden Metadaten auf mögliche Inkonsistenzen hin zu untersuchen und bei Bedarf um neue Metadaten zu erweitern. Dabei wurde - im System SimEnvVis (vgl. Abs. 7.3), welches sich auf Daten im NetCDF-Format spezialisiert hat - im Unterschied zur Erhebung einer großen Anzahl von Metadaten wie im System Metadatum (vgl. Abs. 7.1.3.1), eine alternative Strategie verfolgt: lese alle vorhandenen Metadaten aus NetCDF aus, und erweitere diese bei Bedarf im Laufe des Visualisierungsdesign-Prozesses, falls weitere Informationen benötigt werden. Diese Strategie ist effizient, da die wichtigsten Metadaten bereits nach dem Einlesen der Metadaten aus NetCDF vorliegen, und da Doppelberechnungen so vermieden werden. Abbildung 7.4 illustriert das Vorgehen bei der Erhebung von Metadaten aus dem NetCDF.

Ausgangspunkt ist die Datenerzeugung, bei der je nach Art der zu speichernden Daten und der Anwendung neben den Daten verschiedene Metadaten gespeichert werden. Diese werden dann in einem Einlese/Verarbeitungsprozess in eine interne Metadatenpräsentation überführt. Dies umfasst insbesondere das Einlesen der Metadaten für die Dimensionen (u. a. *Dimensionen, Gitterinformationen*) und der Merkmale (z.B. Minimum, Maximum). Fehlen hierbei relevante Informationen für den VDM-Prozess, werden diese unter Einsatz eines Thesaurus entweder unter einem anderen Namen gesucht (z.B. gibt es unterschiedliche Bezeichner für den Datenwert, der fehlende Daten repräsentiert) oder anhand bereits vorliegender Metadaten der Variable automatisch bestimmt (z.B. Einheit anhand des Variablennamens). Schlägt diese Art der Metadatenersetzung fehl oder es können nur begrenzt geeignete Metadaten bestimmt werden, hat der Nutzer in dieser Phase die Möglichkeit, zur Datenerzeugung zurückzukehren. Im Anschluss werden in den NetCDF-Daten vorliegende Informationen über Merkmale auf Teilgittern des gesamten, durch die Dimensionen aufgespannten Raumes, in eine geeignete Struktur gebracht (Abb. A.13 im Anhang illustriert eine solche aus einem gekoppelten Klimamodell erzeugte Struktur). Letzter Teil des Einlese/Verarbeitungsprozesses ist das Einlesen allgemeiner Metadaten für die Datenmenge aus den globalen Attributen der NetCDF-Datei.

Diese Metadaten bilden die Basis für den anschließenden Design- und Visualisierungsprozess, und können bei Bedarf verfeinert werden. Als erster Schritt erfolgt eine Darstellung der Metadaten. In diese Darstellung kann der Anwender eingreifen und ggf. Metadaten anpassen. Gibt es grundlegende Inkonsistenzen in den Daten, kann auch direkt zur Datenerzeugung zurückgekehrt werden. Im anschließenden (interaktiven) Filterprozess werden nach erfolgter Filterung die Metadaten auf die „Metadaten von Interesse“ eingeschränkt, was insbesondere die aktuell ausgewählte(n) Teilmengen

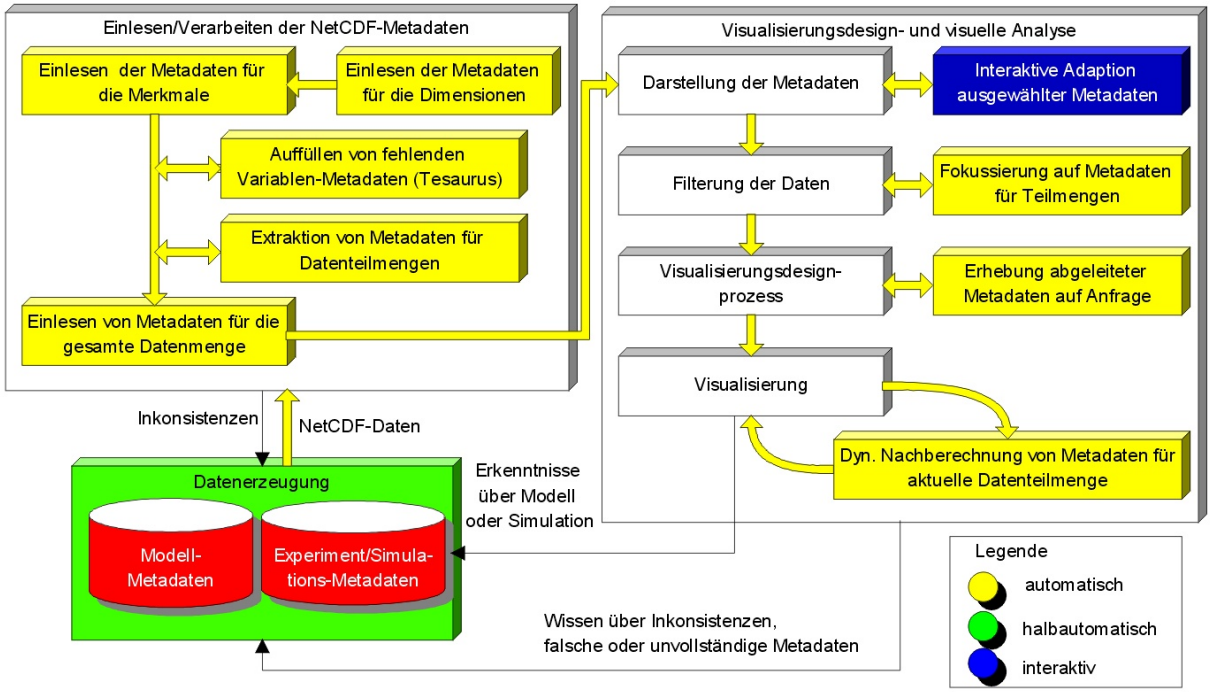


Abbildung 7.4: Schema zur Erhebung von Metadaten aus NetCDF-Daten (für das Framework SimEnvVis)

betrifft. Diese Fokussierung auf (eine) Datenteilmenge(n) von Interesse reduziert den Aufwand bei der Ableitung von Metadaten wesentlich. Dadurch können im anschließenden Visualisierungsdesignprozess Metadaten effizient auf Anfrage berechnet und ggf. für eine spätere Wiederverwendung gespeichert werden. Auch im anschließenden visuellen Analyseprozess können für (zum Teil noch weiter eingeschränkte) Teilmengen dynamisch Metadaten nachberechnet werden (z.B. bei der Bestimmung von Minimal- und Maximalwerten für den aktuell dargestellten Ausschnitt der Daten). Mit den dabei gewonnenen Erkenntnissen kann der Anwender zur Datenerzeugung zurückkehren, um z.B. sein Modell weiter zu verbessern.

7.1.4 Darstellung von Metadaten

Die bei der Erhebung von Metadaten gewonnenen Informationen lassen sich auf vielfältige Art im VDM-Prozess einsetzen. So ermöglichen sie es, einen Überblick über die Eigenschaften der Daten zu erhalten, ohne die Daten direkt zu untersuchen. So lässt sich durch Metadaten-Darstellung expliziter und/oder abgeleiteter Strukturen ein vertieftes Verständnis der Datenmenge gewinnen.

Die explizite Darstellung von Metadaten *in textueller Form* ist ein typisches Vorgehen zur Darstellung grundlegender Dateneigenschaften. Dies ermöglicht neben einem Überblick über die Daten (und deren Konsistenzprüfung) bei Bedarf auch eine Adaption der Metadaten sowie eine Filterung der Daten. Problem hierbei ist, dass die Metadaten (vor allem komplexer Datenmengen) selbst schwer zu verstehen sein können. Als Lösung für dieses Problem lassen sich die Metadaten durch leicht verständliche grafische Metaphern ergänzen. Abbildung 7.5 illustriert dieses Vorgehen.

Die aufgrund variierender Gitterstrukturen in unterschiedliche Teilmengen aufgeteilte Datenmenge soll dabei veranschaulicht werden. Um das zum Teil komplexe Zusammenspiel der beteiligten Dimensionen zu illustrieren, wurden Ikonen entworfen, welche die Dimensionalität einer Teilmenge auf die Anzahl an Kugeln und die Zugehörigkeit zur gleichen/unterschiedlichen Teilmenge auf Farben abbildet. Entsprechend werden die einzelnen Merkmale markiert (Tabelle in Abbildung 7.5a).

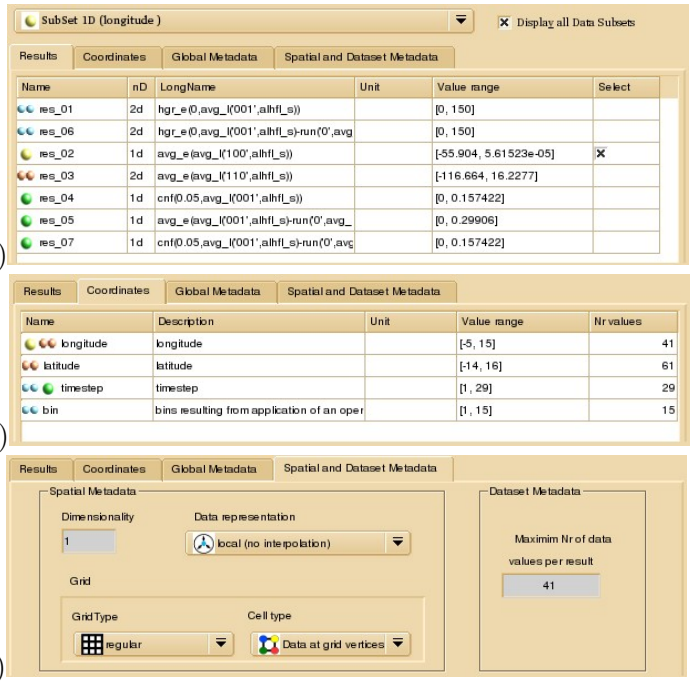


Abbildung 7.5: Illustration von Metadaten in einer textbasierten Nutzerschnittstelle; a) Metadaten für die Merkmale; b) Metadaten für die Dimensionen; c) Allgemeine Datenmenge- und räumliche Metadaten

Auf diese Art lassen sich deren Zugehörigkeit zu Teilmengen veranschaulichen und einzelne Teilmengen filtern (ComboBox über der Tabelle 7.5a,b,c). In einer zweiten Ansicht werden die zugehörigen Dimensionen und deren Beitrag zu einzelnen Teilmengen illustriert (vgl. Abb. 7.5b). Weiterhin lassen sich durch Ikonen die Bedeutung von Metadaten verdeutlichen (vgl. Abb. 7.5c, mit speziellen Ikonen für den *Gittertyp* und den *Wirkungsbereich* der Datenwerte).

Bei komplexen Datenmengen mit einer Vielzahl von Variablen stoßen textuelle Darstellungen an Grenzen der Darstellungsfläche und der Interpretierbarkeit. Außerdem weisen Metadaten zum Teil selbst einen hohen Grad an Struktur auf. So können diese in Form von Matrizen (z.B. gemeinsame Korrelationen), Hierarchien (z.B. aufgrund einer hierarchischen Clusterung oder einer Faktorzerlegung) oder Graphen (z.B. Abhängigkeiten von (Modell-)Variablen aufgrund des Wissens ihrer Erzeugung) vorliegen. Entsprechend lassen sich die für diese Datenklassen entworfenen Visualisierungstechniken einsetzen. Hierbei lassen sich zwei Varianten unterscheiden:

1. kombinierte Visualisierung von Metadaten und Daten: hierbei werden die Darstellungen der Daten um weitere Charakteristika angereichert (vgl. Methoden zur Feature-Visualisierung, z.B. Darstellung von kritischen Punkten in Abb. 5.7b),
2. separate Darstellung von Metadaten: hierunter fallen insbesondere Darstellungen von strukturierten Metadaten (z.B. die den Merkmalen zugrunde liegende Struktur basierend auf einem DGLS in Abb. 6.36 oder die Clusterstruktur der Beobachtungsfälle basierend auf einem hierarchischen Clusterverfahren in Abb. 6.5).

Zur separaten Darstellung von gemeinsamen Informationsgehalten von n-Tupeln von Variablen (vgl. Theisel 1995) wurde im Rahmen dieser Arbeit eine neue Visualisierungstechnik entworfen und umgesetzt (vgl. Abb. 7.6). Diese Technik stellt die relevanten Informationsgehalte $I(i_1, ..., i_n)$ von Kombinationen von Variablen über einem Schwellwert s dar. Hierbei werden Tupel von Variablen mit relevanten gemeinsamen Informationsgehalten auf Würfel abgebildet, die in einer Hierarchie von Kreisen in 3D folgendermaßen angeordnet werden:

- Die Informationsgehalte jeder einzelnen Variable werden auf die oberste Ebene abgebildet (z.B. der Informationsgehalt des Merkmals Salzgehalt (SALZ) in Abb. 7.6).
- Die zweite Ebene enthält Paare von Variablen mit relevanten Informationsgehalten (z.B. die Paare Temperatur und Salzgehalt (TEMP SALZ), Tiefe und Salzgehalt (TIEFE SALZ) u. a.).
- Die dritte Ebene enthält Tripel von Variablen (z.B. das Tripel Temperatur, Salzgehalt und Sauerstoffgehalt (TEMP SALZ O2ML))
- u.s.w.

Die Farbe des Würfels repräsentiert den Informationsgehalt des zugehörigen Tupels (rot für hohe

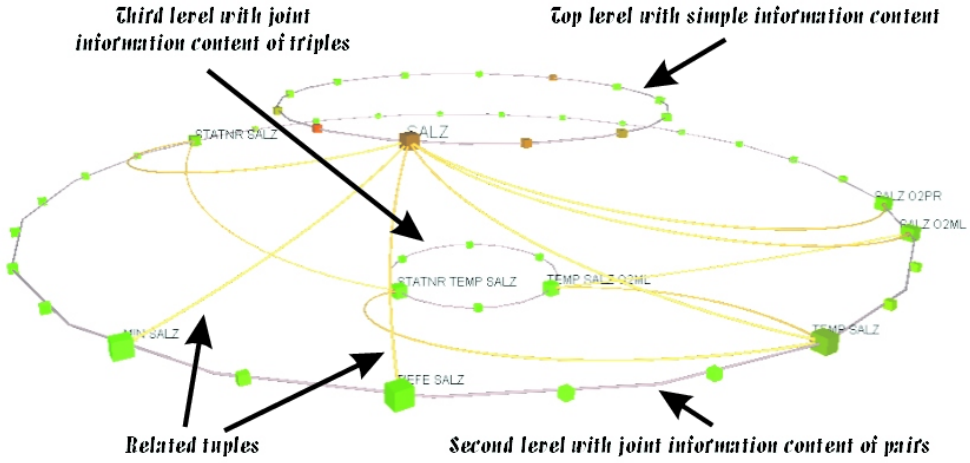


Abbildung 7.6: Metadaten-Visualisierung von gemeinsamen Informationsgehalten für einen Ostsee-Datensatz; die Variable Salzgehalt (SALZ) und die Variablen, die mit ihr Kombinationen über einem Schwellwert bilden, wurde beschriftet

Werte, grün für geringe Werte). Weiterhin werden bei Bedarf die Beziehungen zwischen Tupeln[3] auf Kurven in 3D[4] abgebildet. Um große Strukturen von solchen Informationsgehalten effektiv untersuchen zu können, können einzelne Würfel bei Bedarf ein/und ausgeblendet sowie beschriftet werden.

7.1.5 Diskussion

Die Beschreibung und Verarbeitung der Daten sowie deren Charakteristika bildet eine wesentliche Grundlage für die Entwicklung von Visuellen Data Mining-Systemen. Existierende Taxonomien und Formate fokussieren jedoch auf bestimmte Datenaspekte[5] und konzentrieren sich dabei im allgemeinen entweder auf beschreibende oder auf abgeleitete Metadaten. Um die breite Spannbreite an Metadaten für den Entwurf und die Umsetzung von VDM-Systemen mit integrierter Anwenderunterstützung verfügbar zu machen, wurde in dieser Arbeit eine **allgemeingültige Metadatenspezifikation** entworfen (vgl. auch Nocke 2000; Kreuseler u. a. 2003; Nocke u. Schumann 2002; Lange u. a. 2006). Dies schließt vor allem die Identifikation von Kriterien (u. a. Erweiterbarkeit, Separierung und Nutzerorientierung) an eine solche Spezifikation und die konsequente Einbeziehung heterogener Arten von Metadaten auf verschieden Abstraktionsniveaus ein. Dadurch lassen sich sowohl allgemeine Metadaten über die gesamte Datenmenge und die Variablen als auch Metadaten für spezielle Datenklassen und Anwendungen integrieren. Insbesondere werden eine Vielzahl von beschreibenden und historischen Metadaten systematisch einbezogen. So lassen sich unter anderem in der Literatur bisher eher unterrepräsentierte Aspekte zur Datenqualität auf verschiedenen Ebenen einbeziehen. Als Anwendungsbeispiel wurde die Erweiterbarkeit dieser Spezifikation im Klimaumfeld demonstriert. Weiterhin lässt sie sich einsetzen, um Metadaten zur Kopplung von Daten aus mehreren Quellen zu beschreiben (vgl. Wagenknecht 2006).

Darüber hinaus stellt sich diese Arbeit dem Problem der **effektiven Erhebung von potentiell nützlichen Metadaten**. Zu dessen Lösung wurden (semi-)automatische Ansätze für Daten in

[3] Zwei Tupel stehen genau dann in Beziehung, wenn alle Elemente eines Tupels Teil des anderen Tupels sind.

[4] Dasselbe Vorgehen lässt sich auch in 2D umsetzen, wobei die Knoten des (Informationsgehalt-Graphen) in Ebenen angeordnet und mit Linien verbunden werden.

[5] wie z.B. auf die allgemeine Art und Anzahl der Variablen sowie des Gitters oder auf Metadaten für spezielle Datenklassen

Texttabellen und für Daten in NetCDF-Dateien entworfen und umgesetzt. Hierfür wurden zwei neuartige Konzepte entwickelt: (1) die gegenseitige Unterstützung von Metadatenerhebungsprozessen durch Findung einer geeigneten Reihenfolge, (2) die dynamische Nachberechnung von Metadaten bei Bedarf. Je nach Anwendungshintergrund und verfügbarem Zeitrahmen können so interaktive von zeitaufwendigen Prozessen getrennt und Metadaten in einem iterativen Prozess bei Bedarf verfeinert werden. Mit dieser direkten Einbeziehung von (semi-)automatischen Metadatenberechnungsprozessen in den VDM-Prozess wird die Steuerung dieses Prozesses und die Aufdeckung von Mustern in den Daten wesentlich erleichtert. Dies lässt die Grenze zwischen automatischen Mining-Methoden auf der einen Seite und Metadatenerhebungsverfahren auf der anderen Seite verschmelzen.

Weiterhin lassen sich **Metadaten** einsetzen, um grundlegende Eigenschaften der Daten leicht verständlich zu **illustrieren**. Durch die zum Teil hohe Verdichtung der Informationen können so wichtige Informationen kompakt repräsentiert werden, ohne die eigentlichen Daten selbst darstellen zu müssen. Probleme ergeben sich jedoch, wenn die Metadaten selbst eine (heterogene) Struktur aufweisen oder eine Vielzahl von Variablen oder Datenteilmengen beinhalten. Für die Darstellung solcher Metadaten lassen sich Techniken zur Darstellung von Strukturen verwenden (z.B. Baum- und Graphvisualisierungstechniken). Hierfür stellt die hier vorliegende Arbeit eine neue, interaktive Darstellungstechnik für n-Tupel von gemeinsamen Informationsgehalten vor. Bei der systematischen Darstellung verschiedener Klassen abgeleiteter Strukturen besteht jedoch noch Forschungsbedarf.

Die hier vorgestellten Konzepte für Metadaten wurden konsequent für die Analyse von Klimadaten, aber auch für die Analyse klinischer, fusionierter Daten (vgl. Wagenknecht 2006) eingesetzt und in die drei Frameworks Metadatum, SimEnvVis und VisAna einbezogen. Insbesondere bilden sie eine wichtige Grundlage für vielfältige Entscheidungen im Prozess des Visualisierungsdesigns (vgl. Abs. 7.3).

7.2 Analyseziele für das visuelle Data Mining

Neben den Metadaten ist die explizite Einbeziehung der Analyseziele ein wichtiger Kontextfaktor, um den Analyseprozess gemäß den Absichten der Anwender geeignet zu lenken. Ein Problem in diesem Kontext stellt der breite Interpretationspielraum bei der Bedeutung von textuellen Zielstellungen dar. So liegen in der Literatur eine Vielzahl von Beschreibungen von Zielstellungen vor, die zum Teil erhebliche inhaltliche Überschneidungen aufweisen (vgl. Abs. 7.2.1). Erste Ansätze zur Vereinheitlichung solcher Zielstellungen beziehen jedoch nicht die Breite der bekannten Ansätze ein, sondern konzentrieren sich auf praktikable Lösungen.

Gerade diese Begriffsvielfalt macht es für Anwender aus anderen Gebieten schwer, sich in der Visualisierungsliteratur, aber auch in existierenden Tools zurechtzufinden (vgl. auch Brodlie u. a. 2004). Deswegen ist es Ansatz dieser Arbeit, die Vielzahl in der Literatur betrachteten Zielstellungen einzubeziehen, und dabei anwendungsunabhängige, allgemeine Zielstellungen von anwendungsabhängigen, spezifischen Zielstellungen zu trennen (vgl. hierzu auch Nocke u. Schumann 2004). Dies ermöglicht es, die Lücke zwischen dem Vokabular spezieller Anwendungen und dem allgemeinen, in der Visualisierung üblichen Bezeichnungen zu schließen.

7.2.1 Stand der Forschung

Eine Vielzahl von Ansätzen beziehen die Ziele und Aufgaben der Anwender implizit mit ein, benutzen diese jedoch nicht für eine automatische Generierung von Darstellungen. So bezieht bereits Bertin (1981) elementare Fragen (elementary questions), Fragen mittlerer Granularität (intermediate Questions), übergreifende Fragen (overall question), und darüber hinaus auch viele weitere Zielstellungen implizit mit ein.

Es existieren eine Vielzahl von Ansätzen, welche Analyseziele formalisieren und in Visualisierungssystemen einsetzen (vgl. Schumann u. Müller 2000, für eine Übersicht). Roth u. Mattis (1990) führen Ziele zur Informationsuche ein (z.B. *Werte suchen*, *Werte vergleichen*, *Verteilungen identifizieren* und *Korrelationen finden*). Wehrend u. Lewis (1990) trennen Zielstellungen in einen „Objektteil“ (z.B. die Analyse von skalaren Eigenschaften der Daten („scalar“) oder von Positionen („position“)) und Operationen (z.B. *identifizieren* und *vergleichen*). Keller u. Keller (1993) präsentieren Beispiele für expressive Darstellungen für verschiedene Visualisierungsziele (und Metadaten) und beziehen dabei Datentypen, Operationen, Objekte und den Anwendungskontext ein. Robertson (1990) führt den Ortsbezug einer Analyse ein (ob der Anwender an *globalen*, *lokalen* oder *punktuellen* Informationen über die Daten interessiert ist). Theisel (1994) beschreibt einen abstrakten Formalismus für Zielstellungen, setzt diese miteinander in Beziehung und bildet sie auf konkrete Ziele ab. Ähnlich hierzu formulieren Andrienko u. Andrienko (2006) unter Nutzung einer formalen Notation Zielstellungen basierend auf der Abbildungsfunktion der Daten (von unabhängigen auf abhängige Variable). In Anlehnung an Bertin (1981) werden hierbei sowohl elementare („lockup and relational tasks“) als auch zusammengesetzte Ziele („descriptive and connection discovery tasks“) einbezogen.

In seiner Task-by-Data-Type-Taxonomie führt Shneiderman (1996) verschiedene Aufgaben (z.B. *Überblick*, *Zoom* und *Filtern*) und sieben Datentypen (z.B. 1D, 2D und Hierarchie) ein. In Anlehnung an Wehrend u. Lewis (1990) führen Fujishiro u. a. (2000) Zielobjekte der Analyse („analysis targets“: z.B. Skalare, nominaler Aspekt der Daten, Richtung) ein und separieren diese von den Aktionen der Analyse. Diese Trennung erlaubt beliebige Zielobjekte mit beliebigen Aktionen zu kombinieren. Zusätzlich werden diese beiden Kategorien von Fujishiro u. a. (2000) mit den Zielen von Shneiderman (1996) kombiniert. Amar u. Stasko (2004) untersuchen die Grenzen von Informationsvisualisierungssystemen und identifizieren allgemeine Klassen von Aufgabenstellungen (Unsicherheiten ausdrücken, Beziehungen konkretisieren, Ursache und Effekte bestimmen, Hypothesen bestätigen) und konkretisieren diese in einer Aufgabentaxonomie (Amar u. a. 2005, z.B. Wert extrahieren, Filtern, ...).

Zusammenfassend lässt sich konstatieren, dass explizite Analyseziele nur in Ansätzen in Visualisierungs- und VDM-Systemen eingesetzt werden. Als eine Ursache hierfür lassen sich die verschiedenen Interpretationen der verwendeten Begriffe in unterschiedlichen Anwendungen identifizieren, die Mehrdeutigkeiten der Begriffe verursachen. Um sich diesem Problem zu stellen, müssen verschiedene Sichten auf Zielstellungen etabliert werden, welche die Lücke zwischen der internen Verarbeitung der Ziele, dem Vokabular des Visualisierungsexperten dem Vokabular der Anwendung schließen.

7.2.2 Ein allgemeine Spezifikation für Analyseziele

Ansatz dieser Arbeit ist es, die verschiedenen in der Literatur betrachteten, zum Teil isolierten Aspekte miteinander zu kombinieren:

1. Anwendergetriebene Aspekte:
   (a) Aufgaben (allgemeine Ziele) - Überblick, Extrahieren, Details-on-Demand, - (Shneiderman 1996),
   (b) Aktionen (spezielle Ziele) - Assoziieren, Klassifizieren, Vergleichen, Identifizieren, Aufdecken, ... (Wehrend u. Lewis (1990), Roth u. Mattis (1990), Keller u. Keller (1993), Fujishiro u. a. (2000), Amar u. a. (2005)),
   (c) Analyseort - global, lokal und punktuell (Bertin (1981), Robertson (1990)),
2. Datengetriebene Aspekte:
   (a) Gegenstand der Analyse (Datenmuster und Datentypen von Interesse) - Skalar, Nominal, Richtung, ... (Wehrend u. Lewis (1990), Roth u. Mattis (1990), Keller u. Keller (1993), Fujishiro u. a. (2000)), aber auch Ausreißer und typische Werte,

   (b) Spezialisierung - allgemeine oder datenklassenspezifische Ziele (Shneiderman (1996)),
   (c) Abhängigkeit - abhängige und unabhängige Variable stehen im Fokus des Interesses (Keller u. Keller (1993), dos Santos u. Brodlie (2004), Andrienko u. Andrienko (2006)),

3. Anwendungskontext - neutral und anwendungsabhängig (Keller u. Keller (1993)),
4. Komplexitätsaspekt: elementare und zusammengesetzte Ziele (vgl. z.B. Amar u. Stasko 2004; Amar u. a. 2005; Andrienko u. Andrienko 2006).

Zusätzlich zu diesen Aspekten sind auch deren Beschreibungsformen für eine Spezifikation von Analysezielen relevant. Neben formalen Beschreibungen (vgl. z.B. Theisel 1994; Andrienko u. Andrienko 2006) schließt dies insbesondere auch verbale Beschreibungen und verschiedene Speicherungsformen wie das XML-Format oder interne Formate ein.

Bei der Betrachtung der aufgelisteten Aspekte ergibt sich ein umfassender Blick auf die Spezifikation von Analysezielen. Basierend darauf wird im folgenden eine praktikable, für verschiedene Anwendungsszenarien adaptierbare Spezifikation von Zielen durchgeführt. Herausforderungen hierbei sind neben der Einbeziehung der vielschichtigen Aspekte eine verständliche Präsentation der Ziele in einer leicht bedienbaren Nutzerschnittstelle zu ermöglichen. Um eine Anwenderüberlastung, aber auch einen Verlust wichtiger Ziele zu vermeiden, ist der Ansatz dieser Arbeit diese beiden Aspekte auszubalancieren. Deswegen wurde ein dreistufiges Vorgehen gewählt: (1) wichtige (anwendergetriebene und relevante datengetriebene) Aspekte werden dem Anwender in einem geeigneten Nutzerinterface zur Verfügung gestellt; (2) weitere Aspekte wie die Spezialisierung und der Anwendungskontext werden durch eine geeignete Attributierung hinzugefügt, und (3) unter Einbeziehung des Komplexitätsaspektes werden attributierte Ziele miteinander kombiniert und lassen sich so auf einem vereinfachten Level dem Anwender zur Verfügung stellen.

Entsprechend, werden die anwendergetriebenen Aspekte und der Gegenstand der Analyse als wichtige Aspekte auf der ersten Stufe definiert. Dann lassen sich AO als die Menge aller möglichen Werte, die Analyseorte spezifizieren, und entsprechend AG für den Gegenstand, AA für die Aktion und AF für die Aufgabe der Anaylse definieren. Basierend darauf kann ein elementares Ziel eg als ein Element aus der Vereinigung dieser Mengen definiert werden: $eg \in AO \cup AG \cup AA \cup AF$.

Um eine Vielzahl von nicht sinnvollen Kombinationsmöglichkeiten von elementaren Zielen zu vermeiden, werden im Anschluss daran die zwei Aspekte Spezialisierung und Anwendungskontext separat behandelt[6]: die beschriebenen elementaren Ziele werden im Sinne einer Attributierung um diese zwei Aspekte erweitert, mit SP der Menge aller Spezialisierungen und AK der Menge möglicher Anwendungskontexte. Entsprechend lassen sich dann attributierte, elementare Ziele (aez) wie folgt definieren: $aez = f(eg, sp, ap)$, mit $sp \in SP$ und $ap \in AK$ sowie mit der Attributierungsfunktion $f : EG \times SP \times AP \rightarrow AEG$, mit EG der Menge aller elementaren Ziele und AEG der Menge aller attributierten elementaren Ziele. Datenklassen und datenklassenspezifische Ziele werden in das Spezialisierungsattribut integriert. Dies erlaubt z.B. dem Analysegegenstand „Oberfläche“ die Spezialisierungsinformation „3D-Volumendaten“ hinzuzufügen. Ein Beispiel für ein hinzugefügtes anwendungsspezifisches Ziel ist der Analysegegenstand „Parameterraum“ eines Simulationsexperimentes für den Simulationshintergrund.

Basierend auf diesen Definitionen, lässt sich nun ein zusammengesetztes Ziel zz als eine Untermenge aller attributierten elementaren Ziele aez definieren: $zz \subseteq AEG$. Diese zusammengesetzten Ziele können nun mittels einer Beschriftungsfunktion $n : ZZ \times AK \rightarrow String$ beschriftet werden, wobei ZZ die Menge aller zusammengesetzter Ziele, AK den Anwendungkontext und $String$ die Menge möglicher Zielnamen für bestimmte Anwendungen darstellt. Diese Beschriftungsfunktion ermöglicht es, verschiedene Sichten auf Ziele in verschieden Anwendungsszenarien auszudrücken.

[6]Der Aspekt der Abhängigkeit wird hier nicht speparat einbezogen, da er sich in überwiegendem Maße in den Aktionen und Gegenständen der Analyse wiederfindet.

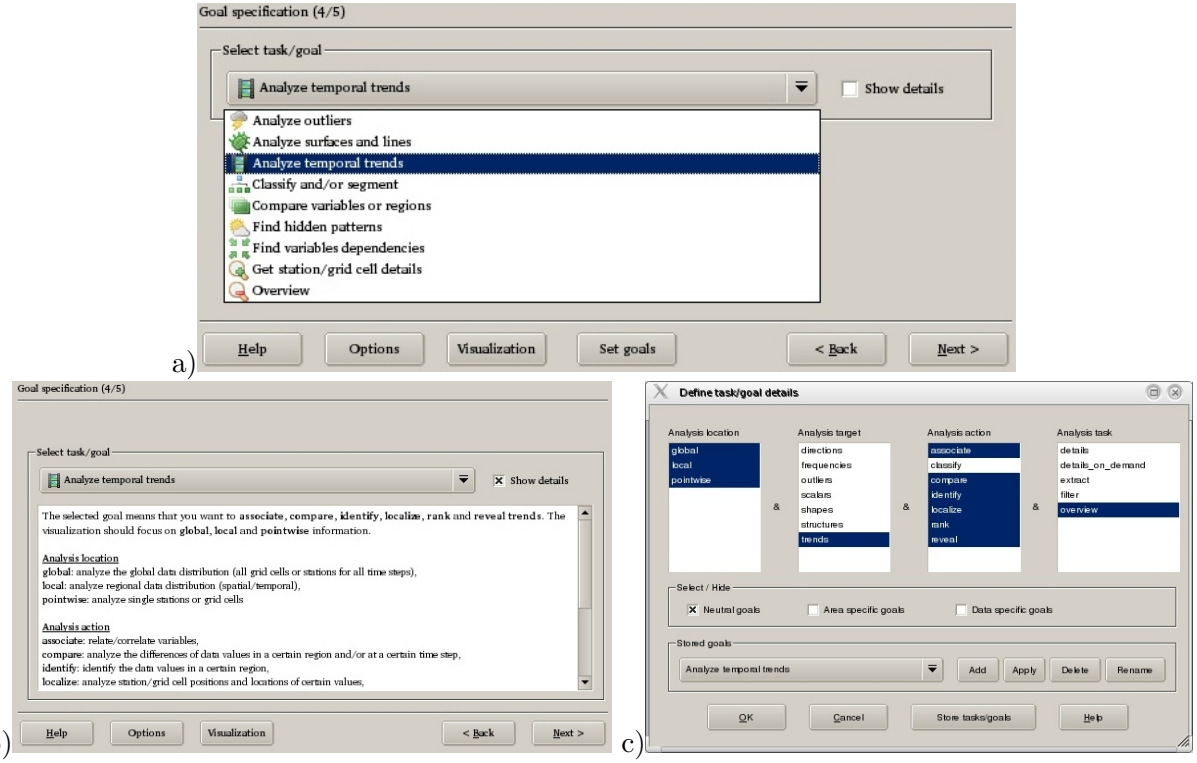


Abbildung 7.7: Erhebung von Zielstellungen; a) Auswahl einer komplexen Zielstellung; b) textuelle Beschreibung der Bedeutungen beteiligter, elementarer Zielstellungen; c) Konstruktion einer komplexen Zielstellung aus elementaren Zielen

Die Frage, welche Werte der spezifizierten Mengen von Zielen aus den Klassen *AO*, *AG*, *AA* und *AF* relevant sind, kann nun von praktischen Erwägungen einer speziellen VDM-Anwendung getrieben werden. Hier treten typischerweise bestimmte Klassen von Fragestellungen auf, die sich mit einer bestimmten Folge von VDM-Methoden beantworten lassen. Zusätzlich ist es sinnvoll, eine Grundmenge von Zielen aus der Literatur bereitzustellen, um allgemeine, anwendungsunabhängige Fragestellungen einzubeziehen. Durch verschiedene Arten der Beschreibung (und Beschriftung) lässt sich dabei eine hohe Verständlichkeit vorgegebener Ziele absichern und das Problem variierender Interpretationen der Zielstellungen reduzieren. Durch die Spezifikation von zusammengesetzten Zielen lässt sich die interne Struktur für Nutzer mit geringer Expertise verstecken. Dies ermöglicht es, interne und abstrakte Details zu verstecken und das Vokabular auf die Anwendung zuzuschneiden, das Visualisierungsdesign aber auf der Ebene elementarer Ziele durchzuführen.

7.2.3 Erhebung und Verwaltung von Analysezielen

Zur Eingabe der Zielstellungen wurde eine spezielle grafische Komponente entworfen und in das Framework SimEnvVis integriert. Die Komponente besteht aus einer Datenstruktur um elementare und zusammengesetzte Ziele zu speichern, einer grafischen Nutzerschnittstelle und I/O-Funktionalität zum Speichern und Laden von Zielen. Abbildung 7.7 zeigt die grafische Oberfläche der Komponente zur Eingabe und zum Management von Zielen. Zwei Ausschnitte aus XML-Dateien, die elementare und zusammengesetzte Ziele speichern, finden sich im Anhang (vgl. B.2).

Grundsätzlich beinhaltet dies ein dreistufigen Zugriff auf Analyseziele (in Abhängigkeit des Nutzerprofils): (1) einfache Auswahl von zusammengesetzten Zielen (Abb. 7.7a), (2) Auswahl von zusammengesetzten Zielen mit Erläuterung der darin enthaltenen elementaren Ziele (Abb. 7.7b) sowie (3) interaktive Komposition von Zielstellungen unter Einbeziehung der Bandbreite elementarer Zielstellungen (Abb. 7.7c).

Um das Verständnis der Bedeutung der Zielstellungen zu verbessern, wurde eine dynamisch generierte Hilfe zum aktuell ausgewählten Ziel (Abb. 7.7b) umgesetzt. Ferner wurden, um die Wiedererkennung und das Verständnis von Zielen zu verbessern, Ikonen für die zusammengesetzten Ziele eingesetzt (Abb. 7.7a,b).

Erfahrene Anwender können auch direkt neue Zielstellungen zusammensetzen (Abb. 7.7c). In den vier Spalten können elementare Ziele aus den vier Kategorien Analyseort (location), Gegenstand (target), Aktion (action) und Aufgabe (task) der Analyse ausgewählt werden. Darunter können je nach Bedarf neutrale, anwendungsspezifische und datenklassenspezifische Ziele ein- und ausgeblendet werden. Die so spezifizierten Mengen von elementaren Ziele können im unteren Teil des Dialoges (Abb. 7.7c) als zusammengefasste Zielstellungen verwaltet werden. Dies schließt das Hinzufügen, Ändern, Löschen, Umbenennen und Speichern zusammengesetzter Ziele ein.

7.2.3.1 Diskussion

In diesem Abschnitt wurden verschiedene Ansätze und Taxonomien aus der Literatur der Visualisierung und des visuellen Data Mining auf ihre Einbeziehung von Analysezielen hin untersucht. Basierend darauf wurden darin wichtige Aspekte systematisiert. Diese Untersuchung bildet die Basis für eine praktikable Spezifikation von Analysezielen. Dabei werden sowohl elementare Zielstellungen (basierend auf den identifizierten Aspekten) als auch zusammengesetzte Zielstellungen einbezogen. Neu an diesem Ansatz ist das dreistufiges Vorgehen und die Einbeziehung einer Beschriftungsfunktion, was einen ausgewogenen Kompromiss zwischen der praktischen Handhabung von Zielen und der Einbeziehung komplexer Zielstellungen ermöglicht. Auf diese Art und Weise lässt sich eine große Menge von Zielstellungen unter Einbeziehung verschiedener Aspekte an verschiedene Anwendungen anpassen und einfach benutzen.

Mit dieser Spezifikation lassen sich verschiedene Entscheidungen im VDM-Prozess unterstützen, von der Auswahl und Parametrierung einzelner VDM-Operatoren bis hin zur Erzeugung ganzer Operatornetzwerke. Sie lässt sich sowohl für allgemeine (general purpose) als auch für auf eine bestimmte Anwendung zugeschnittene Anwendungen (special purpose) einsetzen. Die im Stand der Forschung beschriebenen Ansätze lassen sich damit abbilden.

Es verbleiben jedoch noch Herausforderungen für zukünftige Arbeiten. So wurden, insbesondere mit Hilfe der Beschriftungsfunktion, erste Ziele an einen speziellen Anwendungshintergrund (hier Klimaforschung) angepasst. Eine systematische Untersuchung, inwieweit die Anwender durch den Einsatz der Ziele den VDM-Prozess besser steuern können, steht bisher jedoch noch aus.

Weiterhin wird im Bereich der „human computer interaction“ verstärkt mit komplexen Aufgabenmodellen gearbeitet, um die mit einem Softwaresystem durchführbaren Aufgaben zu modellieren. Der Einsatz solcher Aufgabenmodelle für die Visualisierung ist jedoch noch weitgehend offener Forschungsgegenstand (für einen ersten Ansatz hierfür vgl. Fuchs u. a. 2006).

7.3 Konzeption und Umsetzung eines Visualisierungsdesign-Wizards

In den vorangegangenen Kapiteln wurden eine Vielzahl von Techniken vorgestellt. Auf der einen Seite handelt es sich hierbei um Standardtechniken, die auf die speziellen Bedürfnisse der Anwendung zugeschnitten wurden. Auf der anderen Seite wurden neue Techniken entworfen oder in der Anwendung eher unbekannte Techniken einbezogen. Um einen hohen Grad an Nutzerakzeptanz beim Einsatz dieser Vielfalt an Techniken zu erreichen, sind spezielle Methoden zur deren Auswahl und Parametrisierung erforderlich. Vor allem ist es nicht trivial, für eine spezielle Anwendung ange-

passte Darstellungen für einen bestimmten Datensatz bei einer spezifischen Untersuchungsaufgabe zu finden. Zwar kann sich diese Arbeit auf eine Vielzahl von Ansätzen aus der Literatur stützen (vgl. hierzu auch den Stand der Forschung zum Visualisierungsdesign, Abs. 3.4.2), jedoch ergeben sich aus der Kombination von Anwendungskonventionen und allgemeinem Visualisierungswissen spezielle Probleme.

Ziel dieser Arbeit ist es, ein allgemeines Werkzeug zu entwerfen, welches die Nutzung der vorgestellten Techniken erleichtert, und für die Breite vorhandener Datensätze im Kontext der Konventionen der Anwendung geeignete Startbilder generiert. Ein erster, einfacher Ansatz in diesem Zusammenhang ist es, fest kodierte Metadaten und Ziele zu verwenden, und die verfügbaren Visualisierungen bezüglich dieser Informationen in einer Matrix fest zu verdrahten (vgl. z.B. Wagenknecht (2006)). Dieses Vorgehen ist jedoch nur begrenzt dazu geeignet, die Vielfalt auftretender Daten und die Breite in der Anwendung üblicher Aufgabenstellungen abzubilden.

Deswegen wird in diesem Abschnitt ein Wizard zur Anwenderunterstützung entworfen und dessen Umsetzung skizziert. Dabei werden dynamisch erweiterbare Regeln und Beschreibungen der vorhandenen Techniken ebenso wie eine breite Palette an Metadaten und Analysezielen einbezogen. Herausforderung hierbei ist es, allgemeine Regeln zur Erzeugung von Darstellungen mit anwendungsspezifischen Konventionen und Regeln zur Erzeugung von Darstellungen zu kombinieren. Dabei sollen in der Literatur isoliert betrachtete Ansätze zur Fällung einzelner Entscheidungen im Visualisierungsprozess (z.B. Anordnung von Achsen, Farbabbildung) systematisch behandelt werden.

7.3.1 Konzept

Ansatz dieser Arbeit ist es, den Anwender durch den Visualisierungsprozess zu leiten, ohne ihn dabei zu entmündigen. Dazu soll er dabei unterstützt werden, gute Einstiegsvisualisierungen für einen Problemkontext zu generieren. Beschränkungen der Nutzerfreiheit werden vorgenommen, nur wenn eine Visualisierung eine gewählte Aufgabenstellung nicht lösen kann (z.B. eine Technik kann ein bestimmtes Ziel nicht unterstützen) oder grundsätzlich nicht ausführbar ist (z.B. eine Technik(-implementation) unterstützt die Datencharakteristika nicht).

Ein typisches Vorgehen, einen Entscheidungsprozess interaktiv zu beeinflussen, ist der Einsatz eines Wizards (z.B. auch bei der Excel-Diagrammgenerierung). Dieser ermöglicht es in leicht verständlicher Art und Weise, getroffene Entscheidungen nachzuvollziehen und ggf. wieder rückgängig zu machen. Deshalb wird hier mittels eines solchen Wizards den durch die Techniken und deren Parameter aufgespannten Raum möglicher Darstellungen Schritt für Schritt eingeschränkt. Um die Nutzerautonomie zu gewährleisten, werden hierbei in jedem Schritt automatisch Vorschläge für Entscheidungen im Visualisierungsprozess unterbreitet, die interaktiv durch den Anwender adaptiert werden können. Basierend auf einer Beschreibung der Techniken (Abs. 7.3.2) wird dabei in jedem Schritt die Funktionalität der bereitstehenden Techniken zur Darstellung von Daten bestimmter Datencharakteristik sowie deren Unterstützung bei der Durchführung bestimmter Analyseaufgaben abgeglichen.

Nun muss die Frage beantwortet werden, welche Mechanismen zur Durchführung automatischer Entscheidungen für einen solchen Wizard geeignet sind. Prinzipielle Möglichkeiten sind ein *konstruktives Vorgehen* sowie ein *template-basiertes Vorgehen* (vgl. S. 38ff.). Während beim *konstruktiven Vorgehen* die Darstellungen selbst durch Regeln generiert werden, wird bei einem *template-basierten Vorgehen* aus einer Menge von Techniken ausgewählt und deren Parameter angepasst.

Die im Rahmen dieser Arbeit entworfenen Techniken wurden speziell auf die Klimafolgenforschung zugeschnitten und dabei eine Vielzahl von Interaktionsmöglichkeiten integriert. Diese Techniken sollen auch unabhängig vom Visualisierungsdesign verwendet werden können. Entsprechend ist ein

template-basiertes Vorgehen für das Visualisierungsdesign besonders geeignet. Dabei sollen im Sinne eines pragmatischen Vorgehens wichtige Regeln zu Visualisierungskonventionen aus der Anwendung betrachtet, und die verschiedenen Einflussgrößen und die Beschreibungen der Visualisierungstechniken einbezogen werden. Zusätzlich ist es auch erforderlich - zum Teil wiedersprechendes - anwendungsunabhängiges Wissen aus der Visualisierungsliteratur zu modellieren (vgl. Abs. 7.3.3).

Um die Vielzahl potentiell in Frage kommenden Darstellungen zu strukturieren, wurden die Darstellungstechniken – begrifflich in Anlehnung an das template-basierte Vorgehen beim Visualisierungsdesign – in zwei Ebenen unterteilt: (1) allgemeine Visualisierungstemplates, welche verschiedene Darstellungen von Daten gleicher Datenklassen mit ähnlichen visuellen Metaphern zusammenfassen (z.B. 2D-Darstellungen für Daten auf regulären Gittern), und (2) spezielle Module dieser Visualisierungstemplates, die spezielle Arten von Abbildungen der Daten auf die visuellen Attribute des konkreten Moduls repräsentieren (z.B. Farbe und Isolinien). Der Vorteil dieser Untergliederung liegt darin, dass während der des Visualisierungsdesign lediglich die grundlegende Abbildung der Daten festlegt wird, einzelne Entscheidungen jedoch während der eigentlichen Visualisierung jederzeit rückgängig gemacht können.

Basierend auf dieser Trennung lassen sich die verschiedenen im Visualisierungsdesign durchzuführenden Entscheidungen wie folgt in vier Ebenen gliedern:

- Auswahl von allgemeinen Visualisierungstemplates,
- Auswahl von Visualisierungsmodulen dieser Templates,
- Abbildung der Variablen auf visuelle Attribute und
- Feinparametrisierung (z.B. Auswahl einer speziellen Farbskala).

Somit lassen sich insbesondere Entscheidungen im Mapping und Rendering abbilden. So können in jeder dieser vier Ebenen verschiedene Entscheidungen gefällt werden, welche der Anwender bei Bedarf anpassen kann.

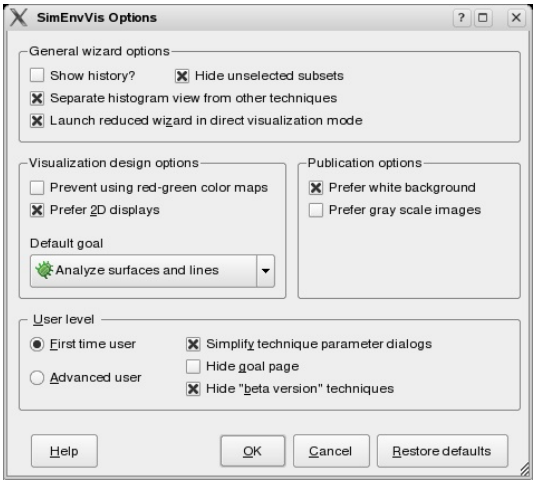


Abbildung 7.8: Dialog zur Verwaltung der Nutzerprofile

Für diese Entscheidungen sind neben Metadaten und Analysezielen auch weiter Einflussgrößen relevant und sollen im folgenden kurz diskutiert werden.

Anwendungskontext. Die Spezifika der Anwendung durchziehen alle Bereiche der Konzeption eines Visualisierungsdienstes für Klimadaten. Insbesondere wurde diese bereits bei der Umsetzung der einzelnen Visualisierungstechniken mit einbezogen. Dies schließt spezielle Namenskonventionen (für den Wizard und die einzelnen Techniken) und Konventionen zur Darstellung von Daten bestimmter Datenklassen ein. Konventionen zur Darstellung werden über Regeln einbezogen.

Nutzerprofile. Die Präferenzen des Anwenders werden explizit mit einbezogen (vgl. Abb. 7.8). Hierbei sollen sich drei grundlegende Klassen von Anwendervorlieben unterscheiden:

1. Präferenzen, welche das Aussehen und die Abfolge des Wizards steuern (z.B. Verstecken der Analyseziele),
2. Präferenzen, welche die Steuerung der Visualisierungstechniken beeinflussen (z.B. Verstecken von Parametern im Nicht-Experten-Modus („first-time user“))
3. Präferenzen und Wahrnehmungsfähigkeiten, welche Entscheidungen bei der Generierung der Darstellung beeinflussen (z.B. Bevorzugen planarer Darstellungen, Vermeidung von Rot-Grün-Farbskalen).

Abbildung 7.9 (links) fasst grundlegende Einflüsse des Wizards bei der Fällung einzelner Entscheidungen zusammen.

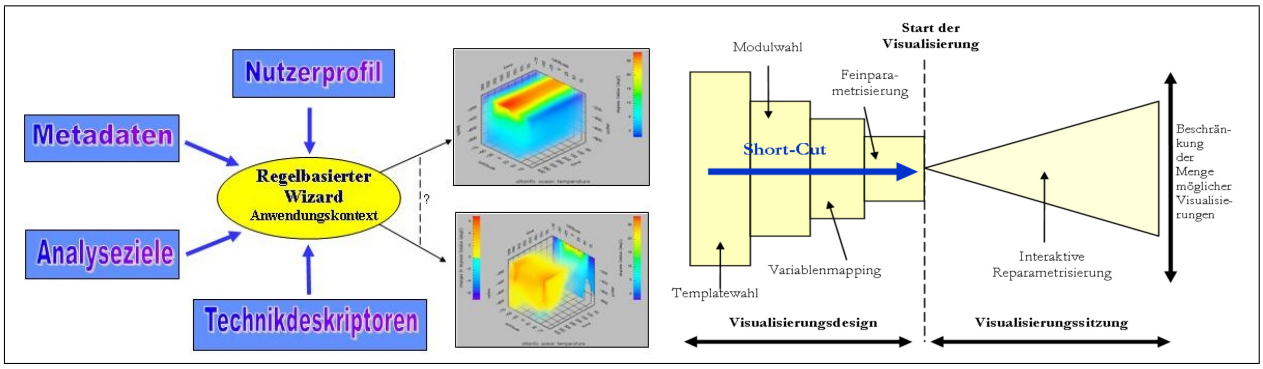


Abbildung 7.9: Konzeption des Visualisierungsdesign-Wizards (vgl. auch Lange u. a. 2006): a) Grundschema des Visualisierungsdesign für Klimadaten; b) Schematischer Ablauf einer visuellen Analyse (in zwei Phasen): Visualisierungsdesign und -sitzung

Durch den Entscheidungsprozess werden schrittweise Visualisierungstemplates eingeschränkt. Während der Visualisierungssitzung kann der Nutzer die dabei getroffenen Entscheidungen interaktiv verändern, und die Menge möglicher Darstellungen nach seinen Kenntnissen und Zielen erweitern.

Um die Anwender mit der Vielzahl an verfügbaren Visualisierungsfunktionalitäten nicht zu überfordern, wurde eine beschleunigte Erzeugung der Startvisualisierung in den Wizard integriert. In jedem Schritt ist es so möglich, basierend auf den bis dahin getroffenen Entscheidungen die nachfolgenden Schritte mit automatisch bestimmten Standardbelegungen zu füllen und die so parametrisierte Visualisierung direkt zu starten. Abbildung 7.9b illustriert, wie der Raum möglicher Visualisierungen basierend auf den vier Entscheidungsebenen zuerst mit Hilfe des Wizards eingeschränkt wird. Mittels eines „Fast-Visualization-Button“ (short cut) können dabei Standardbelegungen für nachfolgende Stufen gewählt werden.

Grundsätzlich soll dieser Prozess so transparent wie möglich gestaltet werden, ohne den Anwender zu überfordern. So sollen zum einen Hilfestellungen für gefällte Entscheidungen und deren Basis (z.B. eine gewählte Zielstellung, vgl. Abb. 7.7) bereitgestellt werden, zum anderen aber auch ungeeignete Techniken ausgeblendet werden.

Im den folgenden beiden Abschnitten sollen nun die beim Design des Wizards konzipierten Beschreibungen von Visualisierungsdeskriptoren (Abs. 7.3.2) und die Modellierung des Designmechanismus (Abs. 7.3.3) im Detail vorgestellt werden.

7.3.2 Beschreibung von Visualisierungstechniken

Für das Visualisierungsdesign ist eine Beschreibung der Eigenschaften der Visualisierungstemplates erforderlich. Dabei muss eine Vielzahl von Informationen verwaltet werden. Im Vordergrund stehen hierbei die flexible Erweiterbarkeit um neue Templateeigenschaften, die leichte Einbindbarkeit neuer Templates und ein flexibles Management der dazu erforderlichen Deskriptoren. Grundlegend lässt sich dabei zwischen

- Schnittstelleneigenschaften: bestimmen, von welchem Typ eine Technik ist und wie diese gestartet und parametrisiert werden kann,
- Designeigenschaften: beschreiben die Eigenschaften der Technik und einzelner Parameter bezüglich Metadaten, Zielen und anderen Einflussgrößen und
- Annotationseigenschaften: beinhalten Hilfestellungen bei der Präsentation der Techniken und Parameter im Entscheidungsprozess (z.B. Technikname, Beschreibung, Ort einer Hilfedatei)

unterscheiden. Diese Eigenschaften finden sich auf jeder der verschiedenen Ebenen Visualisierungstemplate, Visualisierungsmodul, Variablenmapping und Visualisierungsparameter wieder. Ausgehend insbesondere von praktischen Anforderungen des Visualisierungsdesigns bezieht ein Deskriptor eines Visualisierungstemplate die folgenden Eigenschaften ein:

- Basisdaten: Name, Typ (z.B. OpenDX oder OpenGL), Startfunktion/Netzwerk, Hilfedatei, Beschreibung, Entwicklungsstatus, Grundvoraussetzungen zur Ausführung (Grenzen und Eignung des Templates bezüglich wichtiger Metadaten: Anzahlen von unterstützen Dimensionen, Merkmalen und Datensätzen, unterstützte Gittereigenschaften, ...),
- Module:
  - Basisdaten: Name, Beschreibung,
  - Schnittstellenbeschreibung: durch Wahl welcher Parameterwerte wird eine bestimmte Modul angesprochen,
  - Regeln zur Eignungsbestimmung eines Moduls,
  - Variablenmapping
    - visuelle Attribute, auf die Dimensionen und Merkmale abgebildet werden können: Name, Art,
    - Regeln der Eignungsbestimmung eines Variablenmapping,
- Parameter:
  - Basisdaten: Name, Typ (z.B. Selektor/Combobox), Klasse (modulbezogen, allgemein),
  - Parameterwerte: Namen, Werte, Verfügbarkeit,
  - Regeln zur Parameterberechnung.

Die Deskriptoren der Visualisierungstemplates wurden im XML-Format umgesetzt. Der folgende Ausschnitt zeigt beispielhaft den XML-Deskriptor des 3D-Visualisierungstemplates (für mehr Details vgl. Abs. B.3 im Anhang):

```
<Technique text = "3D visualization">
   <VisSystemType text = "OpenDX"/>
   <NetName text = "3DTech.net"/>
   <HelpFile text = "3DVis.html"/>
   <ShortTextDescription text= "The 3D visualization ..."/>
   <DevelopmentState text="PreRelease"/> ...
   <DimensionsDescriptor>
      <generalDimensionality>
        <minoptmax min="3" optmin="3" optmax="3" max="4"/>
      </generalDimensionality> ...
   </DimensionsDescriptor>
   <GridDescriptor>
      <regular text ="yes"/>
      <blockStructured text ="yes"/>
      <scattered text ="no"/> ...
   </GridDescriptor> ...
   <DatSetDescription>
      <NrDataSets>
        <minoptmax min="2" optmin="8" optmax="800000" max="Unlimited"/>
      </NrDataSets>
      <NrAttributes>
        <minoptmax min="1" optmin="1" optmax="1" max="Unlimited"/>
      </NrAttributes>
   </DatSetDescription>
   <Parameters>
      <Parameter name = "GMType" type = "Selector">
         <SelectorValue name = "Volume rendering" value = "4"/>
         <SelectorValue name = "Slices (data)" value = "3"/> ...
      </Parameter>
```

```
      <Parameter name = "ScaleType" type = "Selector"> ... </Parameter>
      <Parameter name = "GridType" type = "Selector"> ... </Parameter>
      <Parameter name = "RenderMode" type = "Selector">
         <SelectorValue name = "Software" value = "1"/>
         <SelectorValue name = "Hardware" value = "2" availability="checkGLXExtension"/>
      </Parameter>
      <Parameter name = "Background_color" type = "Selector" class = "general">
         <SelectorValue name = "black" value = "1"/>
         <SelectorValue name = "dark gray" value = "2"/>
         <SelectorValue name = "gray" value = "3"/>
         <SelectorValue name = "white" value = "4"/>
         <Rules>
            <Rule>[Parameter] hasUserPreference ("PreferWhiteBackground") ? [=4, =2];</Rule>
         </Rules>
      </Parameter>
   </Parameters>
   <VariationSet name = "General mapping">
      <Variation text = "Cutting planes">
         <ParameterKombination>
            <ParameterValue name = "GMType" value = "6"/>
            <ParameterValue name = "ScaleType" value = "1"/>
            <ParameterValue name = "GridType" value = "2"/>
            <ParameterValue name = "RenderMode" value = "1"/>
         </ParameterKombination> ...
         <Rules>
            <Rule>[ModuleVariation] exists(Dimension, "Name", "longitude")
                 AND exists(Dimension, "Name", "lattitude") ? [+=0.05];</Rule> ...
         </Rules>
      </Variation> ...
</Technique>
```

7.3.3 Ein Entscheidungsmechanismus zur Steuerung des Visualisierungsprozesses

Zielstellung bei computergestützten Entscheidungen im Visualisierungsprozess ist es, Anwendungskonventionen und Visualisierungswissen zu formalisieren und zu automatisieren. Im speziellen sollen hierbei Erfahrungs- und Literaturwissen für die spezielle Anwendung explizit modelliert gestellt werden. Herausforderung hierbei ist es, Nichtexperten transparent bei einer Vielzahl von Entscheidungen auf verschiedenen Ebenen zu unterstützen. Probleme treten auf, da es sich hierbei um heterogenes, zum Teil widersprüchliches Wissen handelt, wobei viele Regeln eher unscharf formuliert sind. In der Literatur wurde der Einfluss einzelner Einflussfaktoren auf die Durchführbarkeit einer Visualisierungsaufgabe untersucht, wobei jedoch andere Kontextgrößen weitgehend stabil gehalten wurden. Damit hat zum Beispiel eine Aussage der Art „Eine Darstellungstechnik V unterstützt ein Ziel Z“ nur begrenzte Aussagekraft, da dies auch von den Datencharakteristika und der gewählten Parametrisierung stark abhängen kann.

In diesem Spannungsfeld ist ein flexibler Mechanismus erforderlich, der auch unscharfe und sich widersprechende Aussagen mit einbezieht, diese Entscheidungen jedoch für den Anwender weitgehend transparent bleiben. Entsprechend wurde ein zweigleisiger Mechanismus umgesetzt (vgl. auch Theisel 1994), wobei ein vektor-basiertes Vorgehen mit einem Regelmechanismus unter Einschluss von Fuzzy-und Nichtfuzzy-Regeln kombiniert wird (vgl. hierzu auch Kriterien auf S. 38). Dabei liegt der Fokus auf Entscheidungen im Mappingprozess, wobei insbesondere Metadaten, Ziele und Nutzerpräferenzen einbezogen werden sollen. Hierbei werden die speziell für die Bedürfnisse der Anwendung zugeschnitten interaktiven Visualisierungstechniken mit einem templatebasierten Vorgehen ausgewählt und parametrisiert. Hierzu wurden die beiden folgenden Methoden zur Eignungsbestimmung untersucht:

- Eignungsbestimmung basierend auf einem festkodierten Komponentenvektor, bestehend aus fuzzy-logischen Einzeleignungen bezüglich bestimmter Eigenschaften eines Visualisierungstemplates und
- Eignungsbestimmung basierend auf Regeln für alle anderen Entscheidungen.

Ein solcher Mechanismus muss weiterhin die verschiedenen Kategorien von anwendungsspezifischen Regeln und Konventionen für die Klimafolgenforschung einbeziehen (vgl. auch Abs. 3.2.2):

- Regeln zu allgemeinen Metaphern in diesem Umfeld
  - bei der Abbildung von Merkmalen (z.B. Farbkodierung, Isolinien, elementare Methoden der Vektorfelddarstellung)
  - bei der Abbildung des Raum-/Zeitbezuges (z.B. 2D-Kartendarstellungen, Animation zur Repräsentation der Zeit, Abbildung in verschiedenen Projektionen)
- spezielle Abbildungsregeln für bestimmte Arten von Variablen (z.B. angepasste Farbwahl für die Merkmale Temperatur und Niederschlag)
- Konventionen zur Einblendung zusätzlicher Kontextinformationen bei Bedarf (z.B. Darstellung der Weltkarte oder des Reliefs bei einer Erdoberflächendarstellung)

Darüber hinaus soll der zu entwickelnde Mechanismus den Anwendern aber auch für sie eher unbekannte Metaphern vorschlagen, welche für ein spezielles Problem möglicherweise besser geeignet sind als die Standardmethoden der Anwendung. Deswegen muss auch anwendungsunabhängiges Visualisierungswissen einbezogen werden können. Weiterhin muss modelliert werden können, dass bei einem gewissen Situation eine Technik völlig ungeeignet sein kann, unabhängig vom Rest der Einflussfaktoren („Totschlagkriterien“).

Eignungsbestimmung basierend auf einem festkodierten Komponentenvektor. Für die Auswahl der grundlegenden Visualisierungstemplates steht insbesondere die grundlegende Eignung bezüglich wichtiger Datencharakteristika wie Anzahl der abhängigen und unabhängigen Variablen, der Gittertyp und die Anzahl der darstellbaren Datensätze im Vordergrund. Deswegen wurde für diesen Entscheidungsschritt ein einfaches, vektor-basiertes Vorgehen ausgewählt, welches insbesondere aus die Menge vorhandener, essentieller Metadaten mit den Templateeigenschaften in einem Fuzzy-Vektor bestehend aus n Eignungsparametern x_i ($x_i \in [0,1]$) abspeichert (für einen alternativen Fuzzy-logischen Ansatz vgl. auch Jung 1998). Zusätzlich können die einzelnen Komponenten je nach Bedarf durch n Gewichte w_i ($\sum_{i=1}^{n} w_i = 1$) variabel gewichtet werden. Basierend darauf lässt sich die Eignung E wie folgt berechnen:

$$E = \begin{cases} \frac{1}{m} \sum_{i=1}^{n} b_i \, w_i \, x_i & \text{falls } \forall x_i : \; x_i \neq 0 \\ 0 & \text{sonst.} \end{cases} \tag{7.1}$$

Hierbei legen die Werte $b_i \in \{0, 1\}$ fest, ob eine Eignungskomponente für ein Template überhaupt relevant ist oder nicht, und $m = |\{b_i : b_i = 1\}|$ ist die Anzahl der relevanten Eignungskomponenten. Insbesondere hat die vorgestellte Funktion die wichtige Eigenschaft, dass die Eignung E eines Templates automatisch gleich Null ist, falls eine der Einzeleignungen gleich Null ist, also eine der essentiellen Eigenschaften nicht erfüllt wird.

Weiterhin wurden zur Berechnung der einzelnen Eignungsparameter x_i Funktionen entworfen, welche nicht nur eine grundlegende Eignung, sondern auch eine unscharfe Eignungsbestimmung ermöglichen (vgl. hierzu auch Theisel 1994). Insbesondere betrifft dies eine angepasste Eignungsparameterbestimmung für diskrete Metadaten (z.B. Anzahl unterstützter Datenrekords). Hierbei muss modelliert werden, dass zur Ausführung eines Templates ein bestimmter Metadatenwert einen bestimmten Wert nicht über- oder unterschreiten darf (min, max), sowie dass es einen Präferenzbereich (*minopt*, *maxopt*) gibt, in dem die Eignung immer mit 1 zu bewerten ist. Darüber hinaus kann eine Technik (zumindest theoretisch), auch bezüglich des Maximums (oder auch des Minimums) unbeschränkt sein. Entsprechend wurde die folgende Berechnungsvorschrift von Eignungsparametern

zur Quantifizierung der Eignung einer diskreten Größe g entworfen:

$$x_i = f(g, min, max, optmin, optmax) = \begin{cases} 0 & \text{falls } g < min \vee g > max \\ 1 & \text{falls } g \geq optmin \wedge g \leq optmax \\ \frac{g-min+1}{minopt-min+1} & \text{falls } min \leq g < optmin \\ \frac{max+1-g}{max+1-maxopt} & \text{falls } optmax < g \leq max \wedge max \neq \infty \\ \frac{1}{\sqrt{g-maxopt+1}} & \text{falls } optmax < g \wedge max = \infty \end{cases} \quad (7.2)$$

Innerhalb des Optimalbereiches ist die Eignung gleich Eins, während sie außerhalb der Grenzen min und max gleich Null wird, jedoch genau bei min und max ungleich Null. Im Übergangsbereich zwischen *min* und *optmin* sowie *max* und *optmax* wurde ein linearer Eignungsverlauf modelliert. Falls das Maximum unbeschränkt ist, wird bei Werten von $g \geq optmax$ basierend auf einer umgekehrten Wurzelfunktion ein (im Unendlichen gegen Null konvergierender) asymptotischer Funktionsverlauf erzeugt.

Eignungsbestimmung basierend auf Regeln. Vorteile beim Einsatz eines Regelmechanismus sind neben der leichten Verständlichkeit und Kommunizierbarkeit von Regeln aufgrund einer sprachnahen Form und der flexiblen Einbeziehbarkeit verschiedener Einflussfaktoren insbesondere deren Fähigkeit, widersprechenden Aussagen miteinander zu verknüpfen. Für das Design solcher Regeln lassen sich die folgenden Kriterien identifizieren:

- Reproduzierbarkeit des Einflusses einer Regel auf den Entscheidungsprozess,
- flexible Einbeziehung (aussagen-)logischer Ausdrücke zur Repräsentation komplexer Zusammenhänge,
- flexible Wichtung einzelner Regeln, insbesondere vor dem Hintergrund von anwendungstypischen Konventionen,
- leichte Erweiterbarkeit um neue Regeln,
- einheitliche Syntax für die verschiedenen Arten von Entscheidungen.

Entsprechend wurde in der Umgebung Lex/Yacc eine Grammatik umgesetzt, die eine leicht erweiterbare Palette von leicht verständlichen Regeln umsetzt. Hierbei hat eine Regel die folgende Syntax:

$$Regel :== Entscheidungskontext + Bedingung + Zuweisung. \quad (7.3)$$

Der Entscheidungskontext beinhaltet, ob ob es sich um eine Eignungsbestimmung (eines Templates, eines Moduls oder einer Variablenmappings) oder um eine Parameterberechnung handelt. Die Bedingung stellt einen aussagenlogischen Ausdruck dar, der festlegt, ob die Regel angewendet werden kann oder nicht. Sie ist aufgebaut aus einer Menge elementarer Bedingungen, welche boolsche Anfragen an Technikdeskriptor, Metadaten-, Ziel- und Anwenderpräferenzbeschreibung darstellen. Im Rahmen dieser Arbeit wurde die folgenden zwei Typen von elementaren Bedingungen betrachtet:

1. direkte Anfrage einer Bedingung: z.B. 'exists(Dimension, "Name", "longitude")' oder 'hasGoal ("identify")'
2. Extraktion eines Wertes und Bedingungsauswertung mittels Vergleichsoperationen: z.B. 'getMetaData(Attributes, "Number") > 2'

Die Zuweisung spezifiziert die Anpassung der Eignungs- bzw. Parameterwerte[7]. Hierbei wird insbesondere zwischen einem positiven ('+=W') oder einem negativen ('-=W') Einfluss auf die Eignung bzw. auf den Parameterwert unterschieden. „W" stellt hierbei die Wichtung einer Regel dar.

[7] im Falle, dass die Regelbedingung erfüllt ist

Darüber hinaus lassen sich auch auch andere Anpassungen vornehmen, zum Beispiel die eindeutige Zuweisung eines Wertes basierend auf einer vordefinierten Funktion.

Zur Bestimmung einer Eignung basierend auf einer Menge von Regeln lassen sich verschiedene Berechnungsvorschriften konstruieren. Ein wichtiges Kriterium für die Verständlichkeit einer resultierenden Eignung ist, dass sie den Wert von Eins nicht übersteigt[8]. Umgekehrt lässt sich eine Eignung kleiner-gleich Null auch auf die Aussage „Nichteignung“ abbilden. Weiterhin sollen sowohl positive, neutrale und negative Einflüsse auf die Eignungsbestimmung modelliert werden. Entsprechend wurde eine einfache pragmatische Berechnungsvorschrift für die Eignung basierend auf n $(= n_p + n_n)$ Regeln R_i entworfen:

$$E = f(R_1 \dots R_n) := \frac{\sum_{i=1}^{n_p} b_i \; w_i}{\sum_{i=1}^{n_p} w_i} - \sum_{i=n_p+1}^{n} b_i \; w_i \tag{7.4}$$

Die Eignung wird hierbei als die Subtraktion der Summe negativ gewichteter Regeln $(R_{n_p+1} \dots R_n)$ von der normierten Summe der positiv gewichteten Regeln $(R_1 \dots R_{n_p})$ dargestellt. Die n_p positiv gewichteten Regeln werden, wenn ihre Bedingung b_i erfüllt ist, mit den Gewichten w_i aufsummiert. Diese Summe wird durch die Summe aller positiven Gewichte w_i auf das Intervall $[0, 1]$ normiert, so dass sich im Falle der Erfüllung aller Regelbedingungen b_i für die Eignung ein Wert von Eins, und bei Nichterfüllung aller Regelbedingungen b_i ein Wert von Null ergibt. Zusätzlich zu diesem Term der positiv gewichteten Regeln lassen sich mit dem Subtrahenden im Sinne eines zusätzlichem „Gewichtes“, n_n Regeln definieren, welche die resultierende Eignung bei Bedarf (unnormiert) reduzieren können.

Mit diesem pragmatischen, vektorbasierten Ansatz lassen sich bedingte Regeln mit flexiblen Gewichtungen zur Eignungsbestimmung definieren, wobei die Eignung einer Technik bei bestimmten Kontextbedingungen zusätzlich explizit reduziert werden kann. Dies soll am Beispiel der Regeln für die Technikvariation „TableMode“ (stellt mehrere Kurvendarstellungen in einer Tabelle von Plots dar) des 1D-Techniktemplates illustriert werden:

```
[ModuleVariation] getMetaData(Attributes, "Number") > 3 ? [+=0.7];
[ModuleVariation] hasGoal ( "overview" ) ? [+=0.1];
[ModuleVariation] hasGoal ( "scalar" ) ? [+=0.1];
[ModuleVariation] hasGoal ( "compare" ) ? [+=0.1];
[ModuleVariation] getMetaData(Attributes, "Number") == 1 ? [-=0.3];
```

Hier legt die erste Regel fest, dass die Variation eine Eignung von 0.7 erhält, falls die Anzahl der Merkmale 3 übersteigt. Zusätzlich können im Fall der Wahl der Ziele Überblick, Skalar und Vergleich je weitere 0.1 Eignungspunkte hinzukommen (2.-4. Regel). Andere Ziele oder Metadaten haben keinen Einfluss. Zusätzlich soll modelliert werden, dass in dem Falle, dass lediglich ein Merkmal vorliegt, die Tabellendarstellung unabhängig von den gewählten Zieleinstellungen nicht geeignet ist. Entsprechend wird für diesen Fall der Wert 0.3 abgezogen, so dass selbst im Falle der Wahl aller drei Ziele eine Eignung von Null resultiert. Ein solches Verhalten lässt sich auch auf andere Art als durch eine zusätzliche Subtraktion modellieren (zum Beispiel über eine Erweiterung der Bedingungen der Regeln 2 bis 4), hat sich jedoch für die Regelmodellierung als leicht benutzbares, transparentes Werkzeug herausgestellt.

7.3.4 Umsetzung und Nutzerschnittstelle

Ziel bei der Umsetzung des Visualisierungsdesign-Mechanismus ist es, verschiedenartige Techniken der Komponentenbibliothek (z.B. OpenDX, OpenGL, ...) mit einem einheitlichen Mechanismus anzusteuern, ohne das bei der Benutzung Wissen über die Art der Techniken erforderlich ist. Dazu

[8] Dies hat die Anwender des Systems in einer frühen Version irritiert.

wurde eine Kopplungskomponente umgesetzt, welche es erlaubt, Techniken mit unterschiedlichen Schnittstellen anzusteuern. Dadurch lassen sich unabhängig von der speziellen Entwicklungsumgebung Visualisierungen starten, Parameter einstellen sowie Darstellungen schließen[9].

Um die umgesetzten Mechanismen für die Anwender zugänglich zu machen, ist weiterhin eine geeignete Nutzerschnittstelle erforderlich, welche das Verständnis für die verfügbaren Techniken und damit verbundenen Entscheidungen erleichtert, ohne den Anwender zu überfordern oder einzuschränken. Deswegen wurde, speziell zur Visualisierung von Daten aus der Simulationsumgebung SimEnv (vgl. Flechsig u. a. 2006, 2007) ein Wizard entworfen, der neben den Design-Entscheidungen auch die Auswahl, das Filtering der Daten und die Spezifikation von Metadaten (vgl. Abb. 7.5) sowie die Auswahl von Analysezielen unterstützt (vgl. Abb. 7.7). Dieser Wizard beinhaltet neben den verschiedenen integrierten Visualisierungen die Kernfunktionalität des speziell auf Daten aus dem Bereich der Klima- und Klimafolgenforschung zugeschnittenen Visualisierungsframeworks SimEnvVis (vgl. Lange u. a. 2006; Nocke u. a. 2007).

Auf den Seiten des Wizards werden dem Anwender Vorschläge von verfügbaren Visualisierungsentscheidungen (auf den verschiedenen Ebenen) präsentiert, welche interaktiv angepasst werden können. Neben dynamisch generierten Hilfestellungen zu Problemen bestimmter Entscheidungen (z.B. von eher ungeeigneten Templates) erhält der Anwender hier Vorschaubilder und Hilfetexte zu den verfügbaren Optionen. Abbildung 7.3.4 veranschaulicht die graphische Schnittstelle des Wizards zur halbautomatischen Auswahl von Techniktemplates und Technikmodulen.

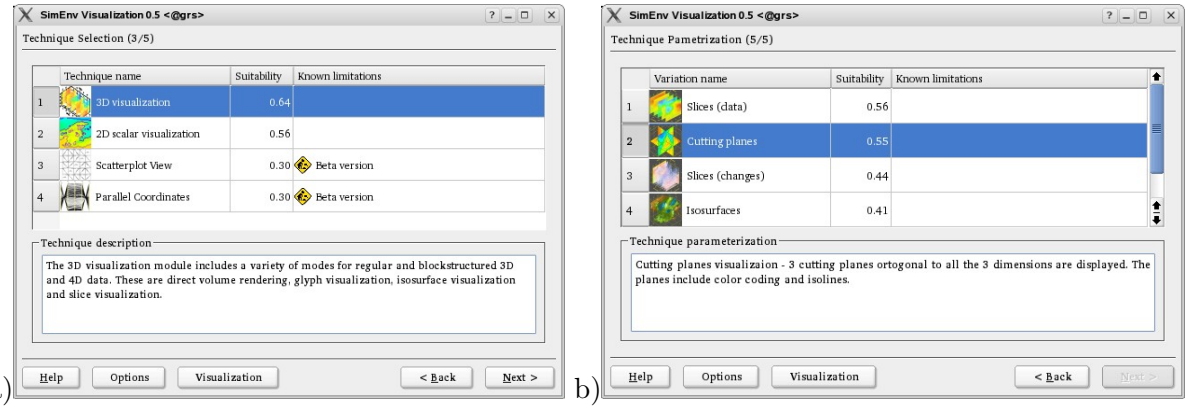


Abbildung 7.10: Screenshots des Visualisierungsdesign-Wizards im Framework SimEnvVis: a) Auswahl von Techniktemplates; b) Auswahl von Technikmodulen

7.3.5 Diskussion

In diesem Abschnitt wurde der Entwurf und die Umsetzung eines Mechanismus zum Visualisierungsdesign speziell für die Anwendung auf Daten aus der Klimafolgenforschung vorgestellt. Als hierfür essentielle Grundlage wurden Deskriptoren für Visualisierungstechniken (bzw. Visualisierungstemplates) beschrieben, welche neben der rein technischen Ansteuerung der eingebundenen Techniken insbesondere Auswahl- und Parametrisierungsentscheidungen auf verschiedenen Ebenen unterstützen. Im Sinne eines pragmatischen Vorgehens wurde ein Entscheidungsmechanismus entworfen und umgesetzt, der fuzzy-logische Elemente und „sprachnahe“ Regeln einbezieht. Dieser Mechanismus beinhaltet ein zweistufiges Vorgehen, in dem in einem ersten Schritt die grundlegende Eignung eines Visualisierungstemplates basierend auf essentiellen Metadaten bestimmt wird, und im anschließenden zweiten Schritt flexibel Regeln für beliebige Eignungsgrößen aufgestellt und behandelt werden können. Speziell an diesem Ansatz ist die einheitliche Behandlung von Entscheidungen

[9] und konzeptionell auch Techniken miteinander koppeln, z.B. über Brushing

auf den Ebenen von der Modulauswahl über das Variablenmapping hin zu einzelnen Parametrisierungen.

Iterativ wurden dabei basierend auf Konventionen aus der Anwendungsliteratur und in Interaktion mit den Klimaforschern wichtige Regeln identifiziert und in den Mechanismus einbezogen. Ziel hierbei war es nicht, Vollständigkeit oder Allgemeingültigkeit zu erreichen, sondern für die Anwender plausibel nachvollziehbare Entscheidungen vorzuschlagen, welche jederzeit interaktiv angepasst werden können.

Speziell wurde der Visualisierungsdesign-Mechanismus für die Visualisierung von Daten aus der Simulationsumgebung SimEnv entwickelt und getestet. Dabei lässt sich entweder der gesamte Mechanismus automatisiert betreiben (vollautomatisch) oder mit Hilfe des Wizards in die einzelnen Entscheidungen eingreifen. Mit Hilfe eines Logging-Mechanismus werden die getroffenen Entscheidungen protokolliert. Als erstes Resultat daraus ergibt sich, dass die Stufe der Automatisierung (Nutzung des Fast-Visualization-Button) stark von der Umgebung[10] und den Vorlieben der einzelnen Anwender abhängt. Um diesen Vorlieben entgegenzukommen, lässt sich der Wizard für jeden Anwender passend parametrisieren, so dass z.B. die manuelle Anpassung elementarer Ziele erst versierten Anwendern angeboten wird. Eine systematische Validierung des Visualisierungsdesigns mit einer größeren Nutzergruppe steht jedoch noch aus.

Darüber hinaus verbleiben hier eine Vielzahl von Herausforderungen für zukünftige Arbeiten. So führt das derzeitige Vorgehen, die Regeln per Hand in XML zu spezifizieren, an Grenzen. Mit steigender Zahl an Regeln pro Entscheidung und verschiedenen einzubeziehenden Einflussfaktoren wird die Spezifikation solcher Regeln zunehmend unübersichtlich, was insbesondere auch das Verständnis von deren Wirkung bei wechselnden Kontextbedingungen einschließt. Als Lösung für dieses Problem bedarf es einer Regelentwurfs- und Testumgebung für den Entwickler oder den versierten Anwender.

Desweiteren wurde die vielversprechende Unterstützung bei der Zuordnung der Variablen zu einzelnen visuellen Attributen[11] im Rahmen der Arbeit untersucht (vgl. betreute Arbeit Klembt u. Krüger (2006)) und im Entscheidungsmechanismus vorgesehen, jedoch bisher noch nicht vollständig umgesetzt und getestet.

Abbildung 7.11: Darstellung der Analysehistorie im Visualisierungsdesign-Wizard: links: sequenzielle Abfolge; rechts: logische Abhängigkeiten

Darüber hinaus bezieht der Mechanismus derzeit nur Entscheidungen für Daten auf einem regelmäßigem Gitter ein, und unterstützt lediglich die Auswahl und Parametrisierung einzelner, ungekoppelter Darstellungen. Die automatische Unterstützung bei der visuellen Kopplung von Datenteilmengen ist ein wichtiger Weg, um Konflikte zwischen Einflussgrößen (z.B. die gleichzeitige Wahl der Ziele „Überblick“ und „Details“) geeignet aufzulösen.

Die Einbeziehung der Analysehistorie ist ein vielversprechender Ansatz, den Anwendern Rückkopplung über durchgeführte Analyseschritte zu geben und erfolgreiche Analysen zu verwalten (vgl. hierzu auch Kreuseler u. a. (2004)). In ersten Tests wurde deswegen untersucht, inwiefern sich die in einem Visualisierungsdesign-Wizard getroffe-

[10] Start der Visualisierung direkt aus dem Simulationssystem bzw. separat mit einer gespeicherten Datei
[11] z.B. die Findung einer geeigneten Reihenfolge der Achsen in einer Parallelen Koordinaten-Darstellung

nen Entscheidungen verwalten, dem Nutzer intuitiv zugänglich und wiederverwenden lassen. Abbildung 7.11 zeigt zwei Darstellungen der im Rahmen einer Analysesitzung mit dem SimEnvVis-Wizard aufgezeichneten Historie. Jede Spalte repräsentiert hier eine Stufe des Wizards wobei die einzelnen Knoten mit Vorschaubildern oder kleinen Ikonen Hinweise über die Art der Entscheidung geben, und die Rahmenfarbe ggf. gestartete Visualisierungen repräsentiert.

7.4 Zusammenfassung

In diesem Kapitel wurde ein neuer Ansatz zum Visualisierungsdesign vorgestellt. Als Basis hierfür wurden neue Konzepte zur Beschreibung der essentiellen Einflussfaktoren Metadaten und Analyseziele eingeführt. Neuland beschreitet diese Arbeit bei der Beschreibung von grundlegenden Metadaten für das visuelle Data Mining. Weiterhin demonstriert die Arbeit innovative Methoden zu deren (halb)automatischer Erhebung. Auch bei der Beschreibung von Zielen werden verschiedene in der Literatur isoliert betrachtete Aufgaben und Zielstellungen integriert. Für diese beiden Einflussfaktoren wurde untersucht, wie sich anwendungsunabhängige und anwendungsspezifische Begriffe in Beziehung setzen und ineinander überführen lassen. Basierend auf diesen Untersuchungen wurde ein pragmatisch orientierter Mechanismus zum Visualisierungsdesign konzipiert und umgesetzt, welcher getrieben von der speziellen Anwendung in der Klimafolgenforschung anwendungsspezifische und anwendungsunabhängige Regeln zur Generierung von Darstellungen miteinander kombiniert, und dabei auch unscharfes und zum Teil widersprüchliches Wissen einbezieht. Neu an diesem Mechanismus ist, dass er verschiedene Entscheidungen im Mapping- und Renderingprozess von der Auswahl von Visualisierungsmetaphern über die Wahl des Variablenmapping bis zu einzelnen Parametrisierungen einbezieht.

Die Grenze dieses Ansatzes besteht darin, dass er sich auf die Auswahl und Parametrisierung einzelner Visualisierungstechniken konzentriert. Für viele praktische Analyseaufgaben ist jedoch eine Verknüpfung mit anderen Methoden erforderlich. Dies beschränkt sich nicht ausschließlich auf die Kopplung mehrerer Visualisierungstechniken (vgl. erste Ansätze hierzu in der betreuten Arbeit von Wagenknecht (2006)), sondern bezieht auch die Vernähung mit automatischen Analyseverfahren mit ein. Abbildung A.14 im Anhang illustriert einen ersten Ansatz, wie (verfahrens-)spezifische Zielstellungen für Clusterfahren sich mit allgemeinen Zielen verknüpfen lassen, und zeigt welche Klassen von Clusterverfahren diesen Zielstellungen zugeordnet werden können.

Darüber hinaus bleiben eine Vielzahl von Herausforderungen für zukünftige Arbeiten. Dies schließt insbesondere eine vertiefte Untersuchung der Verknüpfung von Konzepten und Methoden der Historienverwaltung und des Visualisierungsdesign, die Integration der für die verschiedenen Datenquellen konzipierten Metadatenerhebungsmechanismen in ein einheitliches Werkzeug, sowie die Erweiterung des Einsatzes von Zielstellungen auf allgemeine Aufgabenmodelle im Analyseprozess ein.

Kapitel 8

Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde eine breite Palette von Vorgehensweisen und Techniken aus den Gebieten der Visualisierung und des visuellen Data Mining auf deren Anwendbarkeit hin systematisch untersucht und weiterentwickelt. Dabei lag der Fokus darauf, die aus dem visuellen Data Mining bekannten Methoden für räumliche und/oder zeitliche multi-variate Daten an die Spezifik eines Anwendungsgebietes anzupassen, und dabei die Vielfalt der verfügbaren Techniken für die Anwender leicht zugänglich zu machen. Dies erforderte, grundlegende, anwendungsübergreifende Methodiken zu entwerfen, die es erlauben, sich aufgrund ihres allgemeinen Vorgehens leicht auch auf eine spezielle Anwendung zuschneiden zu lassen.

Als Basis für diese Untersuchungen wurde der aktuelle Forschungsstand in einer Vielzahl von relevanten Teilgebieten zusammengetragen. Dies betrifft neben den Grundlagen der Visualisierung, des visuellen Data Minings und des Visualisierungsdesigns insbesondere Publikationen zu den Gebieten Visualisierung von Wetter- und Klimadaten, Visualisierungs- und visuelle Data Mining-Systeme, vergleichende Visualisierung sowie Metadaten und Analyseziele für das visuelle Data Mining.

Am Beispiel der Klimaforschung demonstriert diese Arbeit, wie eine Kombination aus interaktiven Visualisierungstechniken mit automatischen Analyseverfahren neue Einsichten in die Daten ermöglichen. Im Rahmen einer Kooperation mit dem Potsdam Institut für Klimafolgenforschung wurden eine Vielzahl von Techniken auf ihre Potenz, praktische Fragestellungen zu beantworten, untersucht, und in einer Bibliothek zusammengefasst. Diese Untersuchungen wurden von einer Vielzahl von studentischen Arbeiten begleitet, welche zur Entwicklung verschiedener Softwaretools geführt haben.

Zur Integration der entworfenen Tools wurde im Laufe der Arbeit eine Architektur einer Komponentenbibliothek vorgestellt, welche es erlaubt, wichtige für das visuelle Data Mining identifizierte Bausteine flexibel zu kombinieren und in praktische Systeme einzubeziehen. Insbesondere sind daraus das Framework Metadatum zur flexiblen Erhebung von Metadaten für das visuelle Data Mining aus alpha-numerischen, tabelarischen Textdateien (vgl. Nocke 2000; Nocke u. Schumann 2002), das Framework SimEnvVis zur nutzerunterstützten Visualisierung von Simulationsexperiment-Ausgabedaten unter Einbeziehung eines Visualisierungsdesign-Wizards (vgl. Lange u. a. 2006; Nocke u. a. 2007) sowie der Prototyp VisAna zur interaktiv gesteuerten Kopplung von visuellen Data Mining-Techniken zur Unterstützung des Modellierungs- und Simulationsprozesses (vgl. Nocke u. a. 2003) entstanden. Das in Kooperation mit dem Potsdam Institut für Klimafolgenforschung entwickelte SimEnvVis-Framework ist hierbei besonders hervorzuheben. Es ist ein flexibel einsetzbares Werkzeug zur Visualisierung von Simulationsdaten im Umfeld der Klima- und Klimafolgenforschung. Es ist mittlerweile fest in die Simulationsumgebung SimEnv (vgl. Flechsig u. a. 2006, 2007) integriert und wurde inzwischen für die Auswertung einer Vielzahl von Simulationsexperimenten eingesetzt.

8.1 Innovativer Beitrag

Bei der Lösung der vielfältigen Problemstellungen bei der Analyse von Klimadaten wurden neben der Anpassung von Standardtechniken auch neuartige Methoden entworfen und Methodiken untersucht. Im folgenden seien die wichtigsten Ergebnisse der Arbeit noch einmal zusammengefasst.

Entwurf neuartiger, intuitiver Visualisierungstechniken. Vor dem Hintergrund der Bereitstellung von leicht verständlichen Darstellungen von multi-variaten Daten in Raum und/oder Zeit, wobei vor allem die Suche und Quantifizierung von Mustern (z.B. das Auffinden von Extremen und von Periodizitäten) in den Daten im Vordergrund steht, wurden neuartige Darstellungsmethoden entworfen und umgesetzt. Insbesondere beschreitet diese Arbeit mit der metapherbasierten Darstellung von Ikonen sowie deren Anordnungen basierend auf Mosaikbildern für multi-variate, gestreute 2D-Klimadaten Neuland. Ferner wurden bekannte Darstellungen erweitert, um neue Einsichten in zeitliche Klimadaten zu erhalten (Differenzmethode, Rechteckmethode).

Entwurf einer allgemeinen Methodik zur vergleichenden Visualisierung. Die vorliegende Arbeit untersucht die grundlegenden Möglichkeiten zum Vergleich von Daten (mit dem Fokus auf dem räumlichen Bezug) und stellt hierfür eine allgemeine Methodik bereit. Insbesondere werden dabei Vorgehensweisen zum Vergleich von Daten auf abweichenden Gittern (wie sie z.B. bei der Evaluation von Klimamodellen vorkommen) in unterschiedlichen Bezugsräumen einbezogen, ein Aspekt, der in der gängigen Literatur eher vernachlässigt wird. Die Arbeit beschreibt neue Vorgehensweisen, wie die Probleme, die beim üblicherweise verwendeten Nebeneinanderlegen mehrerer, zumeist ungekoppelter Darstellungen („image level“) vermieden werden können. Eine erste Umsetzung von Verfahren dieses Ansatzes illustrieren die Potentiale bei der vergleichenden Visualisierung für die Modellevaluation.

Kombinatierte visuelle und automatische Analysetechniken. Weiterhin stellt diese Arbeit neue Ansätze zur Darstellung von Ergebnissen aus Cluster- und Hauptkomponentenanalyse vor. Im besonderen wird die Analyse der Eigenschaften von geclusterten Daten in Raum und Zeit sowie der Vergleich von Clusterstrukturen durch die Vielzahl von alternativen Darstellungen - unter anderem durch eine angepasste Farbwahl von Clustern, der multi-variaten Darstellung der Clustereigenschaften im räumlichen Bezug oder dem Vergleich von geclusterten Zeitreihen - deutlich verbessert. Ferner wurde ein neuer systematischer Ansatz zur visuellen Analyse von Trends durch die enge Kopplung von Visualisierungsmethoden mit der Hauptkomponentenanalyse unter Einbeziehung aller Schritte der Visualisierungspipeline vorgestellt. Durch die vorgestellten Techniken wird es Anwendern wesentlich erleichtert, die Ergebnisse aus Cluster- und Hauptkomponentenanalysen zu interpretieren und zu kommunizieren.

Entwurf eines allgemeinen Vorgehensmodells zur Kopplung von Modellierungsfunktionalität mit visuellen Data Mining-Techniken. In dieser Arbeit wurde systematisch untersucht, wie der gesamte Prozess der Modellbildung, -simulation und -evaluation durch VDM-Verfahren unterstützt werden kann. Es wurden neuartige Methoden zur interaktiv gesteuerten Modellspezifikation sowie bei der Modellanalyse vorgestellt und deren Einsatzmöglichkeiten am Beispiel der Reduktion und Evaluation von Klimamodellen (mit dem Schwerpunkt auf gewöhnlichen Differentialgleichungen) demonstriert.

Spezifikation und Erhebung von Metadaten für das visuelle Data Mining. Bisherige Arbeiten zur Spezifikation von Metadaten für Visualisierungszwecke konzentrieren sich auf bestimmte Aspekte bei der Spezifikation und beim Einsatz von Metadaten. Die hier vorliegende Arbeit geht deutlich über diese Ansätze hinaus und führt eine allgemeine Metadatenspezifikation unter Einbeziehung von Metadaten für verschiedene Datenklassen und für die Qualität der Daten, die sich leicht an spezielle Anwendungskontexte anpassen lässt, ein. Neuland beschreitet diese Arbeit beim Entwurf und bei der Umsetzung von Methoden, welchen den Anwender konsequent dabei unterstützen,

für das visuelle Data Mining erforderliche und/oder potentiell nützliche Metadaten zu erheben und diese (visuell) auszuwerten.

Spezifikation von Zielen und Aufgaben. Diese Arbeit systematisiert wichtige Ansätze zu Aufgaben und Zielen und leitet eine praktikable, an verschiedene Anwendungen anpassbare Spezifikation ab. Neu an diesem Ansatz ist das dreistufiges Vorgehen, welches elementare Zielstellungen und komplexe Zielstellungen einbezieht, und die Einführung einer Beschriftungsfunktion, welche einen ausgewogenen Kompromiss zwischen der praktischen Handhabung von Zielen und der Einbeziehung komplexer Zielstellungen ermöglicht.

Einführung eines neuartigen Mechanismus zum Visualisierungsdesign. Im Rahmen der Arbeit wurde ein zweistufiger Mechanismus zum Visualisierungsdesign entworfen und umgesetzt, der sowohl die grundlegende Eignung von Visualisierungen basierend auf essentiellen Metadaten als auch weitere Entscheidungen zur Parametrisierung dieser Visualisierungen basierend auf Regeln für beliebige Eignungsgrößen einschließt. Neu an diesem Ansatz ist die umfassende Sicht auf alle Ebenen der Entscheidungsfindung, angefangen von der Wahl von Visualisierungstemplates über die Modulauswahl und das Variablenmapping bis hin zu einzelnen Parametrisierungen. Dabei ist der Mechanismus um beliebige Regeln erweiterbar.

8.2 Offene Probleme

Über die in der Arbeit entworfenen und umgesetzten Konzepte und Techniken hinaus verbleiben eine Vielzahl von Herausforderungen für weiterführende Forschungsarbeiten. Dies schließt zu allererst die Erweiterung der Visualisierungstechnikbibliothek ein, um die Spannbreite von Problemstellungen im Umfeld der Modellierung und Simulation sowie der Klimaforschung noch besser abzudecken. Ferner besteht die Aufgabe, die Akzeptanz dieser Techniken in der Anwendung weiter zu erhöhen (z.B. durch neue Erkenntnisse), und damit die Chancen des interaktiven visuellen Data Mining verstärkt zu kommunizieren. Insbesondere die Ausweitung des Fokus auch auf Modelle und Daten zur Abschätzung von Klimafolgen sowie auf die Auswertung von komplexen Simulationexperimenten erforderten hierbei zum Teil neue Lösungen (vgl. hierzu auch Nocke u. a. 2007).

In dieser Arbeit wurde die flexible Kopplung von automatischen Verfahren mit Visualisierungstechniken in einem leicht bedienbaren Prototypen VisAna nur in Ansätzen betrachtet. Bisher liegen die automatischen Verfahren hier als isolierte Bausteine vor. Eine systematische Integration aller im Rahmen der Arbeit umgesetzten Techniken in einem allgemeinen Framework zur Unterstützung des Modellierungs- und Simulationprozesses steht jedoch noch aus. Neben dem Problem der Kopplung der verschiedenen dabei beteiligten Komponenten stellen insbesondere die bei Simulations- und automatischen Analyseprozessen auftretenden Zeit- und Resourcenanforderungen eine Herausforderung für ein interaktives System dar.

Weiterhin müssen die Ansätze zur vergleichenden Visualisierung ausgebaut und die Vielzahl untersuchter Ansätze systematisch in Softwarelösungen überführt werden. Dies schließt ein, die vorgeschlagenen Mappingstrategien von Merkmalen und Gittern auf beliebige visuelle Attribute sowie auf verschiedene Renderingstile weiter zu untersuchen und umzusetzen.

Auch bei der Kommunikation von Ergebnissen statistischer Verfahren wurden die dabei auftretenden Probleme nur in Ansätzen gelöst. So bleibt es auch weiterhin eine Herausforderung, die Ergebnisse von Clusterfahren und den aus ihnen resultierenden Objektstrukturen und Clustereigenschaften leicht verständlich - gerade auch im räumlichen und zeitlichen Bezug - darzustellen. Auch der visuelle Vergleich von Resultaten aus mehreren Clusteranalysen und deren multi-variaten Eigenschaften ist ein noch weitgehend offenes Problem.

Im Rahmen dieser Arbeit wurde auf Metadaten für die Visualisierung und das visuelle Data Mining

fokussiert. Ansatzpunkt für zukünftige Arbeiten ist zu untersuchen, wie solche Metadaten und die zugehörigen Methoden zu deren Erhebung und Verwaltung mit Metadaten und Methoden aus anderen Gebieten (z.B. Metadaten im MPEG7-Standard oder in Datenbanken) interagieren können.

Abschließend verbleibt die Aufgabe, den Mechanismus zum Visualisierungsdesign weiter zu testen und auszubauen. Dies schließt ein, ihn um weitere Entscheidungen und neue Regeln, zum Beispiel bei der Feinparametrisierung, zu erweitern und mit Hilfe von Nutzerstudien zu validieren.

8.3 Schlussbemerkungen

Noch immer ist die Benutzung von Visualisierungs- und visuellen Data Mining-Techniken und -systemen in praktischen Anwendungen vielen Begrenzungen und Einschränkungen unterworfen. Die heterogene Landschaft von verfügbaren Systemen mit verschiedenartigen Schnittstellen, die oft für die Anwender einen hohen Einarbeitungsaufwand erfordern, erschweren die Findung neuer Erkenntnisse über die Daten. Die Anbindung dieser Daten und die Findung/Generierung einer interaktiven Darstellung, insbesondere im Unterschied zum Datenbankumfeld, wo sich einheitliche Standards für Datenhaltung und Datenschnittstellen etabliert haben, ist im Visualisierungsumfeld aufgrund der heterogenen Datenquellen verschiedener Anwendungen wesentlich erschwert. So stehen im Umfeld der Klima- und Klimafolgenforschung einer Vielzahl von eingesetzten Systemen zur visuellen Datenanalyse (von low-level skiptbasierten Visualisierungssprachen wie R, über Visualisierungssysteme wie AVS und OpenDX, geographischen Informationssystemen wie ArcGis bis zu allgemeinen Systemen wie Excel und Matlab) eine Vielzahl von Datenformaten (NetCDF, IEEE-kodierte Binärdaten, GRIB, Datenbanken, u.a.) gegenüber.

Zum Schluss dieser Arbeit sollen zwei Visionen formuliert werden, welche darauf abzielen, die Akzeptanz des visuellen Data Mining in diesem Spannungsfeld wesentlich zu erhöhen:

1. **Standardisierte Schnittstellen**: Eine Vision, in deren Richtung diese Arbeit bei der Spezifikation von Metadaten und Visualisierungsdeskriptoren wichtige Schritte geht, ist es, eine Standardisierung von Visualisierungsfunktionalität durchzuführen, welche durch Beschreibungen von Schnittstellen von Visualisierungstechniken für Anwender und Entwickler einen einheitlichen Zugriff zu unterschiedlichen Visualisierungssystemen und damit eine wesentlich erleichterte Benutzbarkeit und Erweiterbarkeit ermöglicht. Dies würde den Einsatz von interaktiven Visualisierungen sowie deren Kopplung untereinander und mit automatischen Methoden in der Praxis stark forcieren.
2. **Nutzerzentrierung**: Eine weitere Vision besteht darin, die in dieser Arbeit vorgeschlagenen Ansätze konsequent weiterzuentwickeln, um ein leicht bedienbares visuelles Data Mining - Framework mit einem hohen Grad an Anwenderunterstützung zu entwerfen und umzusetzen, welches eine direkte Ankopplung an verschiedene Modelle und Experimentierumgebungen ermöglicht. Hierbei sollen sowohl Aspekte der Datenqualität als auch die Unterstützung von explorativen und konfirmativen Analyseaufgaben durchgehend unterstützt werden. Durch den Einsatz von visuellen Data Mining - Techniken in allen Schritten des Modellierungsprozesses kann ein solches Framework über die isolierten Fähigkeiten aktueller Simulations- und Visualisierungssysteme deutlich hinausgehen und neue Wege zur Modellierung und Simulation aufschließen.

Zur Lösung der damit verbundenen Problemstellungen eröffnet sich ein weites Feld von Forschungsaktivitäten.

Anhang A

Weitere Abbildungen

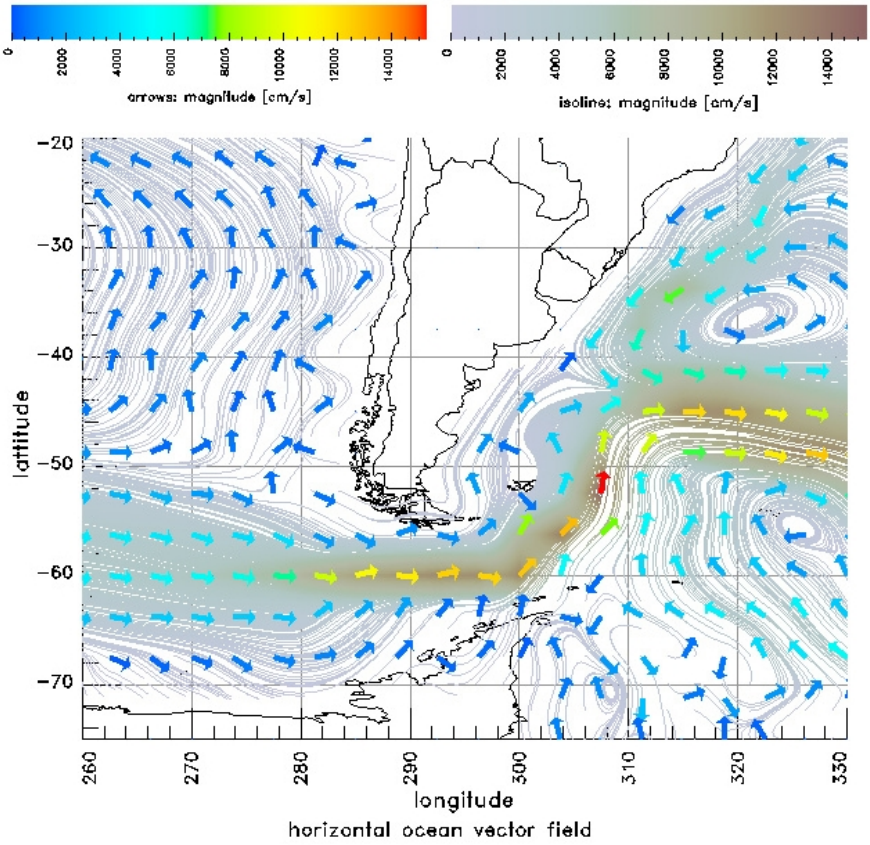


Abbildung A.1: Stromlinien- und Pfeildarstellung des horizontalen Ozeangeschwindigkeitsfeldes generiert mit dem CLIMBER-3-Modell

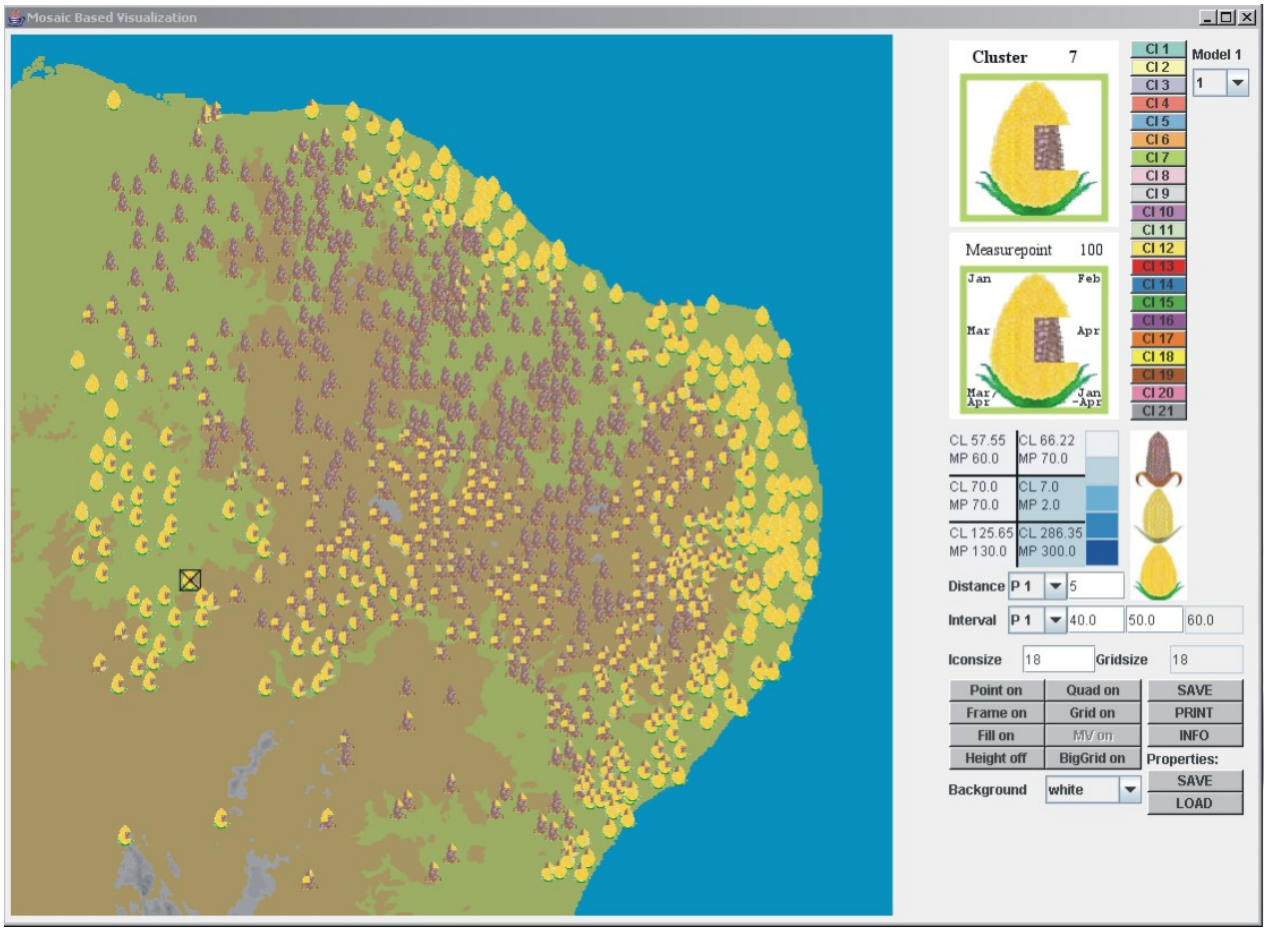


Abbildung A.2: Tool zur metapherbasierten Ikonendarstellung für skalare, gestreute 2D-Klimadaten (entwickelt von Baalcke (2005))

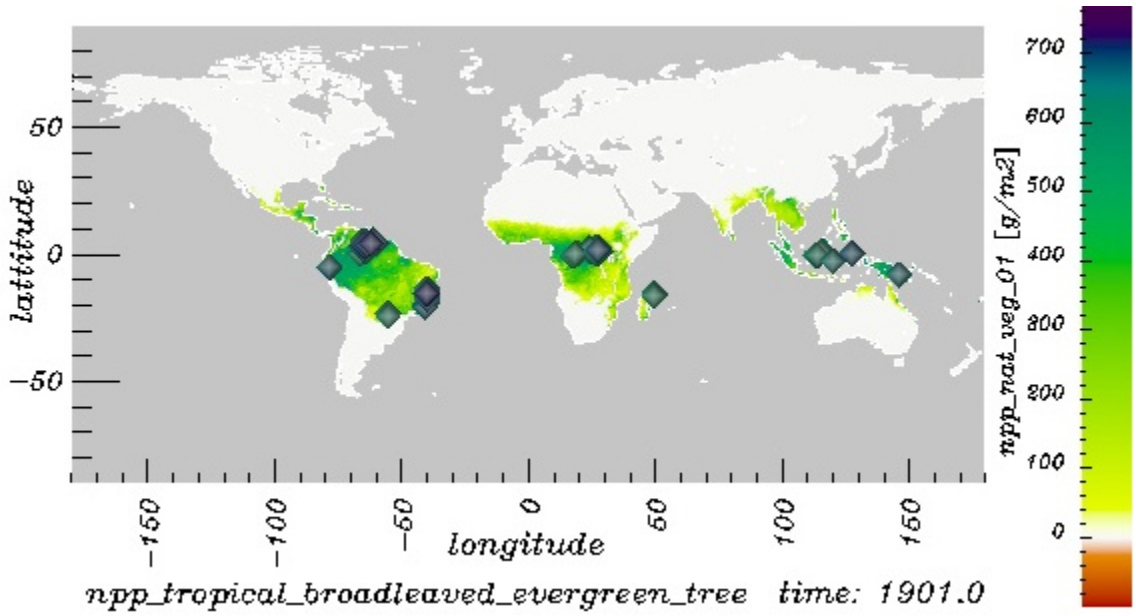


Abbildung A.3: 2D-Farbabbildung der Nettoprimärproduktion von tropischen Hartlaubgewächsen, simuliert mit dem Vegetationsmodell LPJ, Hervorhebung von Extremwerten zwischen 570-770 g/m^2 durch Darstellung von Glyphen.

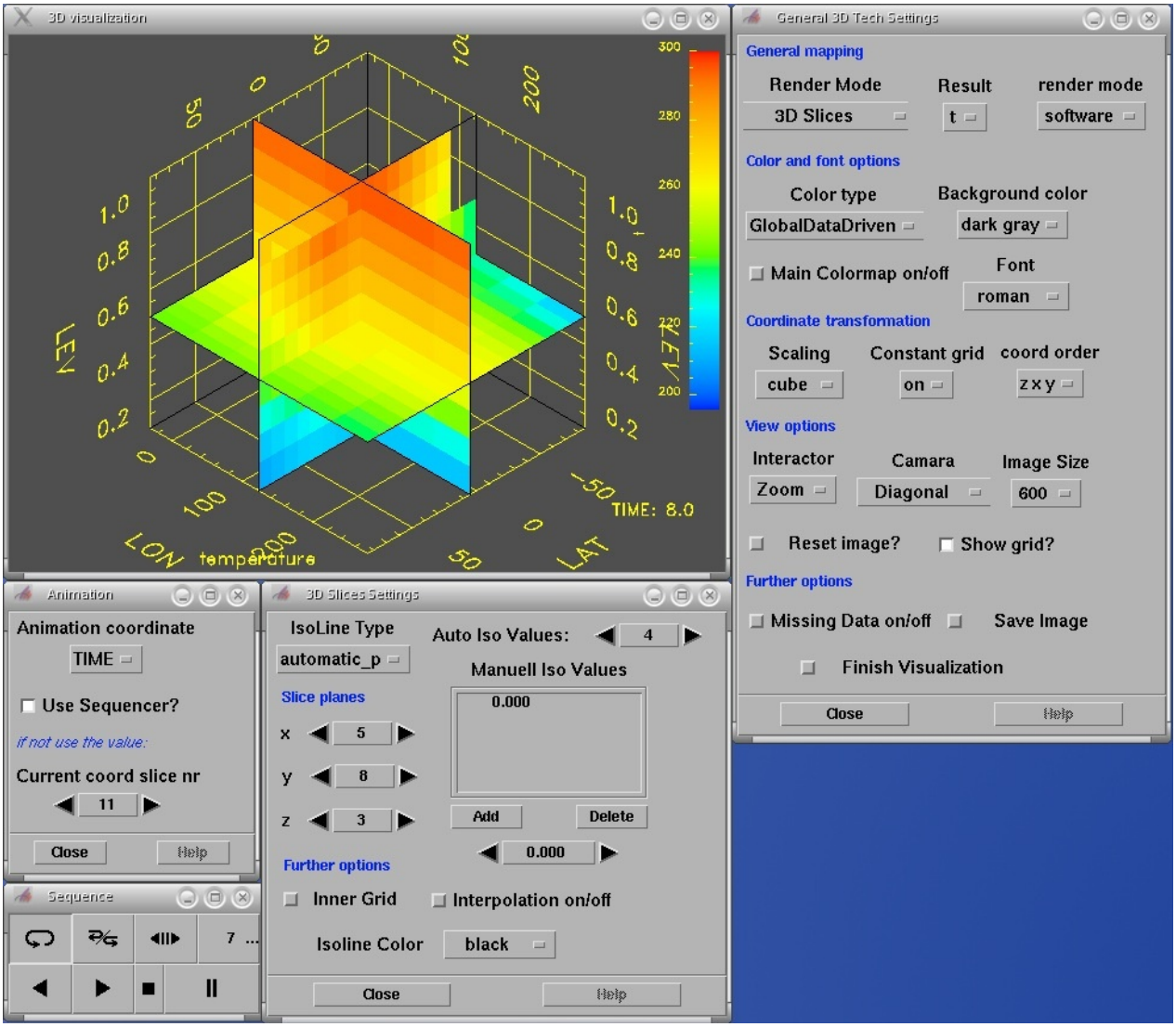


Abbildung A.4: Screenshot des Moduls zur Visualisierung von skalaren Klimadaten auf regulären 3D-Gittern)

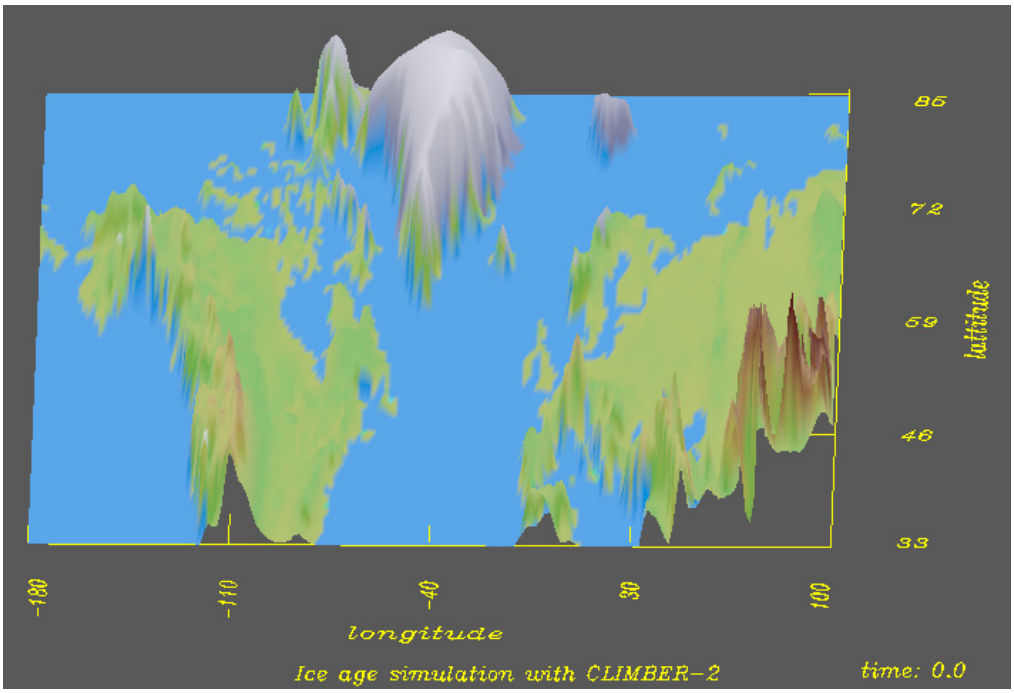


Abbildung A.5: Höhenfelddarstellung einer Eiszeitsimulation mit dem CLIMBER-3-Modell (Ausschnitt aus einer Animation)

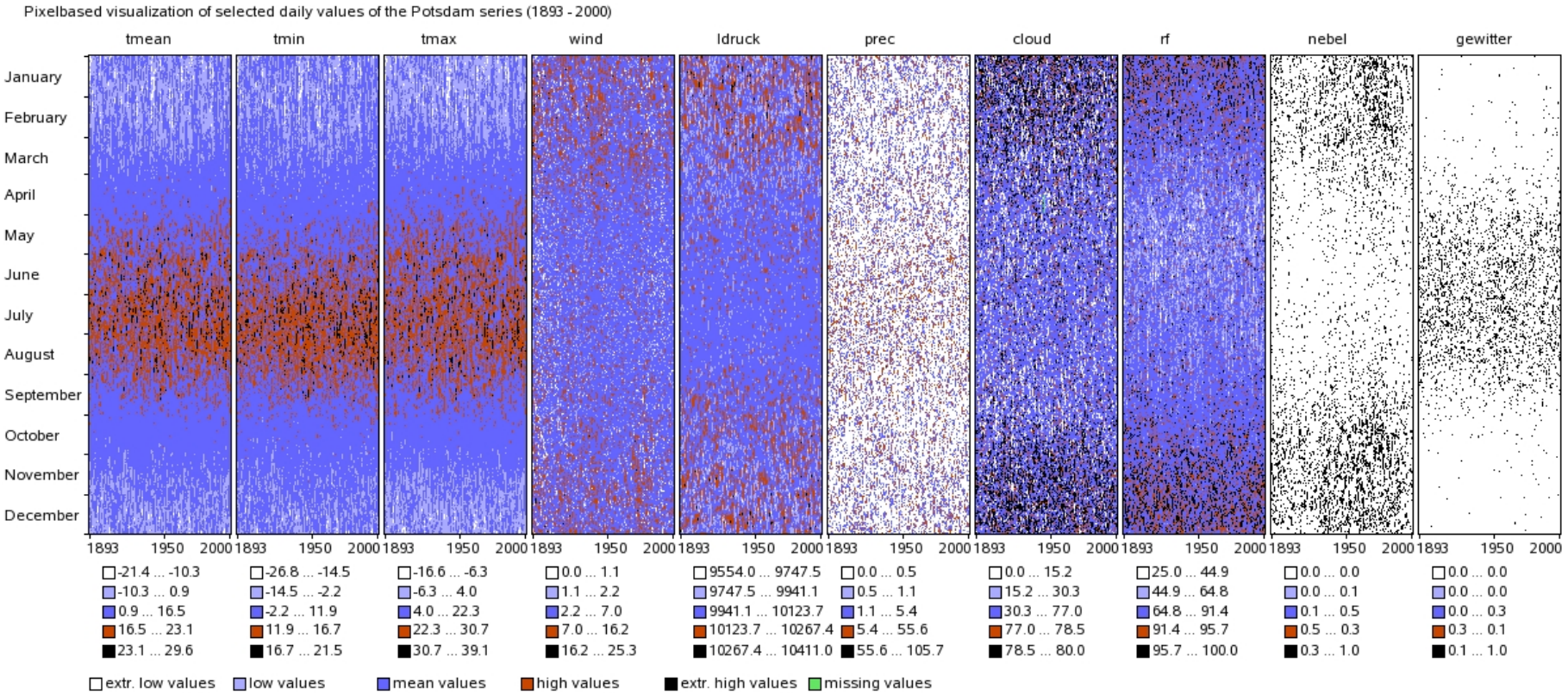


Abbildung A.6: Pixelbasierte Darstellung der Potsdamer Reihe; Darstellung von zehn Merkmalen für jeden Tag von 1893 bis 2000 in einer Weiß-Blau-Orange-Schwarz-Skala

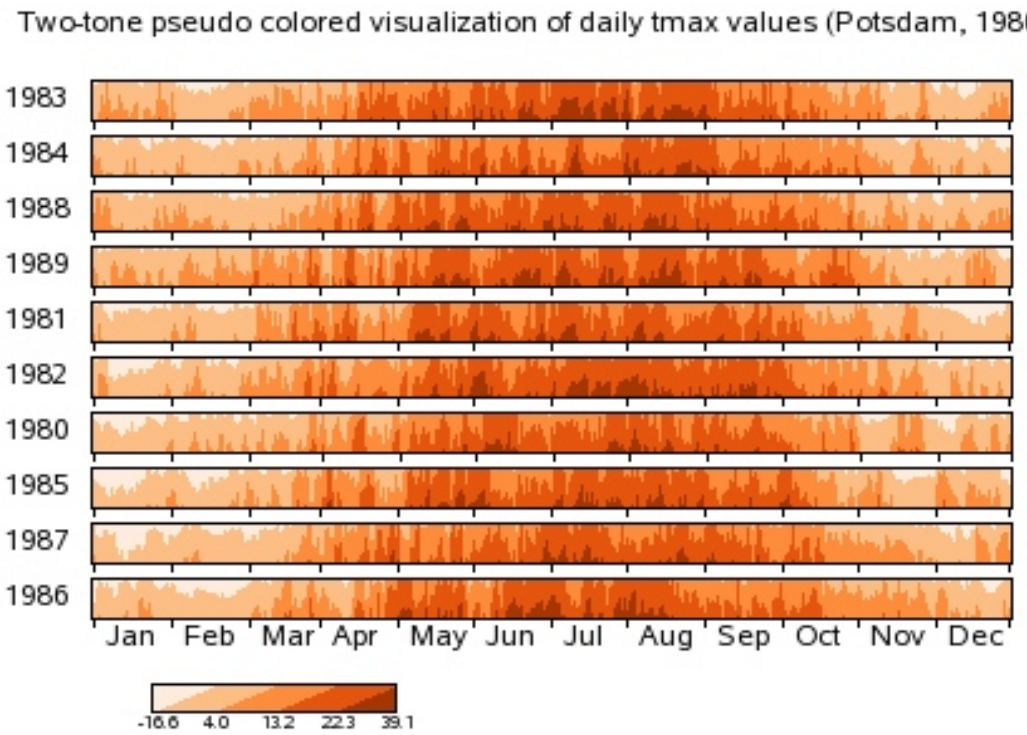


Abbildung A.7: „Two-tone“-Farbabbildung nach Saito u. a. (2005); Darstellung der nach Jahresmittelwert sortierten Jahre 1980-1989; Potsdamer Reihe; Tages-Maximalwerte

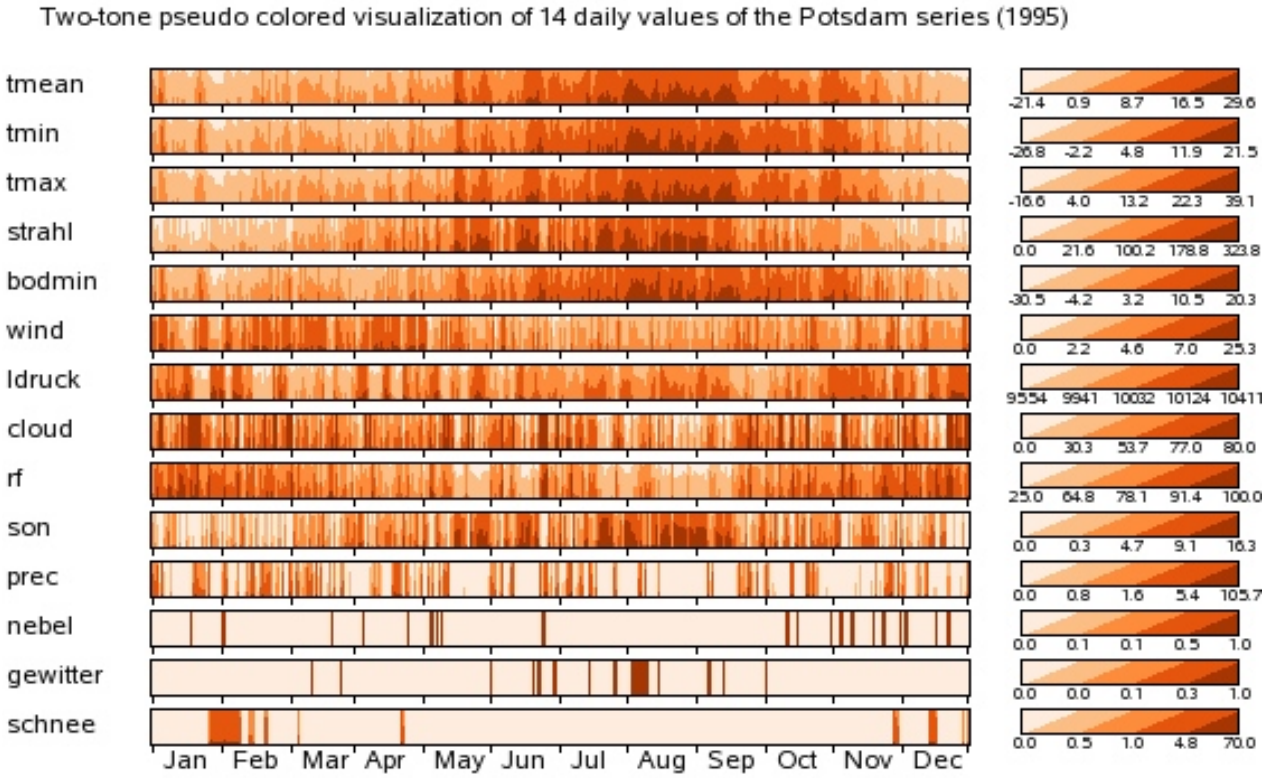


Abbildung A.8: „Two-tone“-Farbabbildung nach Saito u. a. (2005); Darstellung von 14 Merkmalen für das Jahr 1995; Potsdamer Reihe

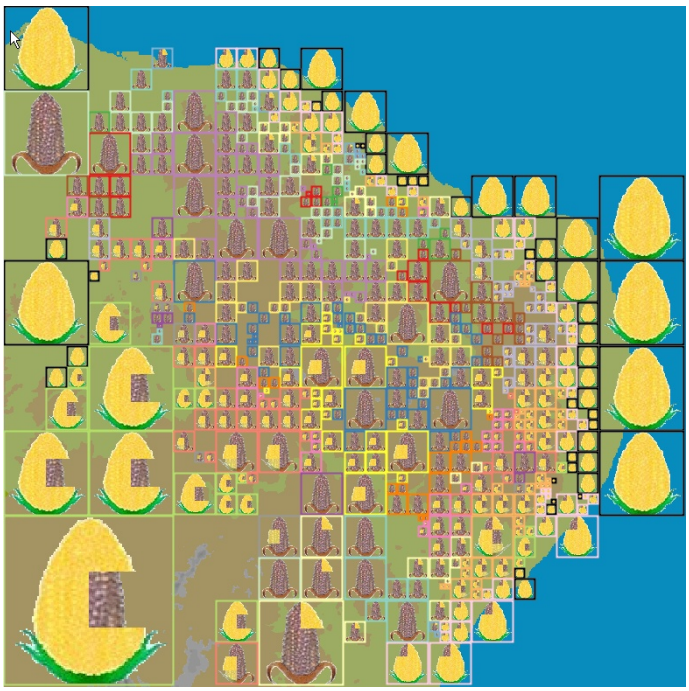


Abbildung A.9: Metapherbasierte Ikonendarstellungen für gestreute, geclusterte 2D-Klimadaten mit zentralen Punkten repräsentiert durch Maisikone und Clusterzugehörigkeit repräsentiert durch farbige Ikonenberandung; Multi-resolution Layout; Hervorhebung eines Clusters durch schwarze Umrandung

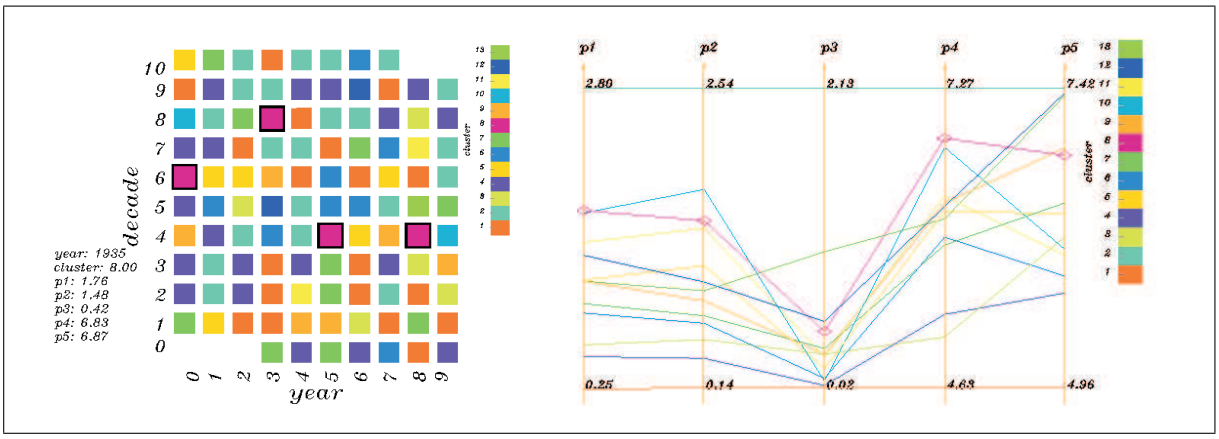


Abbildung A.10: Linking & Brushing von Clustern und Clustereigenschaften (Potsdam Sommerdatensatz); zeitliche Darstellung der Clusterzugehörigkeiten (links) und Parallele Koordinaten-Darstellung der zentralen Punkte (rechts); Cluster 8 (Jahre 1935, 1938, 1950 und 1973) und der zugehörige Linienzug wurden hervorgehoben.

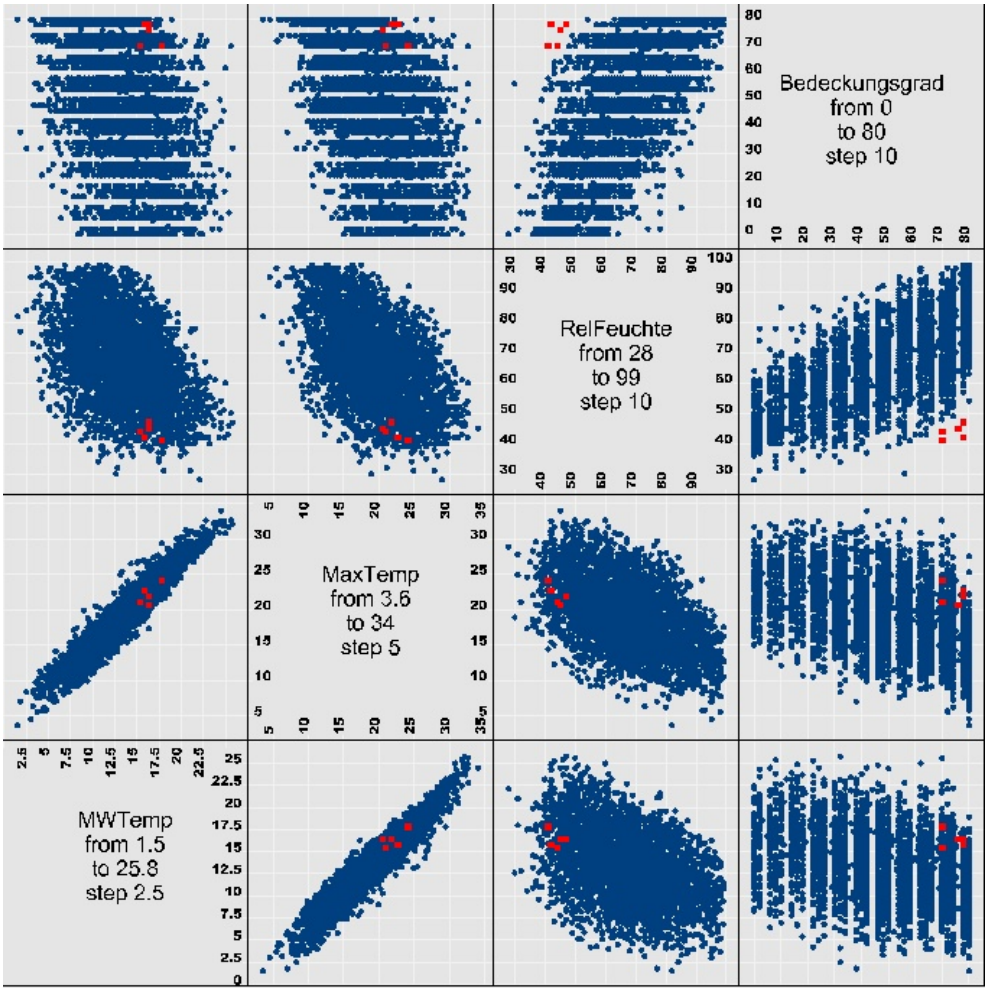


Abbildung A.11: Scatterplot-Matrix (mit dem System InfoVis3D erzeugt) des Datensatzes der Tageswerte aller Maitage von 1893-2003 der Station Potsdam; 4 Merkmale mit Beitrag zur zweiten Hauptkomponente PC2 ausgewählt; Auswahl von fünf Tagen, die bzgl. der *relativen Luftfeuchtigkeit* und dem *Bedeckungsgrad* Abweichungen zum Trend aufweisen

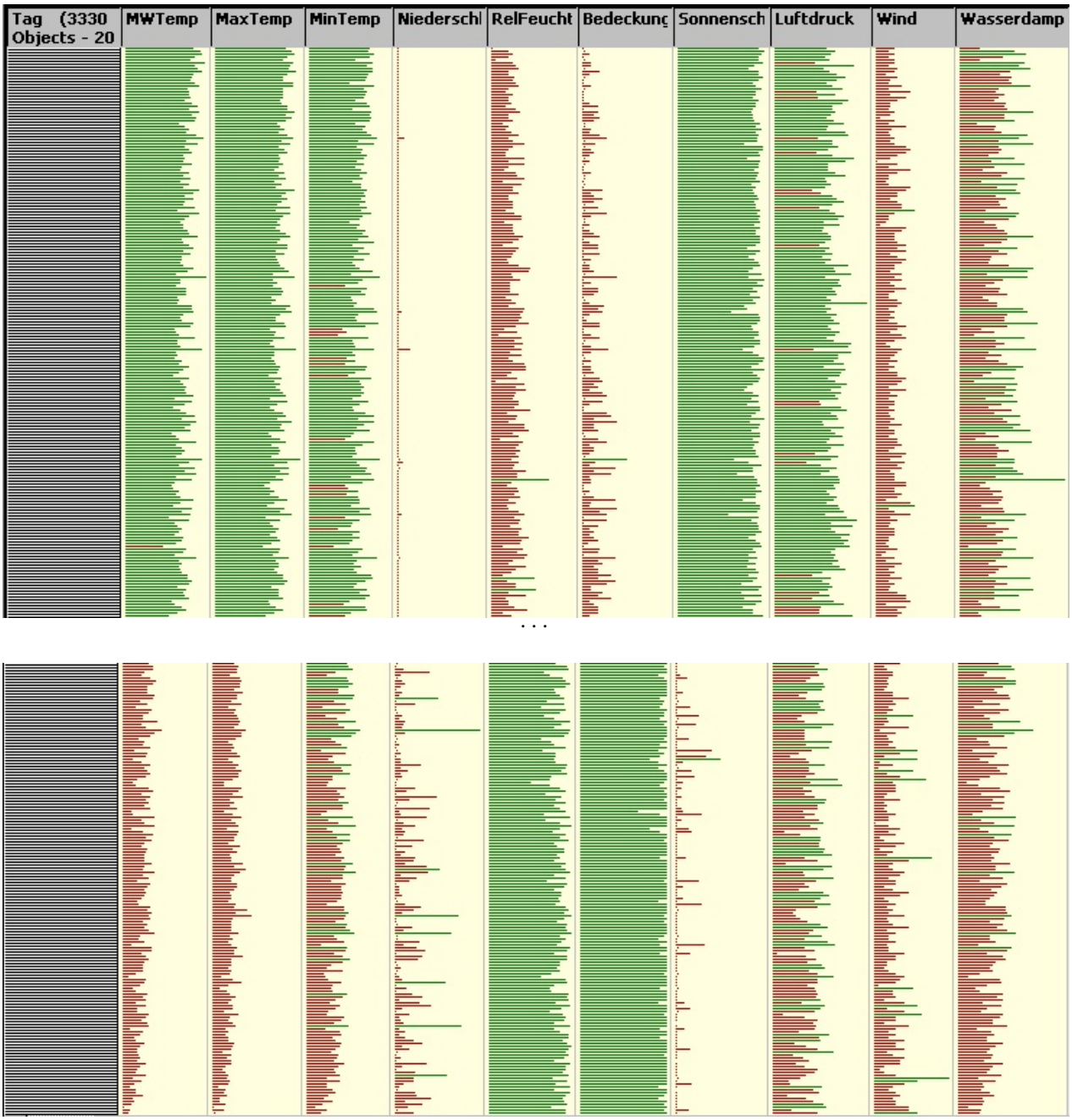


Abbildung A.12: Tabellendarstellung (System InfoVis3D) der Originaldaten des Maitage-Datensatzes sortiert nach der Hauptkomponente PC2

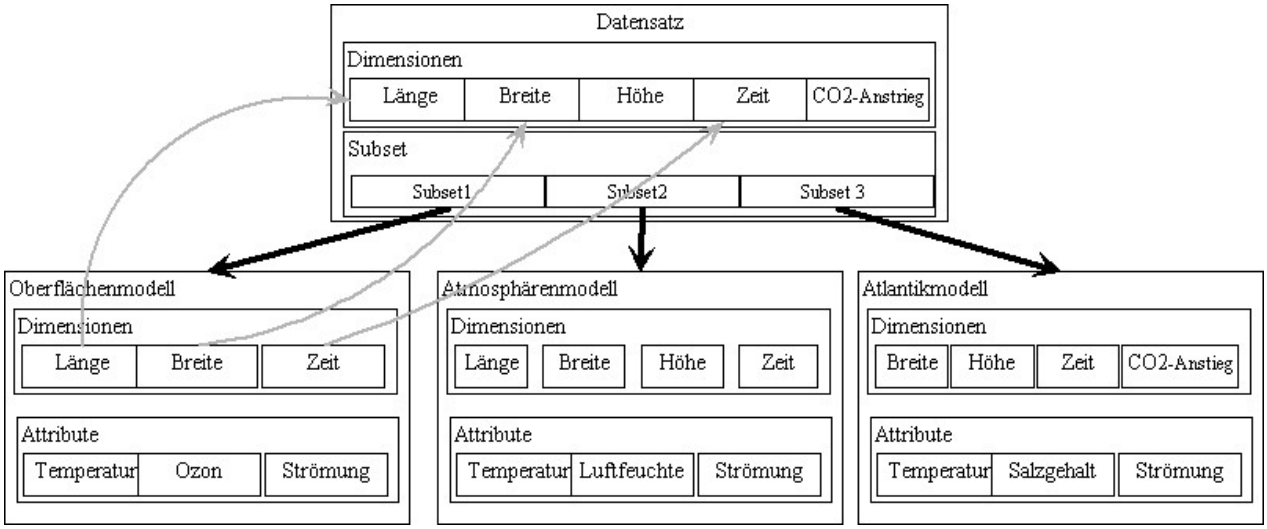
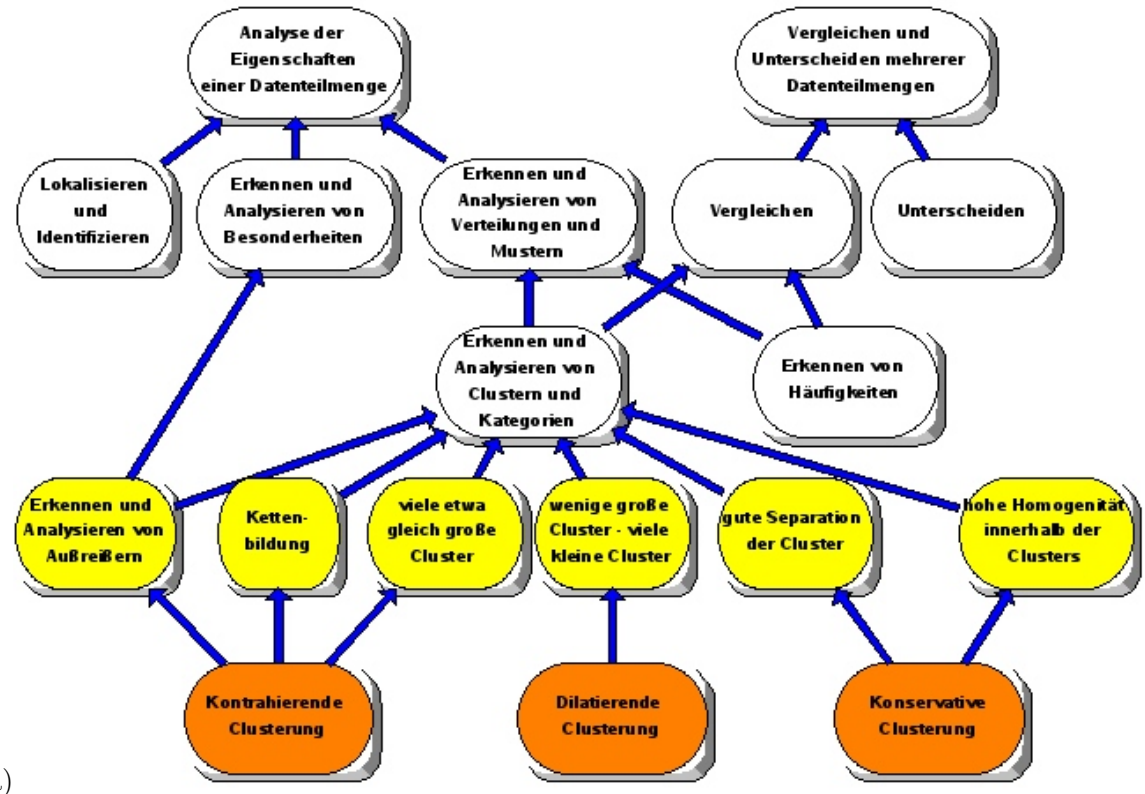


Abbildung A.13: Illustration von Teilmengen in einem gekoppelten Klimamodell



a)

Abbildung A.14: Erster Ansatz zur Spezifikation von Zielstellungen und deren Abhängigkeiten zur Unterstützung bei der Auswahl von Clusterverfahren; weiß: allgemeine Zielstellungen; orange: spezielle Zielstellungen bei der Auswahl eines Clusterfahrens; orange: zugehörige Clusterverfahren

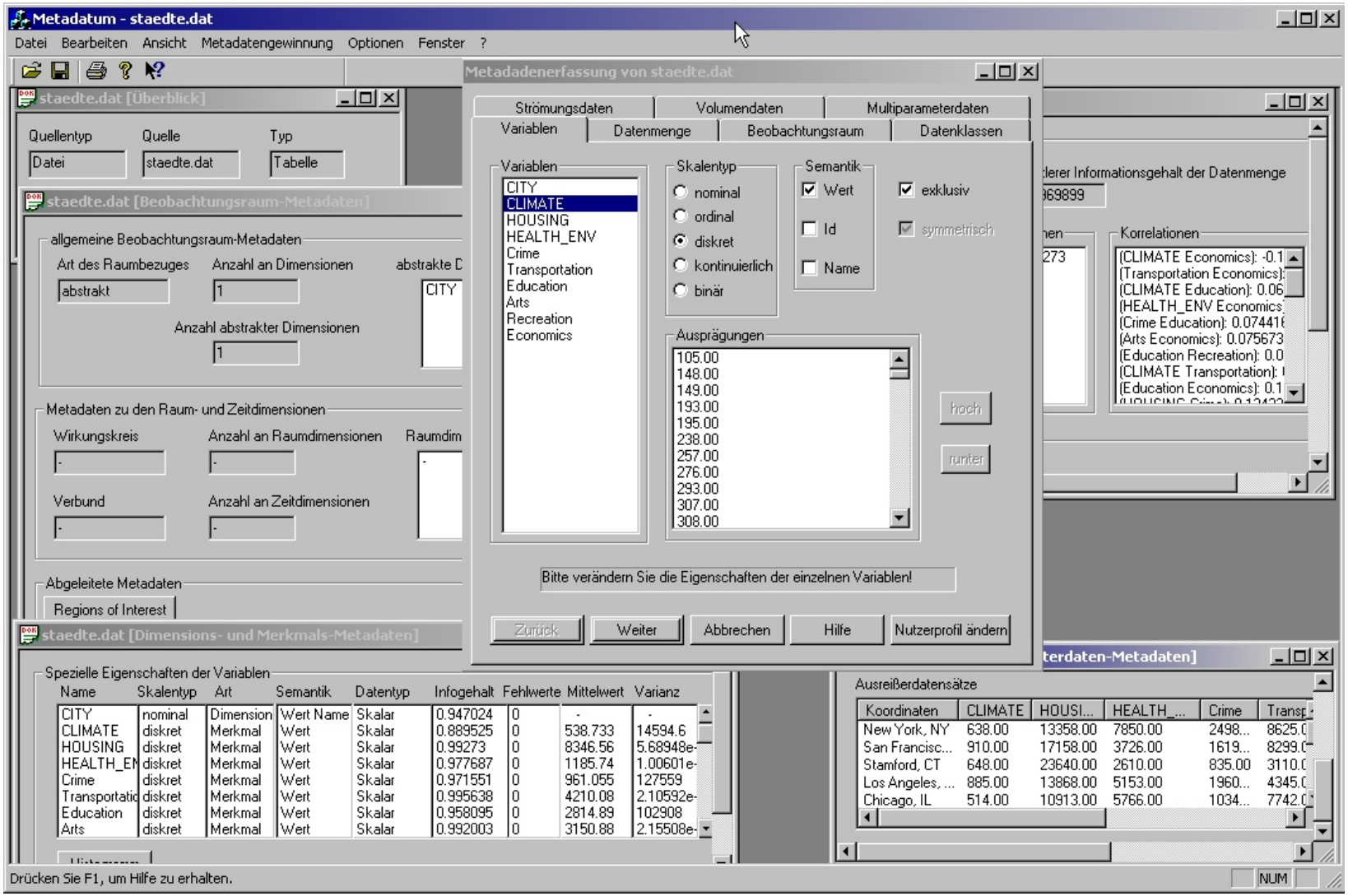


Abbildung A.15: Erhebung und Darstellung von Metadaten im Framework Metadatum

Anhang B

Visualisierungsdesign - die Details

B.1 Allgemeine Spezifikation von Metadaten

Für die in Abschnitt 7.1.2 identifizierten Klassen an Metadaten lassen sich eine Vielzahl von Metadaten untersetzen. Die bei der Metadatenspezifikation betrachteten Datencharakteristika seien im folgenden im Detail aufgelistet (vgl. auch Nocke u. Schumann 2002):

Metadaten für die gesamte Datenmenge

Allgemeine Metadaten für die gesamte Datenmenge

- die Anzahl der Variablen $^{\circ}$
- die Anzahl der Beobachtungsfälle$^{\circ}$(engl.: data records)
- Strukturinformationen über die Variablen* insbesondere
  - Tupel- und Schlüsselinformationen,
  - Korrelationen und gemeinsame Informationsgehalte,
  - Hierarchien von Variablen* um Abhängigkeiten zu beschreiben (e.g. z.B. zwischen Jahr und Monat) oder um eine Gruppierung ähnlicher Variable z.B. durch eine Hauptkomponentenanalyse
- der mittlere Informationsgehalt der Datenmenge*
- die Qualität der Datenmenge$^{\star\triangleright}$, zum Beispiel die Anzahl der Fehlwerte
- der Ursprung einer Datenmenge$^{\triangleright}$unter anderem
  - Art der Datenerhebung (Messung, Simulation oder Beobachtung)
  - Erzeuger der Datenmenge
  - Informationen zur Reproduzierbarkeit der Datenerhebung (z.B. Mess- oder Simulationsbedingungen, Zeit und Ort der Datenerhebung)

Metadaten für relevante Datenteilmengen

- die Anzahl und die Art von Teilmengen von Interesse$^{\circ\star}$
- Metadaten für diese Teilmengen, u.a.
  - die Anzahl und Arbeit der Dimensionen$^{\circ}$
  - die Anzahl der beteiligten Merkmale$^{\circ}$
  - die enthaltenen Datensätze$^{\circ}$
  - die spezifischen Eigenschaften der Teilmenge*(z.B. Verteilung oder Lage von Extremen)

Legende: ∘ BESCHREIBENDE ⋆ ABGELEITETE ▷ HISTORISCHE Metadaten

Metadaten für die Variablen

Allgemeine Metadaten für die Variablen

- der Variablenname$^{\circ}$
- der Skalentyp$^{\circ}$
- die Einteilung in abhängige (Merkmale) und unabhängige (Dimension)$^{\circ}$
- weitere semantische und historische Informationen$^{\circ\triangleright}$
- der Informationsgehalt*, e.g. based on Shannon's entropy
- die Qualität der Variablenwerte$^{\star\triangleright}$ (Informationen über die Unsicherheit einzelner Variablenwerte oder aggregierte Informationen zur Qualität der Variablen (z.B. Anzahl von Fehlwerten))

Metadaten für die Merkmale

- der Datentyp$^{\circ}$
- Verteilungseigenschaften*, z.B. Werteverteilungen, Minima und Maxima, Varianzen oder Mittelwerte

Metadaten für die Dimensionen

- die Art des Raumes, der durch die Dimensionen aufgespannt wird$^{\circ}$(z.B. Kartesischer Raum mit räumlichen und/oder zeitlichen Dimensionen oder ein abstrakter Raum aus Dimensionen beliebiger Art)
- die Dimensionalität desRaumes$^{\circ}$
- die Beschreibung der räumlichen und/oder zeitlichen Dimensionen$^{\circ}$, beinhaltend die Beschreibung von:
  - Art des zugrunde liegenden Gitters$^{\circ}$
  - Wirkungskreis$^{\circ}$(punktuell, regional, global)
- Teilräume von Interesse* mit speziellen Eigenschaften:
  - räumliche Verteilung der Beobachtungspunkte und der Datenwerte
  - Heterogenität von räumlichen Gebieten
  - Qualität der Werte$^{\star\triangleright}$
  - Relevanz und Interesse (in Abhängigkeit von den anderen Eigenschaften)

Metadaten für die Datenklassen

Metadaten für Strömungsdaten

- Zeitabhängigkeit$^{\circ}$
- Dimensionalität der Strömungsvektoren$^{\circ}$
- topologische Eigenschaften des Stömungsfeldes*(Art und Lage von kritischen Punkten und separierenden Regionen)
- weitere allgemeine Eigenschaften des Feldes*(z.B. Krümmung, Divergenz, Wirbel, Wellenfronten, u.a.)

Metadaten für Volumendaten

- Volumendatenspezifische Segmentierungen*des Raumes aufgrund
  - das funktionale Verhalten der unterliegenden Skalarfunktion
  - die Topologie, die Form und die Anzahl von inneren Körpern für relevante Schwellwerte
- die Eigenschaften des Gradientenfeldes*

Metadaten für Multiparameterdaten

- spezielle Beobachtungsfälle* (engl.: data records) wie
  - Ausreißer
  - Beobachtungsfälle mit gleichen Merkmalswerten
  - typische Beobachtungsfälle
- Strukturen auf den Beobachtungsfällen* in Abhängigkeit von
  - der Verteilung und der Größe von Clustern
  - den Eigenschaften der Cluster

Legende: $\circ$ BESCHREIBENDE $\star$ ABGELEITETE $\triangleright$ HISTORISCHE Metadaten

B.2 XML-Repräsentation von Zielstellungen

B.2.1 Elementare Zielstellungen

```
<?xml version="1.0" encoding="UTF-8" ?>
<GoalFile>

    <GoalItem>
        <Value text = "local"/>
        <Type text = "AO"/>
        <Specialization text = "general"/>
        <Context text = "neutral"/>
        <Description text = "analyze regional data distribution
            (spatial/temporal)" />
    </GoalItem>

    <GoalItem>
        <Value text= "global"/>
        <Type text= "AO"/>
        <Specialization text  = "general"/>
        <Context text = "neutral"/>
        <Description text = "analyze the global data distribution (all
            grid cells or stations for all time steps)" />
    </GoalItem>

    <GoalItem>
        <Value text= "pointwise"/>
        <Type text= "AO"/>
        <Specialization text  = "general"/>
        <Context text = "neutral"/>
        <Description text = "analyze single stations or grid cells" />
    </GoalItem>

    <GoalItem>
        <Value text= "scalars"/>
        <Type text= "TA"/>
        <Specialization text  = "general"/>
        <Context text = "neutral"/>
        <Description text = "analyze scalar, continous aspect of data" />
    </GoalItem>

    <GoalItem>
        <Value text= "shapes"/>
        <Type text= "TA"/>
        <Specialization text  = "general"/>
        <Context text = "neutral"/>
        <Description text = "analyze spatial/temporal shapes/objects
            within the data" />
    </GoalItem>
```

```
<GoalItem>
    <Value text= "outliers"/>
    <Type text= "TA"/>
    <Specialization text  = "general"/>
    <Context text = "neutral"/>
    <Description text = "focus on extremes within the data"/>
</GoalItem>

...

<GoalItem>
    <Value text= "associate"/>
    <Type text= "AC"/>
    <Specialization text  = "general"/>
    <Context text = "neutral"/>
    <Description text = "relate/correlate variables"/>
</GoalItem>

<GoalItem>
    <Value text= "classify"/>
    <Type text= "AC"/>
    <Specialization text  = "general"/>
    <Context text = "neutral"/>
    <Description text = "identify groups within the data and display
        them separately"/>
</GoalItem>

<GoalItem>
    <Value text= "compare"/>
    <Type text= "AC"/>
    <Specialization text  = "general"/>
    <Context text = "neutral"/>
    <Description text = "analyze the differences of data values in a
        certain region and/or at a certain time step"/>
</GoalItem>

<GoalItem>
    <Value text= "identify"/>
    <Type text= "AC"/>
    <Specialization text  = "general"/>
    <Context text = "neutral"/>
    <Description text = "identify the data values in a certain region"/>
</GoalItem>

 <GoalItem>
    <Value text= "localize"/>
    <Type text= "AC"/>
    <Specialization text  = "general"/>
    <Context text = "neutral"/>
    <Description text = "analyze station/grid cell positions and
```

```
            locations of certain values" />
    </GoalItem>

    ...

    <GoalItem>
        <Value text= "overview"/>
        <Type text= "AT"/>
        <Specialization text  = "general"/>
        <Context text = "neutral"/>
        <Description text = "provide a general overview about the data"/>
    </GoalItem>

     <GoalItem>
        <Value text= "details"/>
        <Type text= "AT"/>
        <Specialization text  = "general"/>
        <Context text = "neutral"/>
        <Description text = "the focus is to gain data details"/>
    </GoalItem>

   ...

    <GoalItem>
        <Value text= "surfaces"/>
        <Type text= "TA"/>
        <Specialization text  = "volume data"/>
        <Context text = "neutral"/>
        <Description text = "analyze (iso)surfaces in volume data"/>
    </GoalItem>


    <GoalItem>
        <Value text= "segment"/>
        <Type text= "AC"/>
        <Specialization text  = "volume data"/>
        <Context text = "neutral"/>
        <Description text = "segment volumes from each other"/>
    </GoalItem>

    <SenselessCombination>
        <AT value1 = "direction"/>
        <AO value1 = "point"/>
    </SenselessCombination>

</GoalFile>
```

B.2.2 Zusammengesetzte Zielstellungen

```
<StoredGoals>
    <Goal name = "Analyze outliers">
        <Image name = "outliers.png"/>
        <AO v0 = "global" v1 = "local" v2 = "pointwise"/>
        <AC v0 = "associate" v1 = "compare" v2 = "identify" v3 = "localize"
            v4 = "rank" v5 = "reveal"/>
        <TA v0 = "outliers"/>
        <AT/>
    </Goal>
    <Goal name = "Analyze surfaces and lines ">
        <Image name = "lines.png"/>
        <AO v0 = "global" v1 = "local"/>
        <AC v0 = "associate" v1 = "classify" v2 = "compare" v3 = "identify"
            v4 = "localize" v5 = "rank" v6 = "segment"/>
        <TA v0 = "shapes" v1 = "structures" v2 = "surfaces"/>
        <AT/>
    </Goal>
    <Goal name = "Analyze temporal trends">
        <Image name = "trends.png"/>
        <AO v0 = "global" v1 = "local" v2 = "pointwise"/>
        <AC v0 = "associate" v1 = "compare" v2 = "identify" v3 = "localize"
            v4 = "rank" v5 = "reveal"/>
        <TA v0 = "trends"/>
        <AT v0 = "overview"/>
    </Goal>
    <Goal name = "Classify and/or segment">
        <Image name = "classify.png"/>
        <AO v0 = "global" v1 = "local" v2 = "pointwise"/>
        <AC v0 = "classify" v1 = "segment"/>
        <TA v0 = "scalars"/>
        <AT/>
    </Goal>
    <Goal name = "Compare variables or regions">
        <Image name = "compare.png"/>
        <AO v0 = "local"/>
        <AC v0 = "compare"/>
        <TA v0 = "directions" v1 = "frequencies" v2 = "outliers" v3 = "scalars"
            v4 = "shapes" v5 = "structures" v6 = "surfaces" v7 = "trends"/>
        <AT/>
    </Goal>
    <Goal name = "Find hidden patterns">
        <Image name = "Hidden.png"/>
        <AO v0 = "global" v1 = "local"/>
        <AC v0 = "identify" v1 = "reveal"/>
        <TA v0 = "outliers"/>
        <AT v0 = "overview"/>
    </Goal>
    <Goal name = "Find variables dependencies">
```

```
        <Image name = "dependencies.png"/>
        <AO v0 = "global" v1 = "local" v2 = "pointwise"/>
        <AC v0 = "associate" v1 = "compare" v2 = "segment"/>
        <TA v0 = "scalars" v1 = "trends"/>
        <AT/>
    </Goal>
    <Goal name = "Get station/grid cell details">
        <Image name = "detail.png"/>
        <AO v0 = "pointwise"/>
        <AC v0 = "compare" v1 = "identify" v2 = "localize" v3 = "rank"
            v4 = "reveal"/>
        <TA v0 = "directions" v1 = "frequencies" v2 = "scalars" v3 = "shapes"
            v4 = "structures" v5 = "trends"/>
        <AT v0 = "details"/>
    </Goal>
    <Goal name = "Overview">
        <Image name = "overview.png"/>
        <AO v0 = "global"/>
        <AC v0 = "associate" v1 = "identify" v2 = "localize"/>
        <TA v0 = "directions" v1 = "frequencies" v2 = "outliers" v3 = "scalars"
            v4 = "shapes" v5 = "structures" v6 = "trends"/>
        <AT v0 = "overview"/>
    </Goal>
</StoredGoals>
```

B.3 Deskriptoren für Visualisierungstechniken am Beispiel der 3D-Technik

```
<Technique text = "3D visualization">
   <VisSystemType text = "OpenDX"/>
   <NetName text = "3DTech.net"/>
   <ShortName text = "3D Vis"/>
   <HelpFile text = "3DVis.html"/>
   <ShortTextDescription text= "The 3D visualization module includes a variety
      of modes for regular and blockstructured 3D and 4D data. These are direct
      volume rendering, glyph visualization, isosurface visualization and slice
      visualization."/>
   <DevelopmentState text="PreRelease"/>
   <Image filename = "3DTech"/>

   <DimensionsDescriptor>
      <generalDimensionality>
         <minoptmax min="3" optmin="3" optmax="3" max="4"/>
      </generalDimensionality>
      ...
      <TimeSuitablity text = "1.0"/>
   </DimensionsDescriptor>

   <GridDescriptor>
```

```
      <regular text ="yes"/>
      <blockStructured text ="yes"/>
      <scattered text ="no"/>
      <punctual text ="yes"/>
      <local text ="yes"/>
      <global text ="yes"/>
   </GridDescriptor>

   <MapReliefInclusion>
      <MapInclusion text ="yes"/>
      <ReliefInclusion text ="no"/>
   </MapReliefInclusion>

   <DatSetDescription>
      <NrDataSets>
         <minoptmax min="2" optmin="8" optmax="800000" max="Unlimited"/>
      </NrDataSets>
      <NrAttributes>
         <minoptmax min="1" optmin="1" optmax="1" max="Unlimited"/>
      </NrAttributes>
   </DatSetDescription>

   <Parameters>
      <Parameter name = "GMType" type = "Selector">
         <SelectorValue name = "Volume rendering" value = "4"/>
         <SelectorValue name = "Slices (data)" value = "3"/>
         <SelectorValue name = "Slices (changes)" value = "5"/>
         <SelectorValue name = "Glyphs" value = "1"/>
         <SelectorValue name = "Isosurfaces" value = "2"/>
         <SelectorValue name = "3D Slices" value = "6"/>
      </Parameter>
      <Parameter name = "ScaleType" type = "Selector">
         <SelectorValue name = "default" value = "0"/>
         <SelectorValue name = "cube" value = "1"/>
      </Parameter>
      <Parameter name = "GridType" type = "Selector">
         <SelectorValue name = "on" value = "2"/>
         <SelectorValue name = "off" value = "1"/>
      </Parameter>
      <Parameter name = "RenderMode" type = "Selector">
         <SelectorValue name = "Software" value = "1"/>
         <SelectorValue name = "Hardware" value = "2"
              availability="checkGLXExtension"/>
      </Parameter>
      <Parameter name = "Background_color" type = "Selector" class = "general">
         <SelectorValue name = "black" value = "1"/>
         <SelectorValue name = "dark gray" value = "2"/>
         <SelectorValue name = "gray" value = "3"/>
         <SelectorValue name = "white" value = "4"/>
         <Rules>
```

```
            <Rule>[Parameter] hasUserPreference ("PreferWhiteBackground") ?
               [=4, =2];</Rule>
         </Rules>
      </Parameter>
   </Parameters>

   <VariationSet name = "General mapping">
      <Variation text = "Cutting planes">
         <ParameterKombination>
            <ParameterValue name = "GMType" value = "6"/>
            <ParameterValue name = "ScaleType" value = "1"/>
            <ParameterValue name = "GridType" value = "2"/>
            <ParameterValue name = "RenderMode" value = "1"/>
         </ParameterKombination>
         <ShortVariationDescription text= "Cutting planes visualizaion
             - 3 cutting planes ortogonal to all the 3 dimensions are displayed.
     The  planes include color coding and isolines."/>
         <VariationImage name = "3DCuttingSlices.gif"/>
         <Rules>
            <Rule>[ModuleVariation] exists(Dimension, "Name", "longitude")
                 AND exists(Dimension, "Name", "lattitude") ? [+=0.05];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "local" ) ? [+=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "pointwise" ) ? [+=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "scalars" ) ? [+=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "directions" ) ? [+=0.05];</Rule>
            ...
            <Rule>[ModuleVariation] hasGoal ( "segment" ) ? [-=0.1];</Rule>
         </Rules>
      </Variation>
      <Variation text = "Slices (data)">
         <ParameterKombination>
            <ParameterValue name = "GMType" value = "3"/>
            <ParameterValue name = "ScaleType" value = "1"/>
            <ParameterValue name = "GridType" value = "2"/>
            <ParameterValue name = "RenderMode" value = "1"/>
         </ParameterKombination>
         <ShortVariationDescription text= "Data slice visualization - Along
             one dimension semi-transparent planes represent colored cuts in data
             space. The number of cuts can be interactively adapted (overview and
             detail-on-demand)."/>
         <VariationImage name = "3DSlices.gif"/>
         <Rules>
            <Rule>[ModuleVariation] exists(Dimension, "Name", "longitude")
                AND exists(Dimension, "Name", "lattitude") ? [+=0.05];</Rule>
            <Rule>[ModuleVariation] exists(Dimension, "Name", "time") ? [+=0.1];
            </Rule>
            <Rule>[ModuleVariation] hasGoal ( "global" ) ? [+=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "local" ) ? [+=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "scalars" ) ? [+=0.1];</Rule>
            ...
```

```
            <Rule>[ModuleVariation] hasUserPreference ("Prefer2DDisplays") ?
               [+=0.1];</Rule>
         </Rules>
      </Variation>
      <Variation text = "Volume rendering">
         <ParameterKombination>
            <ParameterValue name = "GMType" value = "4"/>
            <ParameterValue name = "ScaleType" value = "1"/>
            <ParameterValue name = "GridType" value = "2"/>
            <ParameterValue name = "RenderMode" value = "2"/>
         </ParameterKombination>
         <ShortVariationDescription text= "Volume rendering - This technique
             casts rays into the 3D dataspace and collects information along this
     ray, resulting a certain color value. The current OpenDX default
     implementation does a simple alpha blending along the ray."/>
         <VariationImage name = "3DVolRend.gif"/>
         <Rules>
            <Rule>[ModuleVariation] exists(Dimension, "Name", "longitude")
                AND exists(Dimension, "Name", "lattitude") ? [-=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "global" ) ? [+=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "local" ) ? [+=0.05];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "pointwise" ) ? [-=0.1];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "scalars" ) ? [+=0.1];</Rule>
            ...
            <Rule>[ModuleVariation] hasGoal ( "surfaces" ) ? [+=0.05];</Rule>
            <Rule>[ModuleVariation] hasGoal ( "segment" ) ? [+=0.1];</Rule>
            <Rule>[ModuleVariation] hasUserPreference ("Prefer2DDisplays") ?
                [-=0.1];</Rule>
         </Rules>
      </Variation>
      ...
</Technique>
```

Literaturverzeichnis

Abello u. a. 2001
Abello, J. ; Finocchi, I. ; Korn, J.: Graph Sketches. In: *IEEE Symposium on Information Visualization (InfoVis'01), San Diego*, 2001, S. 67–72

Abram u. Treinish 1995
Abram, G. ; Treinish, L.: An Extended Data-Flow Architecture for Data Analysis and Visualization. In: *Proceedings of the IEEE Visualization (Vis'95)*, 1995, S. 263–270

Ahlberg u. Shneiderman 1994
Ahlberg, C. ; Shneiderman, B.: Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In: *Human Factors in Computing Systems. Conference Proceedings CHI'94*, 1994, S. 313–317

Ahumada 1993
Ahumada, A.J.: Computational Image Quality Metrics: A Review. In: *Society for Information Display, International Syposium Digest of Technical Papers* 24 (1993), S. 305–308

Alexa u. Müller 1999
Alexa, M. ; Müller, W.: Visualization by Examples: Mapping Data to Visual Representations using Few Correspondences. In: *Proceedings of the Joint Eurographics - IEEE TCVG Symposium on Visualization*, 1999

Amar u. a. 2005
Amar, R. ; Eagan, J. ; Stasko, J.: Low-Level Components of Analytic Activity in Information Visualization. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05)*. Mineapolis, USA, 2005, S. 111–117

Amar u. Stasko 2004
Amar, R. ; Stasko, J.: A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations, 2004, S. 143–149

American Meteorological Society 1993
American Meteorological Society: Guidelines for using Color to Depict Meteorological Information: IIPS Subcomitte for Color Guidelines. In: *Bull. Amer. Meteor. Soc.* (1993), Nr. 74, S. 1709–1713

Andrienko u. Andrienko 1999
Andrienko, G. ; Andrienko, N.: Data Characterization Schema for Intelligent Support in Visual Data Analysis. In: Freksa, C. (Hrsg.) ; Mark, D.M. (Hrsg.): *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science Conference On Spatial Information Theory*. Springer, August 1999, S. 349–366

Andrienko u. Andrienko 2006
Andrienko, N. ; Andrienko, G.: *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach.* Springer-Verlag, Berlin, Heidelberg, 2006

Ankerst 2001
Ankerst, M.: Visual Data Mining with Pixel-oriented Visualization Techniques. In: *Proceedings of ACM SIGKDD Workshop on Visual Data Mining'01; San Francisco*, 2001

Ankerst u. a. 1999
Ankerst, M. ; Breunig, M. ; Kriegel, H.P. ; Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure. In: *Proceedings ACM SIGMOD'99, Int. Conf. on Management of Data, Philadelphia, USA*, 1999, S. 49–60

Ankerst u. a. 2000
Ankerst, M. ; Ester, M. ; Kriegel, H.-P.: Towards an Effective Cooperation of the User and the Computer for Classification. In: *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD'00)*, 2000, S. 179–188

ArcInfo 2007
ArcInfo: http://www.esri.com/software/arcgis/arcinfo/index.html. In: *Internet page*, 2007

Arens u. a. 1993
Arens, Y. ; Hovy, E. ; Vossers, M.: The knowledge underlying multimedia presentations. In: *M. Maybury (ed.): Intelligent Multimedia Intetiaces.* The MIT Press, 1993, S. 280–306

Artero u. a. 2004
Artero, A.O. ; Oliveira, M.C.F. de ; Levkowitz, H.: Uncovering Clusters in Crowded Parallel Coordinates Visualizations. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'04).* Austin, USA, 2004, S. 81–88

Avnur u. a. 1998
Avnur, R. ; Hellerstein, J. ; Lo, B. ; Olston, C. ; Raman, R. ; Roth, T. ; Wylie, K.: CONTROL: Continous Output and Navigation Technology with Refinement On-Line. In: *ACM SIGMOD'98, Seatle, USA*, 1998, S. 567–569

Baalcke 2005
Baalcke, M.: *Mosaic Based Visualization.* Studienarbeit, Universität Rostock, Institut für Informatik, 2005

Bade u. a. 2006
Bade, R. ; Haase, J. ; Preim, B.: Comparison of Fundamental Mesh Smoothing Algorithms for Medical Surface Models. In: *Simulation and Visualization (SimVis'06)*, 2006, S. 289–304

Baker u. Bushell 1995
Baker, M. P. ; Bushell, C.: After the Storm: Considerations for Information Visualization. In: *IEEE Comput. Graph. Appl.* 15 (1995), Nr. 3, S. 12–15. – ISSN 0272–1716

Balzer u. Deussen 2005
Balzer, M. ; Deussen, O.: Voronoi Treemaps. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05).* Mineapolis, USA, 2005, S. 49–56

Barlow u. Neville 2001
Barlow, T. ; Neville, P.: Case Study: Visualization for Decision Tree Analysis in Data Mining. In: *Proc. IEEE Information Visualization'01, San Diego, IEEE Press*, 2001

Bennet u. a. 2003
BENNET, J. ; MAHROUS, K. ; HAMANN, B. ; JOY, K.I.: A Segmentation Approach to Scientific Visualization . In: *Proceedings Spring Conference on Computer Graphics'03, Budmerice*, 2003, S. 11–20

Bergeron u. Grinstein 1989
BERGERON, R.D. ; GRINSTEIN, G.G.: *A Reference Model for the Visualisation of Multidimensional Data*. Proceedings Eurographics '89, 1989

Bergman u. a. 1995
BERGMAN, L. D. ; ROGOWITZ, B. E. ; TREINISH, L. A.: A Rule-based Tool for Assisting Colormap Selection. In: *Visualization '95 (Atlanta)*, IEEE Computer Society Press, 1995, S. 118–125, 444

Bergmann u. a. 1995
BERGMANN, L. D. ; ROGOWITZ, B. E. ; TREINISH, L.: A Rule-based Tool for Assisting Colormap Selection. In: *Proceedings IEEE Visualization'95, Atlanta, USA*, 1995

Berrer u. a. 2000
BERRER, H. ; PATERSON, I. ; KELLER, J.: Evaluation of Machine-Learning Algorithm Ranking Advisors. In: *Proceedings of the IDDM Workshop*, 2000

Bertin 1981
BERTIN, J.: *Graphics and Graphic Informations-Processing*. Walter de Gruyter, Berlin, New York, 1981

Bertin 1983
BERTIN, J.: *Semiology of Graphics*. The University of Wisconsin Press, 1983

Beshers u. Feiner 1993
BESHERS, C. ; FEINER, S.: AutoVisual: Rule-Based Design of Interactive Multivariate Visualizations. In: *IEEE Computer Graphics and Applications* (1993), July

Bier u. a. 1993
BIER, E.A. ; STONE, M.C. ; K.PIER ; BUXTON, W. ; ROSE, T. de: Toolglass and magic lenses: The see-through interface. In: *Proc. SIGGRAPH'93*, 1993, S. 73–80

Bill 1991
BILL, Ralph: *Grundlagen der Geo-Informationssysteme*. Karlsruhe : Herbert Wichmann Verlag, 1991

Bock 1974
BOCK, Hans H.: *Automatische Klassifikation*. Vandenhoeck & Ruprecht, Göttingen, 1974

Bollinger 1996
BOLLINGER, T.: Assoziationsregeln - Analyse eines Data Mining Verfahrens. In: *Informatik-Spektrum* 19 (1996), Nr. 5, S. 257–261

Brandes u. Corman 2002
BRANDES, U. ; CORMAN, S.R.: Visual Unrolling of Network Evolution and the Analysis of Dynamic Discourse. In: *IEEE Symposium on Information Visualization (InfoVis'02), Boston*, 2002, S. 145–151

Brandes u. a. 2005
Brandes, U. ; Fleischer, D. ; Lerner, J.: Highlighting Conflict Dynaics in Event Data. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05)*. Mineapolis, USA, 2005, S. 103–110

Brazdil u. a. 2003
Brazdil, P. ; Soares, C. ; Costa, J.: Ranking Learning Algorithms. In: *Machine Learning* (2003)

Brazdil u. Soares 1997
Brazdil, P.B. ; Soares, C.: A Comparison of Ranking Methods for Classification Algorithm Selection. In: *Machine Learning (ECML'00)* 1810 (1997), S. 63–74

Brewer 1999
Brewer, C.A.: Color Use Guidelines for Data Representation . In: *Proceedings of the section on Statistical Graphics. American Statistical Association. Alexandria VA*, 1999, S. 55–60

Brockhaus 1970
Brockhaus: *Die Enzyklopädie*. 17. 1970

Brockhaus 1996
Brockhaus: *Die Enzyklopädie*. 20. 1996

Brockmann 2004
Brockmann, P.: PRISM - The VTK_Mapper Application / PRISM project page: http://prism.enes.org/Results/Documents/. 2004 (PRISM-Report Series-19). – Forschungsbericht

Brodlie u. a. 2004
Brodlie, K. ; Duce, D. ; Duke, D.: Visualization Ontologies / Report of a Workshop held at the National e-Science Centre. 2004. – Forschungsbericht

Brodlie u. Wood 2001
Brodlie, K. ; Wood, J.: Recent Advances in Volume Visualization. In: *Computer Graphics Forum* 20 (2001), June, Nr. 2, S. 775–792

Brodlie 1992
Brodlie, K. W.: *Scientific Visualisation - Techniques and Applications*. Springer-Verlag, Berlin, 1992

Bruckner u. Gröller 2006
Bruckner, S. ; Gröller, E.: Exploded Views for Volume Data. In: *IEEE Transactions on Visualization and Computer Graphics* 12 (2006), 9, Nr. 5, S. 1077–1084. – ISSN 1077–2626

Brunk u. a. 1997
Brunk, C. ; Kelly, J. ; Kohavi, R.: MineSet: An Integrated System for Data Mining. In: Heckerman, David (Hrsg.) ; Mannila, Heikki (Hrsg.) ; Pregibon, Daryl (Hrsg.) ; Uthurusamy, Ramasamy (Hrsg.): *Proceedings of the third international conference on Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, California, 135–138

Böhm 1999
Böhm, U.: *Eine Methode zur Validierung von Klimamodellen für die Klimawirkungsforschung hinsichtlich der Wiedergabe extremer Ereignisse (in german)*. Dissertation, Freie Universität Berlin, Fachbereich Geowissenschaften, 1999

Böhm u. a. 2005
Böhm, U. ; Gerstengarbe, F.-W. ; Hauffe, D. ; Kücken, M. ; Österle, H. ; Werner, P.C.: Dynamic Regional Climate Modeling and Sensitivity Experiments for the Northeast of Brazil. In: *Global Change and Regional Impacts*. Springer Verlag, Berlin, 2005, S. 428ff

Böhm u. a. 2004
Böhm, U. ; Kücken, M. ; Gerstengarbe, F.-W. ; Werner, P.C. ; Flechsig, M. ; K.Keuler ; Block, A. ; Ahrens, W. ; Nocke, T.: Reliability of regional climate model simulations of extremes and of long-term climate. In: *Natural Hazards and Earth System Sciences* (2004), Nr. 4, S. 417–431

Böttger u. a. 2006
Böttger, J. ; Balzer, M. ; Deussen, O.: Complex Logarithmic Views for Small Details in Large Contexts. In: *IEEE Trans. Vis. Comput. Graph.* 12 (2006), Nr. 5, S. 845–852

Cameron 1995
Cameron, G.: Special Focus: Modular Visualization Environments (MVEs). 29 (1995), Nr. 2, S. 3–60

Card u. Mackinlay 1997
Card, S. K. ; Mackinlay, J.: The structure of the information visualization design space. In: *IEEE Symposium on Information Visualization (InfoVis'97)*, 92–99

Card u. a. 1999
Card, S. K. (Hrsg.) ; Mackinlay, J.D. (Hrsg.) ; Shneiderman, B. (Hrsg.): *Readings in Information Visualization - Using Vision to Think*. San Francisco : Morgan Kaufmann Publishers, 1999

Casner 1991
Casner, S. M.: A Task-Analytic Approach to the Automated Design of Graphic Presentations. In: *ACM Transactions on Graphics* 10 (1991), April, Nr. 2

Cedilnik u. Rheingans 2000
Cedilnik, A. ; Rheingans, P.: Procedural Annotation of Uncertain Information. In: *IEEE Visualization'00*, 2000

Chan u. Stolfo 1997
Chan, P.K. ; Stolfo, S.J.: On the Accuracy of Meta-Learning for Scalable Data Mining. In: *Journal of Intelligent Information Systems* 8 (1997), Nr. 1

Chen 2003
Chen, H.: Compound brushing explained. In: *Palgrave Information visualization* 4 (2003), Nr. 2

Chen 2006
Chen, M.: Feature Aligned Volume Manipulation for Illustration and Visualization. In: *IEEE Transactions on Visualization and Computer Graphics* 12 (2006), Nr. 5, S. 1069–1076. – Student Member-Carlos Correa and Member-Deborah Silver

Chen 1993
Chen, P.C.: A Climate Simulation Study. In: *Proceedings IEEE Visualization'93, USA*, 1993, S. 397–401

Chernoff 1973
CHERNOFF, H.: The Use of Faces to Represent Points in k-Dimensional Space Graphically. In: *Journal of American Statistical Association* 68 (1973), S. 361–368

Chi u. Riedl 1998
CHI, E.H. ; RIEDL, J.T.: An Operator Interaction Framework for Visualization Systems. In: *Proceedings of the Symposium on Information Visualization '98* (1998)

Chi u. Riedl 2000
CHI, E.H. ; RIEDL, J.T.: A Taxonomy of Visualization Techniques using the Data State Reference Model. In: *Proceedings of the Symposium on Information Visualization '00* (2000)

Chuah u. Eick 1998
CHUAH, M. C. ; EICK, S. G.: Information Rich Glyphs for Software Management Data. In: *IEEE Computer Graphics and Applications* 18 (1998), Juli/August, Nr. 4, S. 24–29

Chuah u. Roth 1996
CHUAH, M.C. ; ROTH, S.F.: On the Semantics of Interactive Visualization. In: *Proceedings IEEE Visualization'96, USA*, 1996, S. 29–36

CiteWiz 2007
CITEWIZ: *Project: CiteWiz – Scientific Citation Visualization*. Version: 2007. `http://www.cs.chalmers.se/\%7Eelm/projects/citewiz`. – Web page.

Cooke u. van Noortwijk 2000
COOKE, R.M. ; NOORTWIJK, J.M. van: Graphical Methods. In: SALTELLI (Hrsg.) ; CHAN (Hrsg.) ; SCOTT (Hrsg.): *Sensitivity Analysis*. Wiley, 2000

Cox u. Cox 1994
COX, T.F. ; COX, M.A.A.: *Multidimensional Scaling*. Chapman & Hall, 1994

Crowfis u. Max 1992
CROWFIS, R.A. ; MAX, N.: Direct Volume Visualization of Three-Dimensional Vector Fields. In: *Proceedings Workshop on Volume Visualization, Boston, USA*, 1992

Crowfis u. a. 2000
CROWFIS, R.A. ; SHEN, H.W. ; MAX, N.: Flow Visualization Techniques for CFD Using Volume Rendering. In: *9th International Symposium on flow Visualization*, 2000

Dameris u. a. 2005
DAMERIS, M. ; GREWE, V. ; PONATER, M. ; DECKERT, R. ; EYRING, V. ; MAGER, F. ; MATTHES, S. ; SCHNADT, C. ; STENKE, A. ; STEIL2, B. ; BRÜHL, C. ; GIORGETTA, M.A.: Long-term changes and variability in a transient simulation with a chemistry-climate model employing realistic forcing. In: *Atmosheric Chemistry and Physics Discussions, European Geosciences Union* 5 (2005), S. 2297–2353

Davidson u. a. 2001
DAVIDSON, G.S. ; WYLIE, B.N. ; BOYACK, K.W.: Cluster Stability and the Use of Noise in Interpretation of Clustering. In: *Proceedings IEEE Symposium on Information Visualization 2001 (InfoVis'01)*. San Diego, California, 2001, S. 73–78

Deines u. a. 2006
DEINES, E. ; BERTRAM, M. ; MOHRING, J. ; JEGOROVS, J. ; MICHEL, F. ; HAGEN, H. ; NIELSON, G.M.: Comparative Visualization for Wave-based and Geometric Acoustics. In: *IEEE Trans. Vis. Comput. Graph.* 12 (2006), Nr. 5, S. 1173–1180

Dent 1999
DENT, B. D.: *Cartography, Thematic Map Design.* fifth. McGraw-Hill, 1999

Djurcilov u. a. 2001
DJURCILOV, S. ; KIM, K. ; LERMUSIAUX, P.F.J. ; PANG, A.: Volume Rendering Data with Uncertainty Information. In: *Data Visualisation'01: Proceedings of the EG+IEEE VisSym in Ascona*, 2001

Djurcilov u. Pang 1998
DJURCILOV, S. ; PANG, A.: Visualization Products On-Demand Through the Web. In: *Third Symposium on the Virtual Reality Modeling Language (VRML'98)*, 1998

Doleisch 2005
DOLEISCH, H.: *Visual Analysis of Complex Simulation Data using Multiple Heterogenous Views*, Vienna University of Technology, Institute for Computergraphics and Algorithms, Diss., 2005

Doleisch u. a. 2003a
DOLEISCH, H. ; GASSER, M. ; HAUSER, H.: Interactive Feature Specification for Focus+Context Visualization of Complex Simulation Data. In: *IEEE TCVG - EUROGRAPHICS Symposium on Visualization (VisSym'03)*, 2003

Doleisch u. a. 2003b
DOLEISCH, H. ; GASSER, M. ; HAUSER, H.: Interactive Feature Specification for Focus+Context Visualization of Complex Simulation Data. In: *Proceedings of the 5th Joint IEEE TCVG - EG Symposium of Visualization (VisSym'03)*, 2003, S. 239–248

Doleisch u. Hauser 2002
DOLEISCH, H. ; HAUSER, H.: Smooth Brushing for Focus+Context Visualization of Simulation Data in 3D. In: *International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media (WSCG'02)*, 2002, S. 147–154

Doleisch u. a. 2004
DOLEISCH, H. ; MUIGG, P. ; HAUSER, H.: *Interactive Visual Analysis of Hurricane Isabel with SimVis.* IEEE Visualization (Vis'04) Contest, http://vis.computer.org/vis2004contest/vrvis/, 2004

Doty 2006
DOTY, B.: *The Grid Analysis and Display System, User's Guide.* http://www.iges.org/grads/gadoc/, 2006

Dwyer u. Koren 2005
DWYER, T. ; KOREN, Y.: DIG-COLA: Directed Graph Layout through Constrained Eneergy Minimization. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05).* Mineapolis, USA, 2005, S. 65–72

Earnshaw u. Wiseman 1992
EARNSHAW, R.A. ; WISEMAN, N.: *An Introductory Guide to Scientific Visualization.* Springer, 1992

Eik 1994
EIK, S.G.: Data visualization sliders. In: *Proc. ACM UIST'94)*, 1994, S. 119–120

Eilt u. a. 1995
EILT, M.D. ; JOHNSON, J.T. ; MITCHELL, E. D. ; SANGER, S. ; STUMPF, G. ; WITT, A.

; Hondl, K. ; Thomas, K.: Warning Decision Support System. In: *11th Inter. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, AMS*, 1995, S. 62–67

Engel 2002
Engel, K.: Interactive High-Quality Volume Rendering with Flexible Consumer Graphics Hardware (State-of-the-Art Report). In: *Eurographics'02*, 2002, S. 19–25

Fayyad u. a. 2001
Fayyad, U. ; Grinstein, G. ; Wierse, A.: *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann Publishers, San Francisco, 2001

Fequete 2004
Fequete, J.-D.: The InfoVis Toolkit. In: *IEEE Symposium on Information Visualization (InfoVis'04), Austin*, 2004, S. 167–174

Finkelstein u. Range 1998
Finkelstein, A. ; Range, M.: Image Mosaics / Princeton University, Computer Science Department. 1998 (TR-574-98). – Forschungsbericht

Flechsig u. a. 2006
Flechsig, M. ; Böhm, U. ; Nocke, T. ; Rachimow, C.: *The Multi-Run Simulation Environment SimEnv - User Guide*. Potsdam Institute for Climate Impact Research. `http://www.pik-potsdam.de/software/simenv/simenv`. Version: May 2006

Flechsig u. a. 2007
Flechsig, M. ; Böhm, U. ; Nocke, T. ; Rachimow, C.: SIMENV - A flexible Framework for Sensitivity and Uncertainty Analyses of large-volume Model Output. In: *Poster Presentation at the 5th International Conference on Sensitivity Analysis of Model Output (SAMO'07), Budapest*, 2007

Frank 1998
Frank, A.U.: Different Types of 'Times' in GIS. In: Egenhofer, M.J. (Hrsg.) ; Golledge, R.G. (Hrsg.): *Spatial and Temporal Reasoning in Geographic Information Systems*. New York, Oxford : Oxford University Press, 1998, S. 40–62

Frühauf 1997
Frühauf, T.: *Graphisch-Interaktive Strömungsvisualisierung*. Springer-Verlag, Berlin, 1997

Fuchs u. a. 2006
Fuchs, G. ; Reichart, D. ; Schumann, H. ; Forbrig, P.: Maintenance Support - Case Study for a Multimodal User Interface. In: *S&T/SPIE's 16th Annual Symposium Electronic Imaging: Multimedia on Mobile Devices II, San Jose, California, USA*, 2006

Fujishiro u. a. 2000
Fujishiro, I. ; Furuhata, R. ; Y.Ichikawa ; Takeshima, Y.: GADGET/IV: A Taxonomic Approach to Semi-Automatic Design of Information Visualization Applications Using Modular Visualization Environment. In: *IEEE Information Visualization*, 77ff.

Fujishiro u. a. 1997
Fujishiro, I. ; Takeshima, Y. ; Ichikawa, Y. ; Nakamura, K.: GADGET: goal-oriented application design guidance for modular visualization environments. In: *IEEE Visualization*, 245-252

Furnas 1986
FURNAS, G.: Generalized Fisheye Views. In: *Proceedings ACM CHI'86* (1986)

Gelin 2002
GELIN, L.: Evaluation of Added Value in Multidimensional Visualization for SMHI. In: *European Working Group on Operational Workstations (EGOWS), Rome*, 2002

GeoVISTA 2007
GEOVISTA: http://www.geovistastudio.psu.edu. In: *Internet page*, 2007

Gerstengarbe u. Werner 1999
GERSTENGARBE, F.-W. ; WERNER, P.C.: The complete non-hierarchical Cluster-Analysis / Potsdam Institute for Climate Impact Research (PIK). 1999 (PIK-Report Nr. 50). – Forschungsbericht

Gilliland-Swetland 2000
GILLILAND-SWETLAND, A.J.: *Introduction to Metadata: Setting the Stage*. 2000

Gnanagmari 1981
GNANAGMARI, S.: *Information Presentation Through default Display*, University of Pennsilvania, Diss., 1981

Golovchinsky u. a. 1995
GOLOVCHINSKY, G. ; KAMPS, T. ; REICHENBERGER, K.: Subverting Structure: Data-driven Diagram Generation. In: *IEEE Visualization (Vis'95)* (1995)

Granitzer u. a. 2004
GRANITZER, M. ; KIENREICH, W. ; SABOL, V. ; ANDREWS, K. ; KLIEBER, W.: Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories. In: *IEEE Symposium on Information Visualization (InfoVis'04), Austin*, 2004, S. 127–133

GraphViz 2007
GRAPHVIZ: http://www.graphviz.org. In: *Internet page*, 2007

Graw u. a. 1997
GRAW, K. U. ; LANGE, S. ; CHAVEZ, N. L. ; SCHUMANN, H.: Konzept und Realisierung einer intelligenten Visualisierungshilfe / Universität Rostock. 1997. – Preprint

Griebel u. a. 2004
GRIEBEL, M. ; PREUSSER, T. ; RUMPF, M. ; SCHWEITZER, M.A. ; TELEA, A.: Flow Field Clustering via Algebraic Multigrid. In: *Proceedings IEEE Visualization'04, Austin, USA*, 2004

Griethe u. a. 2005
GRIETHE, H. ; FUCHS, G. ; SCHUMANN, H.: A Classification Scheme for Lens Techniques. In: *International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media (WSCG'05)*. Plzen, Czech Republic, 2005

Griethe u. Schumann 2005
GRIETHE, H. ; SCHUMANN, H.: Visualizing Uncertainty for Improved Decision Making. In: *4th International Conference on Perspectives in Business Informatics Research (BIR'05)*, 2005

Gross u. a. 1997
GROSS, M. H. ; SPRENGER, T.C. ; FINGER, J.: Visualizing Information on a Sphere. In: *IEEE Syposium on Information Visualization (InfoVis'97)*, 1997, S. 11–16

Hake u. Grünreich 1994
HAKE, G. ; GRÜNREICH, D.: *Kartographie*. 7th. Berlin : deGruyter, 1994

van Ham u. van Wijk 2002
HAM, F. van ; WIJK, J. J.: Beamtrees: Compact Visualization of Large Hierarchies. In: *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*. Washington, DC, USA : IEEE Computer Society, 2002, S. 93

van Ham u. van Wijk 2004
HAM, F. van ; WIJK, J. van: Interactive Visualization of Small World Graphs. In: *IEEE Symposium on Information Visualization (InfoVis'04), Austin*, 2004, S. 199–206

Han u. Kamber 2000
HAN, J. ; KAMBER, M.: *Data Mining: Concepts and Techniques*. 8. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, 2000. – ISBN 1–55860–489–8

Hauser u. a. 2001
HAUSER, H. ; MROZ, L. ; BISCHI, G.I. ; GRÖLLER, E.: Two-Level Volume Rendering. In: *IEEE Transactions on Visualization and Computer Graphics* 7 (2001), July-September, Nr. 3

Havre u. a. 2002a
HAVRE, S. ; HETZLER, E. ; WHITNEY, P. ; NOWELL, L.: ThemeRiver: Visualizing Thematic Changes in Large Document Collections. In: *IEEE Transactions on Visualization and Computer Graphics* 8 (2002), Nr. 1, S. 9–20. – ISSN 1077–2626

Havre u. a. 2002b
HAVRE, S. ; HETZLER, E. ; WHITNEY, P. ; NOWELL, L.: ThemeRiver: Visualizing Thematic Changes in Large Document Collections. In: *ACM Transactions on Graphics* 8 (2002), Nr. 1

HCE 2007
HCE: Hierarchical Clustering Explorer: http://www.cs.umd.edu/hcil/hce. In: *Internet page*, 2007

Hearst 1999
HEARST, M. A.: User Interfaces and Visualization. In: BAEZA-YATES, R. (Hrsg.) ; RIBEIRO-NET, B. (Hrsg.): *Modern Information Retrieval*. Harlow : Addison-Wesley, 1999

Heer u. Agrawala 2006
HEER, J. ; AGRAWALA, M.: Software Design Patterns for Information Visualization. In: *IEEE Trans. Vis. Comput. Graph.* 12 (2006), Nr. 5, 853-860. `http://dblp.uni-trier.de/db/journals/tvcg/tvcg12.html#HeerA06a`

Heer u. Boyd 2005
HEER, J. ; BOYD, D.: Vizster: Visualizing Online Social Networks. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05)*. Mineapolis, USA, 2005, S. 33–40

Hege 1992
HEGE, H.-C.: *Handbuch zur Visualisierung am ZIB*. `http://elib.zib.de/preprints/shadows/TR-92-07.html`. Version: 1992

Helgeland u. a. 2004
HELGELAND, A. ; ANDREASSEN, Ø. ; OMMUNDSEN, A. ; REIF, B.A. P. ; WERNE, J.: Visualization of the Energy-Containing Turbulent Scales. In: SILVER, D. (Hrsg.) ; ERTL, T. (Hrsg.) ; SILVA, C. (Hrsg.): *Proceedings of IEEE/SIGGRAPH Symposium on Volume Visualization*, 2004

Herman u. a. 1998
HERMAN, I. ; DELEST, M. ; MELANCON, G.: Tree Visualisation and Navigation Clues for Information Visualisation. In: *Computer Graphics Forum* 17(2) (1998), S. 153–165

Hibbard 2001
HIBBARD, W.: *Vis5D (Documentation).* http://vis5d.sourceforge.net/doc/, July 2001

Hinneburg u. a. 1999
HINNEBURG, A. ; KEIM, D. ; WAWRYNIUK, M.: HD-Eye: Visual Mining of High-Dimensional Data. In: *IEEE Computer Graphics and Applications* 19 (1999), Nr. 5, S. 22–31

Holst 2003
HOLST, M.: *Kopplung von interaktiver Modelldefinition und -visualisierung für die Klimafolgenforschung ?* Studienarbeit, Universität Rostock, Institut für Informatik, 2003

Hotz u. a. 2004
HOTZ, I. ; FENG, L. ; HAGEN, H. ; HAMANN, B. ; JOY, K. ; JEREMIC, B.: Physically Based Methods for Tensor Field Visualization. In: *Proceedings IEEE Visualization'04, Austin, USA*, 2004

Hrdlicka u. a. 2003
HRDLICKA, F. ; SLAVÍK, P. ; GAYER, M.: Real Time Simulation and Visualization using Pre-Calculated Fluid Simulator States. In: *Proceedings IEEE Information Visualization (InfoVis'03), USA*, 2003, S. 440–445

ILOGViews 2007
ILOGVIEWS: http://www.ilog.com/products/views. In: *Internet page*, 2007

InfoVis 2007
INFOVIS: http://ivtk.sourceforge.net. In: *Internet page*, 2007

InSpire 2007
INSPIRE: http://in-spire.pnl.gov. In: *Internet page*, 2007

Iserhardt-Bauer u. a. 2006
ISERHARDT-BAUER, S. ; HASTREITER, P. ; TOMANDL, B. ; ERTL, T.: Evaluation of Volume Growing Based Segmentation of Intracranial Anezrysms Combined with 2D Transfer Functions. In: *Proc. Simulation und Visualisierung (SimVis'06)*, 2006, S. 319–327

Islam u. a. 2004
ISLAM, S. ; DIPANKAR, S. ; SILVER, D. ; CHEN, M.: Spatial and Temporal Splitting of Scalar Fielsds in Volume Graphics. In: SILVER, D. (Hrsg.) ; ERTL, T. (Hrsg.) ; SILVA, C. (Hrsg.): *Proceedings of IEEE/SIGGRAPH Symposium on Volume Visualization*, 2004

ITK 2007
ITK: Medicine Insight Segmentation and Registration Toolkit: http://www.itk.org. In: *Internet page*, 2007

JavaTreeView 2007
JAVATREEVIEW: http://sourceforge.net/projects/jtreeview. In: *Internet page*, 2007

Jiang u. a. 2002
JIANG, M. ; MACHIRAJU, R. ; THOMPSON, D.: Geometric Verification of Swirling Features in Flow Fields. In: *Proc. IEEE Information Visualization 2002, IEEE Press*, 2002

Jiawei 2003
Jiawei, H.: *Reuse of Visization Design Knowledge*, UMIST, U.K., Diss., 2003

Jiawei u. a. 2004
Jiawei, H. ; Bailey, A. ; Sutcliffe, A.: Visualization Design Reuse. In: *Information Visualization (IV'04), London*, 2004

Johansson u. a. 2005
Johansson, J. ; Ljung, P. ; Jern, M. ; Cooper, M.: Revealing Structure Within Clustered Parallel Coordinate Displays. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05)*. Mineapolis, USA, 2005, S. 125–132

Johansson u. a. 2004
Johansson, J. ; Treloar, R. ; Jern, M.: Integration of Usupervised Clustering, Interavtion and Parallel Coordinates for the Explotation of Large Multivariate Data. In: *Information Visualization (IV'04), London*, 2004

Joliffe 1986
Joliffe, I. T.: *Principal Component Analysis. Series in Statistics.* Springer, 1986

Jones u. a. 1995
Jones, R. G. ; Murphy, J. M. ; Noguer, M.: Simulation of Climate Change over Europe using a nested regional climate model; I: Assessment of control climate, including sensitivity to location of lateral boundaries. In: *Q. J. R. Meteorol. Soc.* (1995), Nr. 121, S. 1413–1449

JUNG 2007
JUNG: http://jung.sourceforge.net. In: *Internet page*, 2007

Jung 1996
Jung, V.: A System for Guiding and Training Users in the Visualization of Geographic Data. In: *Proceedings of the 1st Conference of GeoComputation. Leeds, UK*, 1996, S. 470– 482

Jung 1998
Jung, V.: *Integrierte Benutzerunterstützung für die Visualisierung in Geo-Informationssystemen*, TU Darmstadt, Fachbereich Informatik, Fachgebiet GRIS, Fraunhofer IRB Verlag, Stuttgard, Diss., 1998

Kaeding 2006
Kaeding, R.: *Vergleichende Visualisierung von Daten auf variierenden Gittern am Beispiel von Klimadaten*, Diplomarbeit, Universität Rostock, Institut für Informatik, Diplomarbeit, 2006

Kaeding u. Walter 2004
Kaeding, R. ; Walter, M.: *Dokumentation 'Kombinierte Kalender - Cluster - Plot Visualisierung'.* Praktikumsarbeit (KSWS), Universität Rostock, Institut für Informatik, 2004

Kalvin u. a. 2000
Kalvin, A.D. ; Pelah, A. ; Cohen, A. ; Rogowitz, B.E.: Building Perceptual Color Maps for Visualizing Interval data. In: *Proceedings SPIE Conference on Human Vision and Electronic Imaging, San Jose, CA*, 2000

Kamps 1999
Kamps, T.: *Diagram Design. A Constructive Theory.* Springer Verlag, 1999

Kanitsar u. a. 2001
KANITSAR, A. ; WEGENKITTL, R. ; FELKEL, P. ; FLEISCHMANN, D. ; SANDNER, D. ; GROELLER, E.: Computed Tomography Angiography: A Case Study of Peripheral Vessel Investigation. In: *Proc. Visualization (Vis'01)*, 2001, S. 477–480

Kao u. Shen 1999
KAO, D. ; SHEN, H.W.: Automatic Surface Flow Feature Visualization. In: *14th AIAA Computational Fluid Dynamics Conference* (1999)

Keahey 2000
KEAHEY, T. A.: *The Nonlinear Magnification Home Page*. 2000

Keim u. a. 1995
KEIM, D. ; ANKERST, M. ; KRIEGEL, H.P.: Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data. In: *Proceedings IEEE Visualization'95, USA*, 1995, S. 279–287

Keim u. a. 2005
KEIM, D. (Hrsg.) ; KOHLHAMMER, J. (Hrsg.) ; THOMAS, J. (Hrsg.): *Workshop Visual Analytics, Closing Panel: The top 10 Visual Analytics Research Challenges, Darmstadt.* 2005

Keim u. a. 1993
KEIM, D. A. ; KRIEGEL, H.-P. ; SEIDL, T.: Visual feedback in querying large database. In: *Proceedings IEEE Visualization '93*, 1993, S. 158–165

Keim u. Ward 2002
KEIM, D. A. ; WARD, M.: Visual Data Mining. In: BERTHOLD, M. (Hrsg.) ; HAND, D. J. (Hrsg.): *Intelligent Data Analysis - An Introduction*. 2. Harlow : Springer-Verlag, 2002

Keim 2002
KEIM, D.A.: Information visualization and visual data mining. In: *IEEE Transactions on Visualization and Computer Graphics* 8 (2002), Nr. 1, S. 1–8

Keim u. Kriegel 1996
KEIM, D.A. ; KRIEGEL, H.-P.: Visualization Techniques for Mining Large Databases: A Comparison. In: *IEEE Transactions on Knowledge and Data Engineering* 8 (1996), December, Nr. 6, S. 923–938

Keim u. a. 2002
KEIM, D.A. ; MÜLLER, W. ; SCHUMANN, H.: Information Visualization and Visual Data Mining; State of the art report. In: *Proceedings Eurographics 2002, Saarbrücken*, 2002

Keller u. Keller 1993
KELLER, P.R. ; KELLER, M.M.: *Visual Cues. Practical Data Visualization*. IEEE Computer Society Press, Los Alamitos, 1993

Kim u. Pang 1997
KIM, K. ; PANG, A.: Projection-based data level comparisons of direct volume rendering algoriths. In: *Kwansik Kim and Alex Pang. Projection-based data level comparisons of direct volume rendering algoriths. Technical Report UCSC-CRL-97-16, University of California at Santa Cruz, 1997.* (1997)

Kindlmann u. a. 2002
KINDLMANN, G. ; REINGARD, E. ; CREEM, S.: Face-based Luminance Matching for Perceptual Colormap Generation. In: *Proc. IEEE Information Visualization 2002, IEEE Press*, 2002

Klembt u. Krüger 2006
KLEMBT, A. ; KRÜGER, U.: *Intelligente Anordnungsverfahren für Visualisierungstechniken.* Praktikumsarbeit (KSWS), Universität Rostock, Institut für Informatik, 2006

Klimt 2006
KLIMT: http://stats.math.uni-augsburg.de/Klimt. In: *Internet page*, 2006

Kobsa 2001
KOBSA, A.: An Empirical Comparison of Three Commercial Information Systems. In: *IEEE Syposium on Information Visualization (InfoVis'01)*, 2001, S. 11–16

Kohonen 1997
KOHONEN, T.: *Self-organizing maps.* 2. Springer, 1997

Komura u. a. 2004
KOMURA, D. ; NAKAMURA, H. ; TSUTSUMI, S. ; ABURATANI, H. ; IHARA, S.: Multidimensional support vector machines for visualization of gene expression data. In: *Proc. ACM symposium on Applied computing, Nicosia, Cyprus*, 2004, S. 175–179

Koolwaaij u. van Leeuwen 2003
KOOLWAAIJ, J. ; LEEUWEN, P. Fennema D.: SVG for Process Visualization. In: *Online-Referenz: http://www.svgopen.org/2003/papers/ProcessVisualisation* (2003)

Kosara u. Hauser 2002
KOSARA, R. ; HAUSER, H.: The State of the Art in Information Visualization / VRVis Research Center. Vienna, Austria, December 2002 (Technical Report TR-VRVis-2002-043). – Forschungsbericht

Kosara u. a. 2003
KOSARA, R. ; HEALEY, C.G. ; INTERRANTE, V. ; LAIDLAW, D. ; WARE, C.: User Studies: Why, How, and When? In: *IEEE Computer Graphics and Applications* (2003), July/August

Kottek u. Rubel 2003
KOTTEK, M. ; RUBEL, F.: Globale Klimadaten in standardisierter Darstellung. In: *Wiener Meteorologische Schriften* Heft 1 (2003)

Kreuseler 2004
KREUSELER, M.: *Ein flexibles Framework zum Visuellen Data Mining*, Universität Rostock, Institut für Informatik, Diss., 2004

Kreuseler u. a. 2000
KREUSELER, M. ; LOPEZ, N. ; SCHUMANN, H.: A Scalable Framework for Information Visualization. In: *Proceedings of IEEE Information Visualization (InfoVis'00); Salt Lake City; Utah* (2000)

Kreuseler u. a. 2003
KREUSELER, M. ; NOCKE, T. ; SCHUMANN, H.: Integration of Clustering and Visualization Techniques for Visual Data Analysis. In: OPITZ, O. (Hrsg.) ; SCHWAIGER, M. (Hrsg.): *25th Annual Conference of the Gesellschaft für Klassifikation e.V., published in 'Exploratory Data Analysis in Empirical Research'* University of Munich, Springer-Verlag, Heidelberg-Berlin, March 2003, S. 119–132

Kreuseler u. a. 2004
KREUSELER, M. ; NOCKE, T. ; SCHUMANN, H.: A History Mechanism for Visual Data Mining (InfoVis'04), Austin. In: *IEEE Symposium on Information Visualization*, 2004, S. 49–56

Kreuseler u. Schuman 1999
KREUSELER, M. ; SCHUMAN, H.: Information visuallzation using a new Focus + Context Technique in combination with dynamic clustering of information space. In: *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation (NPIVM-99)*. N.Y. : ACM Press, November 6 1999, S. 1–5

Kreuseler u. Schumann 2002a
KREUSELER, M. ; SCHUMANN, H.: A Flexible Approach for Visual Data Mining. In: *IEEE Transactions on Visualization and Computer Graphics* 8 (2002), Nr. 1

Kreuseler u. Schumann 2002b
KREUSELER, M. ; SCHUMANN, H.: A Flexible Approach for Visual Data Mining. In: *IEEE Transactions on Visualization and Computer Graphics* 8 (2002), January-March, Nr. 1

Kruskal u. Wish 1978
KRUSKAL, J.N. ; WISH, M.: *Multidimensional Scaling*. Sage, 1978

Kücken u. a. 2002
KÜCKEN, M. ; GERSTENGARBE, F.-W. ; WERNER, P.C.: Cluster analysis results of regional climate model simulations in the PIDCAP period. In: *Boreal Environment Research* 7 (2002), Nr. 3

Kücken u. a. 1999
KÜCKEN, M. ; SCHÄTTLER, U. ; GERSTENGARBE, F.-W. ; WERNER, P.: Simulation and Visualization of Climate Scenarios on a Distributed Memory Platform. In: ENGQUIST, B. (Hrsg.) ; JOHNSSON, L. (Hrsg.) ; HAMMILL, M. (Hrsg.) ; SHORT, F. (Hrsg.): *Simulation and Visualization on the Grid*. Springer Verlag, Berlin, 1999, S. 242–253

Lamping u. a. 1996
LAMPING, J. ; RAO, R. ; PIROLI, P.: A Focus-Context-Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In: *Proceedings CHI'96, Vancouver, Kanada* (1996)

Lamping u. a. 1995
LAMPING, J. ; RAO, R. ; PIROLLI, P.: A Focus + Context Technique based on Hyperbolic Geometry for Visualising Large Hierarchies. In: *Proceedings of CHI '95 (International Conference on Human Factors in Computing Systems)*, ACM Press, 1995

Landgrebe u. a. 2002
LANDGREBE, J. ; WURST, W. ; WELZL, G.: Permutation-validated principal components analysis of microarray data. In: *Genome Biology* (2002), Nr. 3, research 0019.1-0019.11

Lange u. a. 2006
LANGE, S. ; NOCKE, T. ; SCHUMANN, H.: Visualisierungsdesign - ein systematischer Überblick. In: *Simulation and Visualization (SimVis'06)*, 2006, S. 113–128

Lange 2006
LANGE, Susanne: *Konzeption einer Visualisierungshilfe für multivariate Daten*, Universität Rostock, Institut für Informatik, Diss., 2006

Laramee u. a. 2004
LARAMEE, R.S. ; WEISKOPF, D. ; SCHNEIDER, J. ; HAUSER, H.: Investigating Swirl and Tumble Flow with a Comparison of Visualization Techniques. In: *Proceedings IEEE Visualization'04, Austin, USA*, 2004

Laudien 2000
LAUDIEN, C.: *Ein Meta-Lerner zur automatishen Auswahl von Data-Mining-Verfahren*, Diplom Thesis, University of Rostock, Institute of Computer Science, Diplomarbeit, 2000

Lodha u. a. 1996
LODHA, S.K. ; A.: UFLOW: Visualizing Uncertainty in Fluid Flow. In: *Proceedings Visualization '96, IEEE Computer Society Press, Los Alamitos* (1996), S. 249–254

Lux 1998
LUX, M.: Level of Data - A Concept for Knowledge Discovery in Information Spaces. In: *Proc. Information Visualization (IV'98), London* (1998)

Ma u. Smith 1993
MA, K.-L. ; SMITH, P.J.: Cloud Tracing in Convection-Diffusion System. In: *Proceedings IEEE Visualization'93, USA*, 1993, S. 253ff

MacEachren 1994
MACEACHREN, A. M.: *Some Truth With Maps : A Primer on Symbolization and Design.* Association of American Geographers, 1994 (Resource Publications in Geography)

Mackinlay u. a. 1991
MACKINLAY, J.D. ; ROBERTSON, G.G. ; CARD, S.K.: The Perspective Wall: Detail and COntext Smoothly Integrated. In: *Proc. ACM Conference on Human Factors in Computing Systems (HCI'91)*, 1991, S. 173–180

Mackinlay 1986
MACKINLAY, Jock: Automating the design of graphical presentations of relational information. In: *ACM Transactions on Graphics* 5 (1986), Nr. 2, S. 110–141. – ISSN 0730–0301

Macêdo u. a. 2000
MACÊDO, M. ; COOK, D. ; BROWN, T. J.: Visual Data Mining In Atmospheric Science Data. In: *Data Mining and Knowledge Discovery* 4 (2000), Nr. 1, S. 69–80

Mann u. a. 1998
MANN, M.E. ; BRADLY, S.R. ; MALCOLM, K.: Global-Scale Temperature Patterns and Climate Forcing Over the Past Six Centuries. In: *Nature* (1998), Nr. 392, S. 779ff

Mann u. Rockwood 2002
MANN, S. ; ROCKWOOD, A.: Computing Singularities of 3D Vector Fields with Geometric Algebra. In: *Proc. IEEE Information Visualization 2002, IEEE Press*, 2002

Mao u. a. 2000
MAO, X. ; HATANAKA, Y. ; IMAMIYA, A.: Visualizing Computational Wear With Physical Wear. In: *6th ERCIM Workshop 'User Interfaces for All'*, 2000, S. 12–23

Marchesin u. a. 2004
MARCHESIN, S. ; DISCHLER, S. ; MONGENET, C.: 3D ROAM for Scalable Volume Visualization. In: SILVER, D. (Hrsg.) ; ERTL, T. (Hrsg.) ; SILVA, C. (Hrsg.): *Proceedings of IEEE/SIGGRAPH Symposium on Volume Visualization*, 2004

Matkovic u. a. 2002
MATKOVIC, K. ; HAUSER, H. ; SAINITZER, R. ; GRÖLLER, E.: Real Time Simulation and Visualization using Pre-Calculated Fluid Simulator States. In: *Proceedings IEEE Information Visualization (InfoVis'02), USA*, 2002, S. 67–70

Max u. Crawfis 1995
MAX, N. ; CRAWFIS, R: Advances in Scientific Visualization. In: *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology* Bd. 2410, 1995, S. 340–345

McEachren 1994
MCEACHREN, Alan M.: *Some Truth With Maps: A Primer on Symbolization and Design.* Assn of Amer Geographers, 1994 (Resource Publications in Geography)

McEachren 1995
MCEACHREN, Alan M.: *How Maps Work: Presentaion, Visualization, and Design.* 72 Spring Street, New York : The Guilford Press, 1995

McGrenere u. a. 2002
MCGRENERE, J. ; BAECKER, R. ; BOOTH, K.: An evaluation of a multiple interface design solution for bloated software. In: *Proceedings of the CHI'02, Minneapolis, USA*, 2002, S. 20–24

McKenna u. Arce 2000
MCKENNA, T. ; ARCE, G. R.: New image mosaic structures / Department of Electrical and Computer Engineering, University of Delaware. 2000. – Forschungsbericht

McKitrick u. McIntyre 2005
MCKITRICK, R. ; MCINTYRE, S.: Hockey Sticks, Principal Components and Spurious Significance. In: *Geophysical Research Letters* 32 (2005), Nr. 3

Moorhead u. Zhu 1993
MOORHEAD, R.J. ; ZHU, Z.: Feature Extraction for Oceanographic Data using a 3D Edge Operator. In: *Proceedings IEEE Visualization'93, USA*, 1993, S. 402ff

Mullerworth 2004
MULLERWORTH, S.: The PRISM Data Processing and Visualisation System / PRISM project page: http://prism.enes.org/Results/Documents/. 2004 (PRISM-Report Series-15). – Forschungsbericht

Müller u. Alexa 2004
MÜLLER, W. ; ALEXA, M.: Visual Component Analysis. In: *Proc. of Joint IEEE/EG Symposium on Visualization, VisSym'04, Konstanz, Germany*, 2004, S. 129–136

Müller u. a. 2006
MÜLLER, W. ; NOCKE, T. ; H.SCHUMANN: Enhancing the Visualization Process with Principal Component Analysis to Support the Exploration of Trends. In: *APVis'06*, 2006

Müller u. Schumann 2003
MÜLLER, W. ; SCHUMANN, H.: Visualization Methods for Time-dependent Data - an Overview. In: CHICK, S. (Hrsg.) ; SÁNCHEZ, P. J. (Hrsg.) ; FERRIN, D. (Hrsg.) ; MORRICE, D. J. (Hrsg.): *Proceedings of the 2003 Winter Simulation Conference, New Orleans*, 2003

Nemcsics 1993
NEMCSICS, A.: *Farbenlehre und Farbendynamik - Theorie der farbigen Umweltplanung.* Muster-Schmidt Verlag, Göttingen, Zürich, 1993

Nesbitt 2004
NESBITT, K.V.: Getting to more Abstract Places using the Metro Map Metaphor. In: *Information Visualization (IV'04), London*, 2004, S. 488–493

Nguyen u. Huang 2002
NGUYEN, O.V. ; HUANG, M.L.: A Space-Optimized Tree Visualization. In: WONG, Pak C. (Hrsg.) ; ANDREWS, Keith (Hrsg.): *Proc. IEEE Symp. Information Visualization, InfoVis'02*, IEEE Computer Society, 28–29 Oktober 2002, S. 85–92

Nocke 1999
NOCKE, T.: *Konzeption und Realisierung einer flexiblen Pipeline zur numerischen Vorverarbeitung in der Informationsvisualisierung.* Studienarbeit, Universität Rostock, Fachbereich Informatik, 1999

Nocke 2000
NOCKE, T.: *Metadatengewinnung und -spezifikation für Visualisierungsentscheidungen*, Universität Rostock, Fachbereich Informatik, Diplomarbeit, 2000

Nocke u. a. 2007
NOCKE, T. ; FLECHSIG, M. ; BÖHM, U.: Visual Exploration and Evaluation of Climate-Related Simulation Data. In: *Winter Simulation Conference (WSC'07), Washington D.C., USA*, 2007

Nocke u. a. 2005
NOCKE, T. ; SCHLECHTWEG, S. ; SCHUMANN, H.: Icon-based Visualization using Mosaic Metaphors. In: *Information Visualization (IV'05), London*, 2005, S. 103–109

Nocke u. Schumann 2002
NOCKE, T. ; SCHUMANN, H.: Meta Data for Visual Data Mining. In: *Proceedings Computer Graphics and Imaging (CGIM'02), Kaua'i, Hawaii, USA*, 2002

Nocke u. Schumann 2004
NOCKE, T. ; SCHUMANN, H.: Goals of Analysis for Visualization and Visual Data Mining Tasks. In: *CODATA Workshop Information, Presentation and Design, Prague*, 2004

Nocke u. a. 2004
NOCKE, T. ; SCHUMANN, H. ; BOEHM, U.: Methods for the Visualization of Clustered Climate Data. In: *Computational Statistics* 19 (2004), Nr. 1, S. 75–94

Nocke u. a. 2003
NOCKE, T. ; SCHUMANN, H. ; BÖHM, U. ; FLECHSIG, M.: Information Visualization supporting Modelling and Evaluation Tasks for Climate Models. In: CHICK, S. (Hrsg.) ; SÁNCHEZ, P. J. (Hrsg.) ; FERRIN, D. (Hrsg.) ; MORRICE, D. J. (Hrsg.): *Proceedings of the 2003 Winter Simulation Conference, New Orleans*, 2003

North u. a. 2002
NORTH, C. ; CONKLIN, N. ; SAINI, V.: Visualization Schemas for Flexible Information Visualization. In: *IEEE Syposium on Information Visualization (InfoVis'02)*, 2002

Nowell u. a. 2002
NOWELL, L. ; SCHULMAN, R. ; HIX, D.: Graphical Encoding for Information Visualization: An Empirical Study. In: *IEEE Syposium on Information Visualization (InfoVis'02)*, 2002

Ohl 2005
OHL, S.: *Dokumentation zum Ikonenmodul.* Praktikumsarbeit (KSWS), Universität Rostock, Institut für Informatik, 2005

de Oliveira u. Levkowitz 2003
OLIVEIRA, M.C.F. de ; LEVKOWITZ, H.: From Visualization to Visual Data Mining: A Survey. In: *IEEE Transactions on Visualization and Computer Graphics* 9 (2003), Nr. 3, S. 378–394

Pagendarm u. Post 1995
PAGENDARM, H.-G. ; POST, F. H.: *Comparative Visualization - Approaches and Examples.* In Visualization in Scientific Computing, M. Göbel, H. Müller and B. Urban (Ed.), Springer, Wien, 1995

Pagendarm u. Seitz 1993
PAGENDARM, H.-G. ; SEITZ, B.: *An Algorithm for Detection and Visualization of Discontinuities in Scientific Data Fields Applied to Flow Data with Shock Waves.* In: Scientific Visualization -Advanced Software Techniques, P. Paramidese (Ed.), Ellis Horwood Ltd., 1993

Pagendarm u. Walter 1995
PAGENDARM, H.-G. ; WALTER, B.: Competent, Compact, Comparative Visualization of a Vortical Flow Field. In: *IEEE Transactions on Visualization and Computer Graphics* 1 (1995), June, Nr. 2, S. 142–150

Pajek 2007
PAJEK: http://vlado.fmf.uni-lj.si/pub/networks/pajek. In: *Internet page*, 2007

Pang u. Freeman 1996
PANG, A. ; FREEMAN, A.: Methods for comparing 3D surface attributes. In: *Proc. SPIE, Visual Data Exploration and Analysis III, Georges G. Grinstein; Robert F. Erbacher; Eds.* Bd. 2656, 1996, S. 58–64

Pang u. a. 1997
PANG, A. ; WITTENBRINK, C.M. ; LODHA, S.K.: Approaches to Uncertainty Visualization. In: *The Visual Computer* 13 (1997), Nr. 8, S. 370–390

Pedlosky 1987
PEDLOSKY, J.: *Geophysical Fluid Dynamics.* Springer, Berlin, 1987

Petoukhov u. a. 2000
PETOUKHOV, V. ; A.GANOPOLSKI ; V.BROVKIN ; CLAUSSEN, M. ; ELISEEV, A. ; KUBATZKI, C. ; RAHMSTORF, S.: CLIMBER-2: a climate system model of intermediate complexity. Part I: model description and performance for present climate. In: *Climate Dynamics (2000), Springer Verlag* 16 (2000)

Pfefferer 1996
PFEFFERER, L.: *Objektzentrierte Visualisierung mehrdimensionaler Daten als Erweiterung konventioneller Datenbankmodelle*, Universität München, Diss., 1996

Piccolo 2006
PICCOLO: http://www.cs.umd.edu/hcil/piccolo. In: *Internet page*, 2006

Polaris 2007
POLARIS: http://graphics.stanford.edu/projects/polaris. In: *Internet page*, 2007

Post u. a. 2003
POST, F. H. ; VROLIJK, B. ; HAUSER, H. ; LARAMEE, R. S. ; DOLEISCH, H.: The State of the Art in Flow Visualization: Feature Extraction and Tracking. In: *Computer Graphics Forum* 22 (2003), Nr. 4, S. 775–792

Prefuse 2007
PREFUSE: http://prefuse.sourceforge.net. In: *Internet page*, 2007

Rahmstorf u. Ganopolski 1999
RAHMSTORF, S. ; GANOPOLSKI, A.: Simple theoretical model may explain apparent climate instability. In: *Journal of Climate* 12 (1999), S. 1349–1352

Rao u. Card 1994
RAO, R. ; CARD, S. K.: The Table Lens: Merging Graphical and Symbolical Representations in an Interactive Focus+Context Visualization for Tabular Information. In: *Proc. ACM Conference on Human Factors in Computing Systems (HCI'94)*, 1994

Reinders u. a. 2001
REINDERS, F. ; POST, F.H. ; H.J.W.SPOELDER: Visualization of Time-Dependent Data using Feature Tracking and Event Detection. In: *The Visual Computer* 17(1) (2001), February, S. 55–71. – ISSN 0178–2789

Reinders 2001
REINDERS, K.F.J.: *Feature-Based Visualization of Time-Dependent Data*, Delft University of Technology, Diss., March 2001

Rew u. a. 1993
REW, R. ; DAVIS, G. ; EMMERSON, S.: *NetCDF user's Guide. An Interface of Data Access.* UCAR, 1993

Ribarsky u. a. 2002a
RIBARSKY, W. ; FAUST, N. ; WARTELL, Z. ; SHAW, C. ; JANG, J.: Visual Query Of Time-Dependent 3D Weather In A Global Geospatial Environment. In: LADNER, R. (Hrsg.) ; SHAW, K. (Hrsg.) ; ABDELGUERFI, M. (Hrsg.): *Mining Spatio-Temporal Information Systems.* Kluwer, Amsterdam, 2002

Ribarsky u. a. 1999
RIBARSKY, W. ; KATZ, J. ; HOLLAND, A.: Discovery Visualization Using Fast Clustering. In: *IEEE Computer Graphics and Applications* 19 (1999), Nr. 5, S. 32–39

Ribarsky u. a. 2002b
RIBARSKY, W. ; SHAW, C. ; WARTELL, Z. ; FAUST, N.: Building the Visual Earth. In: *SPIE 16th International Conference on Aerospace/Defense Sensing, Simulation, and Controls*, 2002

Riley u. a. 2003
RILEY, K. ; EBERT, D. ; HANSEN, C. ; LEVIT, J.: Visually Accurate Multi-Field Weather Visualization, 2003

Robertson u. a. 1991
ROBERTSON, G. ; MACKINLAY, J. ; CARD, S.: Cone trees: Animated 3d visualization of hierarchical information. In: *ACM Proceedings of Computer-Human Interaction (CHI'91)*, 1991, S. 189–194

Robertson 2006
ROBERTSON, J. (Hrsg.): *Proceedings of Multiple Views.* London, UK, 2006

Robertson 1990
ROBERTSON, P. K.: A Methodology for Scientific Data Visualisation: Chosing Representations Based on a Natural Scene Paradigm. In: *Procedings Visualisation '90, IEEE Computer Society Press, Los Alimatos* (1990), S. S. 114–123

Robertson u. Hutchins 1997
ROBERTSON, P. K. ; HUTCHINS, M.A.: An Approach to Intelligent Design of Color Visualisation. In: *In: G. Nielson, H. Hagen, H. Müller: Scientific Visualisation. IEEE Computer Society, Los Alamitos* (1997), S. S. 179–190

Ropinski u. Hinrichs 2004
ROPINSKI, T. ; HINRICHS, K.: Real-Time Rendering of 3D Magic Lenses having arbitrary convex Shapes. In: *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'04), Pilsen, CZ*, 2004

Rosario u. a. 2004
ROSARIO, G. u. a.: Mapping Nominal Values to Numbers for Effective Visualizations. In: *Palgrave Information Visualization* 3 (2004), Juni, Nr. 2, S. 80–95

Ross u. Chalmers 2003
ROSS, G. ; CHALMERS, M.A.: Visual Workspace for Hybrid Multidimensional Scaling Algorithms. In: *Proceedings IEEE Symposium on Information Visualization 2003 (InfoVis'03)*. Seattle, Washington, 2003, S. 247–257

Ross u. a. 1997
ROSS, T.F. ; MANNS, D. ; FAAS, W.M.: Climvis - A Cool Way To Visualize Noaa's Climate Data. In: *Preprints. Sixth Symposium on education*, 1997

Roth u. a. 1996
ROTH, S. A. ; LUCAS, P. ; SENN, J. A. ; GOMBERG, C. C. ; BURKS, M. B. ; STROFFOLINO, P. J. ; KOLOJEJCHICK, J. A. ; DUNMIRE, C.: Visage: A User Interface Environment for Exploring Information. In: *IEEE Symposium on Information Visualization (InfoVis'96), San Francisco*, 1996, S. 3–12

Roth u. Mattis 1991
ROTH, S. F. ; MATTIS, J.: Automating the Presentation of Information. In: *Proceedings of the IEEE Conference on AI Applications*, 1991

Roth u. Mattis 1990
ROTH, S.F. ; MATTIS, J.: Data Characterization for Intelligent Graphic Presentation. In: *Proceedings of the Computer-Human Interaction Conference (CHI '90)*, 1990, S. 193–200

Rushmeier u. a. 1997
RUSHMEIER, H. u. a.: Perceptual Measures for Effective Visualizations. In: *Proceedings IEEE Visualization'97, Phoenix, USA*, 1997

Rushmeier u. a. 1995
RUSHMEIER, H. ; WARD, G. ; PIATKO, C. ; SANDERS, P. ; RUST, B.: Comparing Real and Synthetic Images: Some Ideas About Metrics. In: *Proceedings of Sixth Eurographics Workshop on Rendering, Dublin, Ireland,*, 1995, S. 82–91

Saito u. a. 2005
SAITO, T. ; MIYAMURA, H.N. ; YAMAMOTO, M. ; SAITO, H. ; HOSHIYA, Y.: Two-Tone Pseudo Coloring - Compact Visualization for One-DImensional Data. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05)*. Mineapolis, USA, 2005, S. 173–180

dos Santos u. Brodlie 2004
SANTOS, S. dos ; BRODLIE, K.: Gaining Understanding of Multivariate and Multidimensional Data through Visualization. In: *Computers & Graphics* 18 (2004), Nr. 3

Scanlon 1994
SCANLON, C.H.: Cockpit graphical weather information shown to enhance efficiency, safety, and situational awareness. In: *Proceedings of the 39th Annual Corporate Aviation Safety Seminar (CASS)*, Flight Safety Foundation, 1994, S. 83–94

Schmidt u. a. 2004
SCHMIDT, H. ; BRASSEUR, G.P. u. a.: Mesospheric Dynamics, Energetics and Chemistry (MEDEC). 3 (2004), Nr. 7

Schmidt 2004
SCHMIDT, M.: *Darstellung und Verwaltung von Operatorgraphen und Analyseinformationen für das visuelle Data Mining in Klimadaten.* Studienarbeit, Universität Rostock, Institut für Informatik, 2004

Scholtz 2006
SCHOLTZ, D.: *Einsatz von Kohonen-Karten für Klimadaten.* Praktikumsarbeit (KSWS), Universität Rostock, Institut für Informatik, 2006

Schröder 1997
SCHRÖDER, F.: *Visualisierung meteorologischer Daten.* Springer, 1997

Schröder u. Wagenknecht 2003
SCHRÖDER, M. ; WAGENKNECHT, K.: *Dokumentation zum SimEnvVis - Projekt & dem KSWS, Version 0.5.* Praktikumsarbeit (KSWS), Universität Rostock, Institut für Informatik, 2003

Schulz 2005
SCHULZ, H.-J., Universität Rostock, Fakultät für Informatik und Elektrotechnik, Diplomarbeit, 2005

Schulz u. a. 2006a
SCHULZ, H.-J. ; NOCKE, T. ; SCHUMANN, H.: A Framework for Visual Data Mining of Structures. In: *29th Australasian Computer Science Conference (ACSC'06), Hobart, Australia*, 2006

Schulz u. a. 2006b
SCHULZ, H.-J. ; SCHUMANN, H. ; NOCKE, T.: Visualizing and Analyzing Large Systems of Differential Equations. In: *Poster Presentation at the Winter Simulation Conference (WSC'06) in Monterey, U.S.A.*, 2006

Schulze-Wollgast u. a. 2005
SCHULZE-WOLLGAST, P. ; TOMINSKI, C. ; SCHUMANN, H.: Enhancing Visual Exploration by Appropriate Color Coding. In: *International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media (WSCG'05).* Plzen, Czech Republic, 2005

Schumann u. Müller 2000
SCHUMANN, H. ; MÜLLER, W.: *Visualisierung, Grundlagen und allgemeine Methoden.* 1. Berlin Heidelberg : Springer-Verlag, 2000. – ISBN ISBN 3–540–64944–1

Senay u. Ignatius 1994
SENAY, H. ; IGNATIUS, E.: A Knowledge-Based System for Visualization Design. In: *IEEE Computer Graphics & Applications* 14 (1994), Nr. 6, S. 36–47

Shen u. Eades 2004
SHEN, X. ; EADES, P.: Using MoneyTree to Represent Financial Data. In: *Proceedings of IEEE Information Visualization, IV'04, London*, 2004, S. 285–289

Shneiderman 1992
SHNEIDERMAN, B.: Tree Visualization with Treemaps: A 2D Space Filling Approach. In: *ACM Transactions on Graphics* 11 (1992), Nr. 1, S. 92–99

Shneiderman 1996
SHNEIDERMAN, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. In: *Proceedings IEEE Symposium on Visual Languages '96, IEEE, Los Alamos*, 1996

Shneiderman 2002
SHNEIDERMAN, B.: Inventing discovery tools: combining information visualization with data mining. In: *Palgrave Information Visualization* 1 (2002), Nr. 1, S. 5–12

Silver 1997
SILVER, D.: Feature Visualisation. In: *In: G. Nielson, H. Hagen, H. Müller: Scientific Visualisation. IEEE Computer Society, Los Alamitos* (1997), S. 279–293

Silver u. a. 2004
SILVER, D. (Hrsg.) ; ERTL, T. (Hrsg.) ; SILVA, C. (Hrsg.): *Proceedings of IEEE/SIGGRAPH Symposium on Volume Visualization*. Austin, USA, 2004

Silver u. Zabusky 1993
SILVER, D. ; ZABUSKY, N.J.: Quantifying Visualizations for Reduced Modelling in Nonlinear Science: Extracting Structures from Datasets. In: *Journal of Visual Communication and Image Representation* 4 (1993), March, Nr. 1, S. 44–61

SimVis 2007
SIMVIS: http://www.vrvis.at/simvis. In: *Internet page*, 2007

Singh u. Silver 2004
SINGH, V. ; SILVER, D.: Interactive Volume Manipulation with Selective Rendering for Improved Visualization. In: SILVER, D. (Hrsg.) ; ERTL, T. (Hrsg.) ; SILVA, C. (Hrsg.): *Proceedings of IEEE/SIGGRAPH Symposium on Volume Visualization*, 2004

Spence 2001
SPENCE, R.: *Information Visualization*. Harlow : Addison-Wesley, 2001

Spirkovska u. Lodha 2002
SPIRKOVSKA, L. ; LODHA, S.K.: Audio-Visual Situational Awareness for General Aviation Pilots. In: *Proceedings Eurographics'02*, 2002

Spotfire 2007
SPOTFIRE: Spotfire DecisionSite: http://www.spotfire.com. In: *Internet page*, 2007

SPSS 2007
SPSS: http://www.spss.com/. In: *Internet page*, 2007

Stasko u. a. 1997
STASKO, J. (Hrsg.) ; DOMINGUL, J. (Hrsg.) ; BROWN, M.H. (Hrsg.) ; PRICE, B.A. (Hrsg.): *Software Visualization*. MIT Press, 1997

Stasko u. a. 2004
STASKO, J. ; MILLER, T. ; PLAUE, C. ; POUSMAN, Z. ; ULLAH, O.: Personalized Peripheral Information Awareness through Information Art. In: *Proceedings of the International Conference on Ubiquitous Computing*, 2004, S. 18–35

Stasko u. Zhang 2000
STASKO, J. T. ; ZHANG, E.: Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations. In: *IEEE Symposium on Information Visualization '00*, 57-

Steinacker u. a. 2001
STEINACKER, A. ; GHAVAM, A. ; STEINMETZ, R.: *Metadata Standards for Web-Based Resources.* 2001

Stier u. a. 2005
STIER, P. ; FEICHTER, J. ; KINNE, S. ; KLOSTER, S. ; VIGNATI, E. ; WILSON, J. ; GANZEVELD, L. ; TEGEN, I. ; WERNER, M. ; BALKANSKI, Y. ; SCHULZ, M. ; BOUCHER, O. ; MINIKIN, A. ; PETZOLD, A.: The aerosol-climate model ECHAM5-HAM. In: *Atmosheric Chemistry and Physics, European Geosciences Union* 5 (2005), S. 1125–1156

Strothotte u. a. 1999
STROTHOTTE, T. ; MASUCH, M. ; ISENBERG, T.: Visualizing Knowledge about Virtual Reconstructions of Ancient Architecture. In: *Proceedings Computer Graphics International, IEEE Computer Society*, 1999, S. 36–43

Strothotte u. Schlechtweg 2002
STROTHOTTE, T. ; SCHLECHTWEG, S.: *Non-Photorealistic Computer Graphics: Modeling, Rendering, and Animation.* Morgan Kaufmann, 2002

Struck u. Marczok 2003
STRUCK, T. ; MARCZOK, H.: *Dokumentation CalViz (Kombinierte Kalender - Cluster - Plot Visualisierung).* Praktikumsarbeit (KSWS), Universität Rostock, Institut für Informatik, 2003

Tang u. a. 2004
TANG, D. ; STOLTE, C. ; BOSCH, R.: Design choices when architecting visualizations. In: *Information Visualization* 3 (2004), Nr. 2, S. 65–79. – ISSN 1473–8716

Taylor u. Bendford 1998
TAYLOR, I.M. ; BENDFORD, S.D.: Data Containers for Relation Based Visualization. In: *IEEE Information Visualization (IV'98)* (1998)

TeCoMed 1998
TECOMED: http://wwwicg.informatik.uni-rostock.de/Projekte/TeCoMed. In: *Internet page*, 1998

Theisel 1994
THEISEL, H.: *Automatische Auswahl geeigneter Visualisierungstechniken für allgemeine wissenschaftliche Datensätze.* Diplomarbeit, Universität Rostock, Fachbereich Informatik, 1994

Theisel 1995
THEISEL, H.: Analyse und Visualisierungshilfe für mehrdimensionale wissenschaftliche Daten. In: *Informatik, Forschung und Entwicklung (in german)* 10 (1995), Oktober, S. S. 91–98

Theisel u. Kreuseler 1998
THEISEL, H. ; KREUSELER, M.: An Enhanced Spring Model for Information Visualisation. In: *Computer Graphics Forum, Scientific Visualisation (Proceedings Eurographics '98)* Bd. 17, 1998

Thiede u. Krüger 2004
THIEDE, C. ; KRÜGER, A.: *Design und Umsetzung einer auf Quadtrees basierenden Baum-Datenstruktur zur Verwaltung von Höhenkarten.* Praktikumsarbeit (KSWS), Universität Rostock, Institut für Informatik, 2004

Thomas 2005
THOMAS, J.: Visual Analytics: a Grand Challenge in Science - Turning Information Overload into the Opportunity of the Decade. In: *IEEE Syposium on Information Visualization (InfoVis'05)*, 2005

Thurston u. a. 2003
THURSTON, J. ; LINDNER, M. ; LASCH, P.: The "Brandenburg Study": Visualizing the regional impacts of climate change on forests in the Federal state of Brandenburg, Germany. In: *gg*, 2003

TimeSearcher 2007
TIMESEARCHER: http://www.cs.umd.edu/hcil/timesearcher. In: *Internet page*, 2007

Todorovski u. Dzeroski 2003
TODOROVSKI, L. ; DZEROSKI, S.: Combining Classifiers with Meta Decision Trees. In: *Machine Learning* (2003)

Tollis u. a. 1999
TOLLIS, I. ; EADES, P. ; BATTISTA, G. di: *Graph Drawing - Algorithms for the Visualization of Graphs.* Prentice Hall, 1999

Tominski u. a. 2004
TOMINSKI, C. ; ABELLO, J. ; SCHUMANN, H.: Axes-based visualizations with radial layouts. In: *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing.* New York, NY, USA : ACM Press, 2004. – ISBN 1–58113–812–1, S. 1242–1247

Tominski u. a. 2003
TOMINSKI, C. ; SCHULZE-WOLLGAST, P. ; SCHUMANN, H.: Visualisierung zeitlicher Verläufe auf geografischen Karten . In: *Proceedings GeoVis'2003, Hannover*, 2003

Tominski u. Schumann 2004
TOMINSKI, C. ; SCHUMANN, H.: An Event-Based Approach to Visualization. In: *IV '04: Proceedings of the Information Visualisation, Eighth International Conference on (IV'04).* Washington, DC, USA : IEEE Computer Society, 2004. – ISBN 0–7695–2177–0, S. 101–107

Tory u. Möller 2004a
TORY, M. ; MÖLLER, T.: Human Factors in Visualization Research. In: *IEEE Transactions on Visualization and Computer Graphics* 10 (2004), Januar/Februar, Nr. 1

Tory u. Möller 2004b
TORY, M. ; MÖLLER, T.: Rethinking Visualizations: A High-Level Taxonomy. In: *IEEE Symposium on Information Visualization (InfoVis'04), Austin, USA*, 2004, S. 151–158

Touchgraph 2007
TOUCHGRAPH: http://www.touchgraph.com. In: *Internet page*, 2007

Trafton u. a. 2002
TRAFTON, J.G. ; MARSHALL, S. ; MINTZ, F. ; TRICKETT, S.B.: Extracting Explicit and Implict Information from Complex Visualizations. In: HEGARTY, Mary (Hrsg.) ; MEYER, Bernd (Hrsg.) ; NARAYANAN, N. H. (Hrsg.): *Diagrammatic Representation and Inference, Second International*

Conference, Diagrams 2002, Callaway Gardens, GA, USA, April 18-20, 2002, Proceedings Bd. 2317, 2002. – ISBN 3–540–43561–1, S. 206–220

Trapp u. Pagendarm 1996
TRAPP, J. ; PAGENDARM, H.-G.: Data Level Comparative Visualization in Aircraft Design. In: *Proceedings IEEE Visualization'96, USA*, 1996, S. 393ff

Treinish 1994
TREINISH, L. A.: Severe Rainfall Events in Northwestern Peru: Visualization of Scattered Meteorological Data. In: O'CONNOR, Lisa (Hrsg.) ; STORMS, Penny (Hrsg.): *Proceedings of Visualization'94*, 1994

Treinish 1999
TREINISH, L.A.: Task-Specific Visualization Design. In: *IEEE Computer Graphics and Applications* (1999), September/October, S. 72–77

Treinish u. a. 1992
TREINISH, L.A. ; BUTLER, D.M. ; HIKMET, S. ; GRINSTEIN, G.G. ; BRYSON, S.T.: Grand Challenge Problems in Visualization Software. In: *Proceedings IEEE Visualization'92, USA*, 1992

Treinish u. a. 2003
TREINISH, L.A. ; PRAINO, A.P. ; CHRISTIDIS, Z.D.: Implementation Of Mesoscale Numerical Weather Prediction For Weather-Sensitive Business Operations. In: *19th Conference on IIPS*, 2003

Trembilski 2003
TREMBILSKI, A.: *Naturalistische Methoden zur Visualisierung meteorologischer Daten in Augmented Video*, TU Darmstadt, Fachbereich Informatik, Diss., 2003

Tufte 1983
TUFTE, E.R.: *The Visual Display of Quantitative Information.* Conneticut : Graphics Press, 1983

Tulip 2007
TULIP: http://www.tulip-software.org. In: *Internet page*, 2007

Unwin u. Theus 2004
UNWIN, A. ; THEUS, M.: Special issue on data visualization. In: *Computational Statistics* 19 (2004), Nr. 1

Unwin u. a. 1990
UNWIN, A. ; WILLS, G. ; HASLETT, J.: REGARD — Graphical Analysis of Regional Data. In: *'1990 Proceedings of the Section on Statistical Graphics', American Statistical Association*, 1990, S. 36–41

VHE 2007
VHE: Visible Human Explorer: http://www.cs.umd.edu/hcil/visible-human/vhe.shtml. In: *Internet page*, 2007

Viola u. a. 2004
VIOLA, I. ; KANITSAR, A. ; GRÖLLER, E.: Importance-Driven Volume Rendering. In: *Proceedings IEEE Visualization'05, Austin, USA*, 2004

Visage 2006
VISAGE: http://www.cs.cmu.edu/%7Esage/visage.html. In: *Internet page*, 2006

VisAxes 2007
VISAXES: http://wwwicg.informatik.uni-rostock.de/%7Ect/VisAxesNET.html. In: *Internet page*, 2007

VisDB 2007
VISDB: http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html. In: *Internet page*, 2007

VisServer 2007
VISSERVER: http://www.inxight.com. In: *Internet page*, 2007

Wagenknecht 2006
WAGENKNECHT, K.: *Metadaten-gesteuerte Visualisierung von fusionierten, klinischen Daten in verteilten Umgebungen*, Universität Rostock, Fakultät für Informatik und Elektrotechnik, Diplomarbeit, 2006

Wang u. Mueller 2004
WANG, L. ; MUELLER, K.: Generating Sub-Resolution Detail in Images and Volumes Using Constrained Texture Synthesis. In: *Proc. Visualization (Vis'04)*, 2004, S. 75–82

Ward 2002
WARD, M.O.: A taxonomy of glyph placement strategies for multidimensional data visualization. In: *Palgrave Information Visualization* 1 (2002), S. 194–210

Ware 2000
WARE, C.: *Information Visualization: Perception for Design*. San Francisco : Morgan Kaufmann Publishers, 2000

Weber u. a. 2001
WEBER, M. ; ALEXA, M. ; MÜLLER, W.: Visualizing Time-Series on Spirals. In: *IEEE Symposium on Information Visualization '01*, 2001, S. 21–28. – ISBN 0-7695-1342-5

Wehrend u. Lewis 1990
WEHREND, S. ; LEWIS, C.: A problem-oriented classification of visualization techniques. In: *IEEE Visualization'90*, 1990, S. 139–143

Weihai u. Zesheng 1994
WEIHAI, C. ; ZESHENG, T.: A highly interactive meteorological visualization system MeteoVis. In: *Pacific Graphics'94/CADDM'94*, 1994

Weinkauf u. a. 2004
WEINKAUF, T. ; THEISEL, H. ; HEGE, H.-C. ; SEIDEL, H.-P.: Boundary switch connectors for topological visualization of complex 3d vector fields. In: *Proceedings Data Visualization*, 2004

West u. Machiraju 1998
WEST, J.E. ; MACHIRAJU, R.: A Numerical Imaging Approach to Comparative Visualization. In: *Proceedings IEEE Visualization'98*, 1998

Westphal u. Blaxton 1998
WESTPHAL, C. ; BLAXTON, T.: *Data Mining Solutions - Methods and Tools for Solving Real-World Problems*. 8. New York : John Wiley & Sons, Inc, 1998

van Wijk u. van Selow 1999
Wijk, J. J. ; Selow, E. R.: Cluster and Calendar Based Visualization of Time Series Data. In: *IEEE Symposium on Information Visualization'99*, 1999, S. 4–9

van Wijk u. van de Wetering 2000
Wijk, J. J. ; Wetering, H. van d.: Cushion Treemaps: Visualization of Hierarchical Information. In: *IEEE Symposium on Information Visualization (InfoVis'00), San Francisco*, 2000, S. 73–78

Wilkinson 1999
Wilkinson, L.: *The Grammar of Graphics*. Springer-Verlag, 1999

Wills 1999
Wills, G. J.: *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, 1999

WilmaScope 2007
WilmaScope: http://sourceforge.net/projects/wilma. In: *Internet page*, 2007

Wise u. a. 1995
Wise, J.A. ; Thomas, J. ; Pennock, K. ; Lantrip, D. ; Pottier, M. ; Schur, A. ; Crow, V.: Visualizing the Non-Visual: Spatial Analysis and Interaction and Interaction with Information from Text Documents. In: *IEEE Syposium on Information Visualization (InfoVis'95)*, 1995, S. 51–58

Witten u. Frank 2000
Witten, I. ; Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco : Morgan Kaufmann Publishers, 2000

Wong u. Bergeron 1997
Wong, P. C. ; Bergeron, R. D.: 30 Years of Multidimensional Multivariate Visualisation. In: *In: G. Nielson, H. Hagen, H. Müller: Scientific Visualisation. IEEE Computer Society, Los Alamitos* (1997), S. 3–33

Wong 1999
Wong, P.C.: Visual Data Mining. In: *IEEE Computer Graphics & Applications* (1999), September/October, S. 20–21

Wong u. a. 2000
Wong, P.C ; Foote, H. ; Leung, R. ; Jurrus, E. ; Adams, D. ; Thomas, J.: Vector Fields Simplification - A Case Study of Visualizing Climate Modeling and Simulation Data Sets. In: *Proceedings of the IEEE Visualization (Vis'00)*, 2000, S. 485–488, 596

Woodring u. Shen 2006
Woodring, J. ; Shen, H.-W.: Multi-variate, Time-varying, and Comparative Visualization with Contextual Cues. In: *Proceedings IEEE Visualization'06*, 2006

Wright 1995
Wright, W.: Information Animation Applications in Capital Markets. In: *IEEE Syposium on Information Visualization (InfoVis'95)*, 1995, S. 19–25

Wrobel 2004
Wrobel, M.: *Multidimensionale, heterogene, visualisierbare Datenräume: Anforderungen,*

Entwurf und Implementierung einer adaptiven und interaktiven Schnittstelle für transdisziplinäre wissenschaftliche Daten im Kontext der Erdsystemanalyse, Freie Universität Berlin, Diss., 2004

XGobi 2006
XGobi: http://www.research.att.com/areas/stat/xgobi. In: *Internet page*, 2006

Xue u. a. 2004
Xue, D. ; Zhang, C. ; Crawfis, R.: Rendering Implicit Flow Volumes. In: *Proceedings IEEE Visualization'04, Austin, USA*, 2004

Yang u. a. 2004
Yang, J. ; Patro, A. ; Huang, S. ; Mehta, N. ; Ward, M.O. ; Rundensteiner, E.A.: Value and Relation Display for Interactive Exploration of High Dimensional Datasets. In: *IEEE Symposium on Information Visualization (InfoVis'04), Austin*, 2004, S. 73–80

Yang u. a. 2002
Yang, J. ; Ward, M.O. ; Rundensteiner, E.A.: InterRing: An Interactive Tool for Visually Navigating and Manipulating Hierarchical Structures. In: Wong, P.C. (Hrsg.) ; Andrews, K. (Hrsg.): *Proc. IEEE Symposium on Information Visualization (InfoVis'02)*, IEEE Computer Society, 2002, S. 77–84

Yang u. a. 2003
Yang, J. ; Ward, M.O. ; Rundensteiner, E.A. ; Huang, S.: Visual hierarchical dimension reduction for exploration of high dimensional datasets. In: *Proc. of Joint IEEE/EG Symposium on Visualization, VisSym'03, Grenoble, France*, 2003, S. 19–28

Yang u. a. 2005
Yang, J. ; Ward, M.O. ; Rundensteiner, E.A. ; Ribarsky, W.: Value and Relation Display: A General Framework for Interactive Exploration of High Dimensional Datasets. In: *IEEE ?* (2005)

Zhao u. a. 2005
Zhao, S. ; McGuffin, M.J. ; Chignell, M.H.: Elastic Hierarchies: Combining Treemaps and Node-Link Diagrams. In: *Proceedings IEEE Symposium on Information Visualization (InfoVis'05)*. Mineapolis, USA, 2005, S. 57–64

Zhou u. a. 2002a
Zhou, H. ; Chen, M. ; Webster, M.: Comparative evaluation of visualization and experimental results using image comparison metrics. In: *Proceedings IEEE Visualization'02*, 2002, S. 315–322

Zhou u. a. 2002b
Zhou, M.X. ; Chen, M. ; Feng, Y.: Building a Visual Database for Example-based Graphics Generation. In: *IEEE Syposium on Information Visualization (InfoVis'02)*, 2002

Zhou u. Feiner 1996
Zhou, M.X. ; Feiner, S.K.: Data Characterization for Automatically Visualizing Heterogeneous Information. In: *Proceedings of the IEEE Information Visualization (InfoVis'96)*, 1996, S. 13–20

Zornow 2006
Zornow, D.: *Visualisierung des Verhaltens von gewöhnlichen Differenzialgleichungen im Umfeld der Klimaforschung*, Universität Rostock, Fakultät für Informatik und Elektrotechnik, Diplomarbeit, 2006

Selbstständigkeitserklärung

Ich erkläre, dass ich die eingereichte Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Rostock, d. 06. 07. 2007

Thesen

1. Das visuelle Data Mining (VDM) ermöglicht durch eine enge Verknüpfung von automatischen Analyseverfahren und interaktiven Visualisierungstechniken neue Wege der Datenanalyse, indem es die Fähigkeiten der menschlichen Wahrnehmung mit den Fähigkeiten moderner Computersysteme koppelt. Diese Verknüpfung verstärkt das Verständnis und das Vertrauen in die beteiligten Verfahren und eröffnet neue Einsichten in die Daten.

2. Am Beispiel der Analyse von Klimadaten wurden in dieser Arbeit die breitgefächerten Einsatzmöglichkeiten des VDM demonstriert. Im besonderen erlauben interaktive Visualisierungstechniken und deren Kopplung mit automatischen VDM-Verfahren die in diesem Umfeld auftretenden heterogenen, multi-variaten Datensätze auf geeignete Art und Weise zu untersuchen und neue Erkenntnisse (z.B. versteckte Muster oder Trends) abzuleiten.

3. Visualisierungstechniken haben ein großes Potential, multi-variate räumliche/zeitliche Daten zu analysieren. Zur visuellen Analye von Klimadaten eignen sich eine Vielzahl von Darstellungstechniken, die in üblichen Systemen zur Wetter- und Klimavisualisierung z.T. nicht eingesetzt werden. In der vorliegenden Arbeit wurden insbesondere neue metapherbasierte Darstellungen räumlicher Daten, spezielle Darstellungstechniken des Zeitbezuges, Methoden zur Analyse des Merkmalsraumes und Verfahren zur vergleichenden Visualisierung vorgestellt.

4. Automatische Analyseverfahren wie die Cluster- und Hauptkomponentenanalyse erlauben es, große Datenmengen zu explorieren und zu strukturieren. Um die Ergebnisse solcher Verfahren – welche zum Teil auch im räumlichen und/oder zeitlichen Bezug gegeben sind – besser interpretieren und kommunizieren zu können, lassen sich speziell darauf zugeschnittene Visualisierungstechniken einsetzen. In der vorliegenden Arbeit wurde der Ansatz, die verschiedenen, bei den genannten Analyseverfahren auftretenden Ergebnisse in alle Schritte des Visualisierungsprozesses einzubeziehen, systematisch untersucht und umgesetzt.

5. Die Einbeziehung von Verfahren und Methodiken des visuellen Data Mining für den gesamten Prozess der Modellbildung, -simulation und -evaluation ist ein vielversprechender Ansatz. Er erlaubt den Prozess der Generierung und Validierung von Modellen wesentlich zu unterstützen und ermöglicht neue Vorgehensweisen bei der Modellierung in der Klimaforschung.

6. Die breite Anwendung moderner Visualiserungstechniken erfordert einen hohen Grad an Unterstützung. Die (halb-)automatische Auswahl und Parametrisierung ist ein wichtiger Mechanismus, um die Lücke zwischen dem Anwenderwissen über Visualisierung auf der einen Seite und dem Wissen aus der Visualisierungsforschung auf der anderen Seite zu schließen. Der in dieser Arbeit vorgestellte zweistufige Mechanismus zum Visualisierungsdesign und dessen Skalierbarkeit auf spezielle Anwendungsszenarien vereinfacht den Zugang der Anwender zu für sie z.T. unbekannten Techniken und die Generierung aussagekräftiger Darstellungen. Der Einsatz von Deskriptoren zur Beschreibung der Eigenschaften der Visualisierungstechniken sowie die Verwendung von flexibel erweiterbaren Regeln ist hierbei ein effektives Vorgehen. Wichtige Einflussfaktoren auf die Auswahl und Parametrisierung einer Visualisierung sind Metadaten und Analyseziele. Die Spezifikation dieser Einflussfaktoren muss sowohl anwendungs- und datenklassenunabhängige als auch -abhängige Aspekte einbeziehen. In der Arbeit wurden Mechanismen entworfen und umgesetzt, um Metadaten und Ziele unter einem hohen Grad an Nutzerunterstützung erheben und verwalten zu können.

7. Um die im Rahmen der Arbeit bereitgestellten VDM-Techniken flexibel einsetzen und verknüpfen sowie Verfahren zur Unterstützung der Anwender mit einbeziehen zu können, wurde eine flexibel einsetzbare Komponentenbibliothek entworfen und umgesetzt. Essentielle Bestandteile einer solchen Komponentenbibliothek sind neben einer Bibliothek von VDM-Techniken auch Methoden zur Verknüpfung dieser Techniken, zur Verwaltung von Einflussfaktoren auf das VDM, zum Analyseprozessmanagement und zur halbautomatischen Erzeugung von geeigneten Visualisierungen. Damit bildet diese Komponentenbibliothek die Basis für verschiedenartige VDM-Tools, insbesondere für die Systeme SimEnvVis und VisAna.

8. Es wurde eine Vielzahl von prototypischen Softwaretools zur Validierung der aufgestellten Thesen konzipiert und realisiert. Dabei hat sich das Framework SimEnvVis als Visualisierungsschnittstelle zum Simulationsystem SimEnv über den prototypischen Status hinaus entwickelt, und wird – unter Bereitstellung einer großen Bandbreite an Visualisierungstechniken – am Potsdam Institut für Klimafolgenforschung für die Auswertung von Simulationsexperimenten eingesetzt. Gerade die Einbeziehung von multi-variaten Methoden und das integrierte Visualisierungsdesign ermöglichen es den Anwendern, geeignete Darstellungen zu erzeugen und neue Einsichten in ihre Modelle zu erlangen.