INT. J. GEOGRAPHICAL INFORMATION SCIENCE, 1999, VOL. 13, NO. 4, 311–334

Research Article

Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods*

ALAN M. MACEACHREN, MONICA WACHOWICZ, ROBERT EDSALL, DANIEL HAUG
GeoVISTA Center, Department of Geography, 302 Walker, Penn State University, University Park, PA 16802 USA

and RAYMON MASTERS
Center for Academic Computing, Department of Architecture, and GeoVISTA Center, Penn State University, PA, USA

Abstract. We present an approach to the process of constructing knowledge through structured exploration of large spatiotemporal data sets. First, we introduce our problem context and define both Geographic Visualization (GVis) and Knowledge Discovery in Databases (KDD), the source domains for methods being integrated. Next, we review and compare recent GVis and KDD developments and consider the potential for their integration, emphasizing that an iterative process with user interaction is a central focus for uncovering interest and meaningful patterns through each. We then introduce an approach to design of an integrated GVis-KDD environment directed to exploration and discovery in the context of spatiotemporal environmental data. The approach emphasizes a matching of GVis and KDD meta-operations. Following description of the GVis and KDD methods that are linked in our prototype system, we present a demonstration of the prototype applied to a typical spatiotemporal dataset. We conclude by outlining, briefly, research goals directed toward more complete integration of GVis and KDD methods and their connection to temporal GIS.

1. Introduction

Large environmental data sets represent a major challenge for both domain and information sciences. The domain sciences, most of which developed under data poor conditions, must now adapt to a world that is data rich—so data rich that large volumes of data often remain unexplored while the media they are stored upon deteriorate or become obsolete. The information sciences, most of which developed in a pre-computer era or when batch processing by computer was the norm, must now adapt to a world that is not only digital but highly dynamic—in which there is a potential for computers to produce answers in real time as an analyst explores data and poses 'what if' questions. Much of the environmental data being generated

*Web supplement located at: http//www.geovista.psu.edu/ijgis.htm

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online ©1999 Taylor & Francis Ltd
http://www.tandf.co.uk/JNLS/gis.htm
http://www.taylorandfrancis.com/JNLS/gis.htm

today (e.g. from the Earth Observation System, from monitoring efforts in endangered ecosystems, from meteorological stations, etc.) includes georeferencing. The spatial aspects of these data are, in fact, often a primary focus of analysis—for studies of pollutant dispersal, forest fragmentation, and other applications. Repeated observation is critical to answering the most important environmental science questions (those related to environmental process), thus environmental data sets typically have temporal as well as spatial components.

It is in this context of both rapidly evolving computing technologies and increasing spatiotemporal data availability that we see a substantial challenge for research in methods for spatiotemporal data analysis. The challenge is two-fold: to extend GIS, spatial analysis, and visualization methods (developed for application to static spatial data) into a spatiotemporal domain *and* to integrate these components of geographic information science to produce new methods (and associated tools) that facilitate environmental science and environmental policy decisions. This overall challenge is the focus of the Apoala Project underway in our laboratory.

In this paper we address one aspect of the problem, the development and integration of data analysis and visualization methods designed to facilitate identification and interpretation of spatial and spatiotemporal features. Toward this end, we focus on methods associated with the new but expanding fields of Geographical Visualization (GVis) and Knowledge Discovery in Databases (KDD). GVis has been defined as 'the use of concrete visual representations—whether on paper or through computer displays or other media—to make spatial contexts and problems visible, so as to engage the most powerful of human information-processing abilities, those associated with vision (MacEachren 1992, p. 101). A primary focus of GVis research over the past decade has been the role of highly interactive visualization tools in facilitating identification and interpretation of patterns and relationships in complex data. KDD has been defined as: 'the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data' (Fayyad *et al.* 1996, p. 6). As with GVis (see below) KDD is characterized as a multistep *process*, a process in which 'data mining' algorithms (algorithms through which patterns are extracted from data) play a central role.

In the next section, we elaborate upon these definitions of GVis and KDD and build the case for their integration. Section three outlines our approach to an integrated *knowledge construction* system (GKConstruct), with emphasis on integration of GVis and KDD methods, and describes key features implemented in an early prototype. Knowledge construction is defined here as the active process of manipulating 'data' (which can include numerical and other abstract representations of real world phenomena) to arrive at abstract models of relationships among phenomena in the world that facilitate our understanding of those phenomena and, ultimately, of the world. In section four, we demonstrate how the linked GVis and KDD methods can be applied to a sample time series of georeferenced climate data. We conclude, in the final section, with a brief discussion of our more ambitious effort to link these knowledge construction tools more directly to temporal GIS.

2. Seeking patterns in data: complimentary approaches

Particularly when applied to scientific data, GVis and KDD have similar goals. They differ, however, in the extent to which they rely upon human vision or computational methods to process data. In this section, we review, separately, the underlying principles and key developments of the past decade in both fields (each builds on

longer traditions, but has been identified as a distinct research stream for about a decade).

2.1. *Geographic visualization*

GVis builds from a base in cartography, geographical information systems, image analysis, and spatial analysis and has strong ties to related efforts in scientific and information visualization more generally (e.g. JPL 1987, Kaufman and Smarr 1993, Treinish 1993) and to exploratory data analysis efforts in statistics (e.g. Cook *et al.* 1997, Carr *et al.* 1998)—see MacEachren and Kraak (1997) for a recent review and bibliography. Our focus here is on the common themes linking various GVis research activities. Key among these is a view of GVis as a *process*, part mental and part concrete (involving human visual thinking, computer data manipulation, and human computer interaction), in which vast quantities of georeferenced information are sifted and manipulated in the search for patterns and relationships. Among the first process oriented perspectives on GVis is DiBiase's (1990) characterization of visualization as a four-stage process that facilitates science, seen itself as a process. A complimentary view focusing on the perceptual-cognitive process of interpreting and understanding georeferenced visual displays was offered in MacEachren and Ganter (1990). Others who have adopted a process oriented approach to GVis include Monmonier (1992) and Openshaw *et al.* (1994), with their ideas about use of map animation in the process of spatial analysis and Mitas *et al.* (1997) who emphasize the use of GVis in the process of developing and applying complex landscape simulations and land use optimizations.

Beyond the emphasis on GVis as a process, several other common threads link aspects of GVis research. Among these are: (1) the iterative nature of successful human interpretation of visual displays (e.g. MacEachren 1995, Buttenfield and Mackaness 1991, Wood 1994), (2) the importance of interactivity that facilitates both iteration and access to multiple perspectives on information (MacEachren 1994, Dykes 1997, Kraak 1998), and (3) the overarching goal for visualization methods (when applied to science) of finding patterns and relationships in data (MacEachren and Ganter 1990, Dorling 1992, Wilbanks *et al.* 1997).

Most GVis research, whether by geographical information scientists or others, has focused on overcoming hurdles involved in applying the latest technology to the visual display and analysis of spatial (and spatiotemporal) data. It is our contention, that fundamental advances in GVis will depend as much on establishing a solid theoretical basis for GVis methods (and to linking that theory with related information sciences theory development) as it does on application of advances in technology (e.g. those leading to increasingly realistic displays).

A coherent theoretical framework for GVis is just beginning to emerge. That framework integrates the formalism of semiotics as an approach for understanding and modeling representational relationships with a cognitive perspective on the process of using visualization methods to facilitate scientific understanding (MacEachren 1995). The goal for this integrated perspective is to develop a conceptualization of GVis as a *process* that involves humans achieving insight by interacting with data through use of manipulable visual displays that provide representations of these data and of the operations that can be applied to them. From semiotics, we gain tools for understanding abstract representations of phenomena and processes (i.e. representations in digital, visual, and other forms) as well as methods for explaining how meaning is brought to the representations by their creators and

users. From the study of cognition we gain a perspective on the ways in which human information users conceptualize problem domains, process visual displays, and link mental schemata to actions through interface tools (figure 1). This integrated cognitive-semiotic perspective serves as a base from which to consider three categories of visualization meta-operations that are at the heart of the data exploration components of the Apoala Project: feature identification, feature comparison, and feature interpretation. These three operation forms are defined below, briefly and, in §3, linked to three categories of KDD operations (introduced in §2.2). The three visualization operations presented here are a revision of a similar set presented in MacEachren (1995). A distinction made initially, between spatial and spatiotemporal features, is dropped here and replaced by addition of a higher level process, feature interpretation. For a detailed explanation of the initial visualization operations and methods through which each can be implemented, see MacEachren (1995, pp. 361–434). The GVis operations delineated have parallels in traditional cartographic approaches to map reading, where tasks are often broken into three levels. These include Keates' (1982) delineation of three levels of map symbol perception (detection, discrimination, and identification), Olson's (1976) three levels of map symbol interpretation (comparison of individual symbol characteristics, assessment

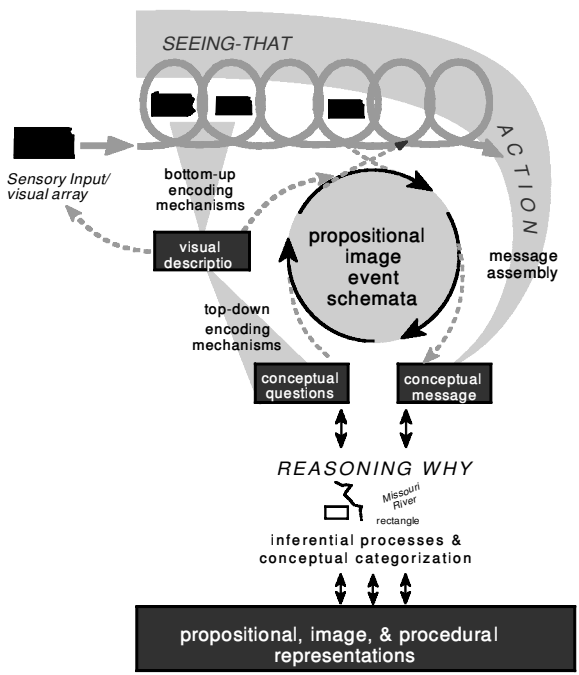

Figure 1. Extended feature ID model for map-based visualization. This diagram attempts to integrate what we know about human perception and cognition related to the interpretation of visual information displays. Emphasis is given to the iterative nature of human experts examining a representation and attempting to interpret that representation and use it as a prompt to insight (a process that cycles between seeing or noticing 'features' in the display and interpreting those features by matching what is seen with what is known. The model suggests that knowledge is stored in the form of cognitive representations that are drawn upon to generate knowledge schemata (methods for matching what is sensed with prior knowledge).

of symbol groupings, and use for decision-making or knowledge building), Meuhrcke's (1986) three levels of map use (reading, analysis and interpretation), and Bertin's (1983) three levels of information (elementary, intermediate, and overall). Key distinctions between the three GVis operations presented here and past characterizations of map reading levels include an emphasis on 'features' in the data (rather than symbols to be translated) and the linking of an object-oriented perspective on features with an integrated cognitive-semiotic theory of geo-representation.

2.1.1. *Feature identification*

Feature identification focuses on the task of finding instances of identifiable 'features' in spatial or spatiotemporal data. Emphasis is on examining the distribution of data in all of its dimensions in an effort to notice any distinct object, regularity, anomaly, hot spot, etc. Features can range from individual 'objects' to patterns of objects and can vary in size and distinctness. Patterns of objects may be seen as a whole, with the 'pattern' not obvious (e.g. a 'forest' composed of tree objects) or the pattern itself may be the key aspect of a feature (e.g. karst topography). Features can also vary in the extent to which they reflect the state of things (i.e. are static over the time frame of interest) or exist due to an action or process (e.g. a street intersection versus a traffic accident). From a semiotic perspective, emphasis in feature identification is on the sign-vehicles of the display (i.e. the display's graphic expressions or carriers of meaning), on their conjunctions with one another, and on the organization among sign-vehicles (sign system syntactics). Feature identification as a visualization operation, thus, emphasizes what is noticed in a display (at an abstract level) rather than what the display represents. Visualization methods that are particularly useful for noticing abstract spatial features and patterns include focusing, sequencing, multivariate glyphs, and remappings from one space to another. Methods that facilitate noticing abstract spatiotemporal features and patterns include animation—though time and space, space-time cubes, small multiples (with each representing a particular time), plus any of the feature-noticing methods (cited above) when applied to the spatiotemporal aspects of the data.

2.1.2. *Feature comparison*

Feature comparison extends consideration to multiple objects or patterns. The goal is to 'enhance the likelihood that an analyst will see not only features, but relationships among features' (MacEachren 1995, p. 401). Feature comparisons can focus on any of the attributes used to characterize each feature. Location is the most obvious for geographical features, thus tools for noticing spatial co-occurrence are particularly important, but comparison might be on the basis of shape, degree of pattern clustering, temporal correspondence, or any of a number of other attributes. Semiotically, emphasis here is also on sign-vehicles, but particularly on the syntactics of relationships among sign-vehicles or sign-vehicle sets to one another. Visualization methods that facilitate feature comparison include: scatterplot matrices, parallel coordinate plots, small multiples, map overlay, multivariate colour schemes, and linked brushing using any of the above.

2.1.3. *Feature interpretation*

A feature can be both a member of an abstract 'object class' (from the perspective of object-oriented data modeling) and a member of a 'category' of real world entity (from the perspective of cognitive category theory). The goal of feature interpretation

is to merge the two by bringing domain knowledge to bear on the identified features and their relationships. From a semiotic perspective, the goal is to link the 'sign-vehicle' of an identified feature (e.g. a visually distinct grouping in a map distribution or an anomaly in multidimensional data space, etc.) with a real world 'referent' (phenomenon) through a shared 'interpretant' (a meaning relationship through which we characterize the aspects of the referent that are modeled by the sign-vehicle). Feature interpretation tools, thus, provide connections between abstract data representations, metadata (data about those data), an analyst's prior knowledge, and knowledge sources external to the data set being explored (e.g. digital libraries). Through these connections, feature behavior and relationships of features can be related to behavior of and relationships among real-world phenomena. Visualization methods that facilitate interpretation of spatiotemporal features include all methods mentioned above combined with methods of information visualization—methods for visualizing knowledge constructs, verbal information, etc. Examples of the latter include cone trees (Robertson 1991), spider diagrams (Armstrong *et al.* 1992), and spatialization of information (Skupin and Buttenfield 1996).

2.2. *Knowledge discovery in databases*

The development of KDD coincides with an exponential increase in digital data generated by and available to science, government, and industry. The definition cited above for KDD, foscusing on a *process* of identifying patterns in data that are not only valid and useful but also understandable and novel, has been generally accepted (see Frawley *et al.* 1991, Ester *et al.* 1995, Brachman and Anand 1996). While various authors have proposed somewhat different delineations of the process, we find that by Fayyad *et al.* (1996) particularly suited to the application of KDD in a spatio-temporal data context. These authors describe the KDD process as consisting of five steps:

— Data selection—having two subcomponents: (*a*) developing an understanding of the application domain and (*b*) creating a target dataset from the universe of available data.
— Preprocessing—including data cleaning (such as dealing with missing data or errors) and deciding on methods for modeling information, accounting for noise, or dealing with change over time.
— Transformation—using methods such as dimentionality reduction to reduce data complexity by reducing the effective number of variables under consideration.
— Data mining—having three subcomponents: (*a*) choosing the data mining task (e.g. classification, clustering, summarization), (*b*) choosing the algorithms to be used in searching for patterns, and (*c*) the actual search for patterns (applying the algorithms).
— Interpretation/evaluation—having two subcomponents: (*a*) interpretation of mined patterns (potentially leading to a repeat of earlier steps), and (*b*) consolidating discovered knowledge, which can include summarization and reporting as well as incorporating the knowledge in a performance system.

Although the list above might suggest a linear process, KDD in practice is anything but linear. Brachman and Anand (1996, p. 39), for example, contend that '... knowledge discovery is a *knowledge-intensive* task consisting of *complex interactions*, *protracted over time*, between a *human* and a (large) database, ...' Fayyad

et al. (1996) emphasize that KDD involves both interaction and iteration, that humans repeat analysis steps repeatedly as knowledge is being refined—thus our contention above that KDD is really about knowledge *construction* rather than *discovery*.

Several KDD methods have recently emerged from the literature and they differ in the conceptualizations developed, reflecting (in part) their separate developments in fields such as database systems, machine learning, statistics, and artificial intelligence. We have grouped them into three categories of KDD meta-operations of increasing complexity: *concept hierarchy and structure extraction*: (a process in which data abstractions are derived and linked at multiple conceptual levels), *categories extraction and classification* (a process of deriving classes, clusters, rules, and/or patterns from target data and using the result to assign data to classes that result from the categorization process), and *phenomenon extraction* (a process of deriving representations of real-world phenomena from a target data set). Each is discussed below, briefly.

2.2.1. *KDD methods for concept hierarchy and structure extraction*

Concept hierarchy and structure extraction is a KDD process that models attributes of a multidimensional data set at multiple levels (e.g. for the spatial component, at multiple scales, resolutions, or levels). Techniques used involve multi-level data generalization, summarization, and characterization. This process has much in common with cartographic generalization—particularly as cartographic generalization has been formalized with a goal of building hierarchically structured multi-resolution databases (Frank and Timpf 1994).

Two concept hierarchy and structure extraction techniques suggested in the KDD literature are the Data Cube Approach (Harinarayan *et al.* 1996, Gray *et al.* 1997) and the Attribute-Oriented Induction Approach (Han 1995, Han and Fu 1996). The core of both is creation of a concept hiearchy (or lattice) of attributes. In the Data Cube Approach, concept hierarchies are specified using aggregation functions (Group by, transitive binary relationships) and database view (table view, object manager view) definitions. Attribute-Oriented Induction methods are based on machine learning research developments in which concept hierarchies are specified according to the relations among attributes by using rules definition.

Concept hierarchies can relate to spatial, temporal, or attribute components of a data variable. A spatial concept hierarchy, for example, might specify a nested series of drainage basins to which a water quality sample is linked or a hierarchy of political units to which a census sample belongs. A complementary pair of simple temporal concept hierarchies is illustrated in figure 2. Several authors have investigated methods for extracting concept hierarchies for spatial data (see: Wang *et al.* 1997, Koperski *et al.* 1998, and Han *et al.* 1997).

2.2.2. *KDD methods for categories extraction and classification*

A KDD categories extraction and classification process involves the search for common attributes among a set of objects, and then the arrangement of these objects into classes, clusters, or patterns according to a meaningful partitioning criteria, model or rule. An object can be a physical feature (e.g. stream-flow measured at irregularly-spaced Gauging stations), an abstract feature (e.g. precipitation deficit—the deviations from climate means) or an event (e.g. a drought occurring over a

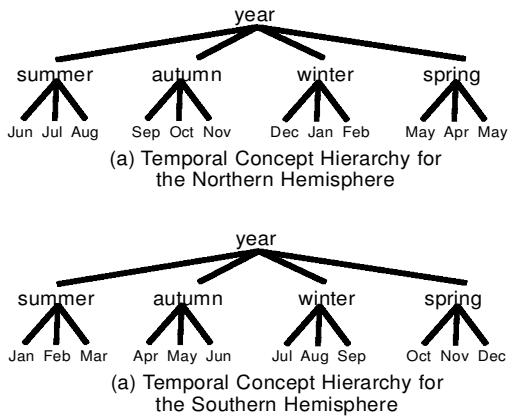

Figure 2. Simple concept hierarchy for climate analysis—one (*a*) appropriate for study of northern hemisphere climate, the other (*b*) appropriate to the southern hemisphere.

spatial and temporal extent). In general, methods for categories extraction and classification can be grouped into two broad types: symbolic and statistical.

Symbolic methods address the issue of producing sets of statements about local dependencies among objects in a rule form. A vast literature is available that describes mining algorithms based on symbolic methods for implementing rule induction tools in KDD (see Chen *et al.* 1996). Some example symbolic method algorithms are CLARANS—Clustering Large Applications based upon Randomized Search (Ng and Han 1994) and BIRCH (Balanced Iterative Reducing and Clustering) (Zhang *et al.* 1996). Based on CLARANS, two spatial data mining algorithms were developed SD-CLARANS (spatial dominant algorithm) and NSD-CLARANS (non-spatial dominant algorithm) (Ester *et al.* 1995).

In statistical methods, the focus is on exploiting statistical approaches (probability distributions, hypothesis testing, model estimation and scoring) for performing the mining-task of extracting discriminators from a data set (Hosking *et al.* 1997). Statistical techniques applied for extracting categories are based on supervised/unsupervised learning, cluster analysis, and related methods. We are particularly interested in unsupervised methods that can be used to uncover unknown spatiotemporal patterns in large data sets. One potentially appropriate data mining tool of this type is AutoClass (Cheeseman and Stutz 1996), a public domain software package that offers unsupervised classification based upon Bayesian classification theory. The AutoClass mining algorithm is designed to work under an assumption that the class labels for target classes are *a priori* unknown. The classification problem is formulated as an optimization problem that deals with maximizing the likelihood probabilities of a set of models which are regarded as a set of assumptions. The result is a fuzzy classification where each object has a probability of membership in each class.

2.2.3. *KDD methods for phenomenon extraction*

Phenomenon extraction involves finding instances of an existing object class that represents a real world phenomenon (e.g. thunderstorm) and/or instances of a previously unknown phenomenon type (that will be used to define a new object class). A key distinction between phenomenon extraction and previously defined operations

is the emphasis on linking data 'features' to real world 'phenomena' through the application of domain knowledge. For spatiotemporal environmental data (the data of interest here), the task becomes one of extracting environmental phenomena from observational and simulated data sets. The Content-based Querying in Space and Time (CONQUEST) system represents one attempt at spatiotemporal environmental phenomenon extraction (Slortz *et al.* 1995). CONQUEST has been used to extract (from linked data sources) two canonical climate phenomena that (as seems true for most spatiotemporal phenomena) have imprecise concept definitions in the literature, cyclones and 'blocking features'.

2.3. *Common themes and potential for integration*

It is clear that GVis and KDD share perspectives related to both goals and approach. For each, a primary goal is to find, relate, and interpret interesting, meaningful, and unanticipated features (objects or patterns) in large data sets. In both cases, knowledge construction (or discovery) is viewed as a complex process and researchers have recognized the important role of the human domain expert in guiding the process and in interpreting results. In addition, methods in GVis and KDD both emphasize iteration as central to their effective application. Neither a single visual representation of a data set nor a single data mining run is expected to result in profound insight. It is only by repeated application of methods, with systematic changes in parameters, that a coherent picture is expected to emerge.

3. Bringing GVis and KDD methods together

The KDD literature contains frequent mention of the importance of visualization (e.g. Brachman and Anand 1996, Uthurusamy 1996). In most cases, however, visualization is considered only as a tool to facilitate the interpretation-evaluation stage of KDD. (Slortz *et al.* 1995 and Simoudis *et al.* 1996 are two exceptions). Our goal, in contrast, is a more complete integration of GVis and KDD methods. In order to achieve this goal, we approach system integration at three levels: conceptual, operational, and implementational (Howard and MacEachren 1997).

3.1. *Conceptual level*

At the conceptual level both GVis and KDD share the same high-level goal, the construction of knowledge that can be used to advance science, increase business profits, manage environmental resources, and related applications. Critical issues that underlie the knowledge construction process are: what kind of spatiotemporal data are meant to be visualized and mined (e.g. environmental and/or socioeconomic, spatially discrete and/or continuous), what particular kinds of outcomes are required from the process (e.g. hypotheses about relationships, predictions of a future state), and who are the users of knowledge obtained (e.g. domain scientists, policy analysts). Decisions about issues at this stage act as constraints on GVis-KDD integration at the operational level.

3.2. *Operational level*

The operational level deals with specification of appropriate methods, and combinations of methods, for achieving conceptual level goals. We contend that integration of GVis and KDD methods is essential in order to take full advantage of the strengths of human analysts (domain expertise and visual pattern recognition abilities) and computers (raw processing power). We propose a 3×3 conceptualization

of integrated GVis and KDD methods that links the meta-operation categories delineated for each above (figure 3).

As a start toward comprehensive integration of GVis and KDD methods associated with these meta-operation pairings, we have begun the process of building and linking tools that support the data mining and interpretation/evaluation steps of KDD. Emphasis at present is on row one and two, column two, of the meta-operation matrix (use of feature identification and comparison tools to facilitate categories extraction and classification). In the subsection below, we introduce the specific GVis and KDD methods we are developing (and/or adapting) and in section four we demonstrate application of these methods using our initial prototype.

3.3. *Implementational level*

At the implementational level, choices are made about specific tools that meet operational level goals, about appropriate algorithms that underlie those tools, and about specific software/hardware environments within which the algorithms can be realized. In developing and implementing GVis methods and tools to support the GVis meta-operations detailed above, we have adapted and extended a variety of visualization and EDA techniques introduced by us and by others over the past decade. In relation to KDD methods, we rely exclusively on methods provided in AutoClass. As a result, discussion below focuses primarily on the GVis methods we integrate to complement AutoClass.

3.3.1. *GVis methods*

The GVis operation categories detailed above are considered meta-operations—because each suggests an overall operational goal that is best approached through

GVis Operations	KDD Operations: concept hierarchy and structure extraction	KDD Operations: categories extraction and classification	KDD Operations: phenomenon extraction
feature identification	can help an analyst notice levels at which features/patterns appear, thus the key levels to be modeled in the concept hierarchies	helps to find category bounds, explore their fuzziness, identifies specific data individuals that are prototypical of a category and those that are anomolous	can be used to find exemplars of real world phenomena that match criteria derived in the category extraction operation
feature comparison	can help to relate concept hierarchies for different attributes a step toward creating an appropriate overall multidimensional structure for relating them	facilitates comparison of categories to determine the key attribute dimensions along which they relate or differ, determine which categories are similar	can be used to compare phenomena to derive rules that describe relationships (perhaps causal ones) between them
feature interpretation	facilitates linking an internal version of concept hierarchy and/or concept structure to domain knowledge as part of the process of refining each to fit a particular theme	facilitates the explanation of categories that are extracted in relation to physical parameters of the real world	can be used to help understand real world phenomena by developing abstract characterization of those phenomena and modeling their behaviors

Figure 3. Integration of GVis and KDD methods. With each pairing of operation categories, we cite possible outcomes of the merger as it relates to knowledge construction goals. The possibilities are intended as examples from a larger potential list. Examples cited are derived from a perspective of what we gain by adding GVis to KDD operations. A similar matrix could be derived taking the alternative perspective.

the iterative application of a set of more specific visual analysis operations. In GKConstruct, dynamic 'interactors' allow users to apply these visual operations to particular representation forms and the representation forms are related to one another through dynamic linking. With dynamic linking, data objects that share some aspect (e.g. spatial location) are linked across display views (thus across representation forms in our system) so that an action on a data object in one view will be reflected by a complementary action on corresponding data objects in other views.

In our initial GKConstruct prototype we have implemented three dynamically linked general *representation forms*: geoviews, 3D scatterplots, and parallel coordinate plots. Each is described below:

— *Geoviews*: three-dimensional windows in which geographic space is mapped to display space in at least two of the dimensions (figure 4). The third dimension is used to represent either the third spatial dimension (elevation/depth) or to represent time (with time treated as a linear dimension).
— *3D scatterplots*: representations of the relationships between three variables, with one variable plotted on each axis of a cube (figure 5). 3D scatterplots are a simple form of 'spatialization' in which non-spatial data 'dimensions' are mapped to the two- or three-dimensions of a display space. Spatialization suggests, and takes advantage of, the metaphor of near = similar | far = different. Using linked views, the 'mapping' of non-spatial aspects of georeferenced data to the spaces of a display can be grounded in more intuitive representations that map geographic space to display space (our geoviews).
— *Parallel coordinate plots*: data representations that contain several parallel

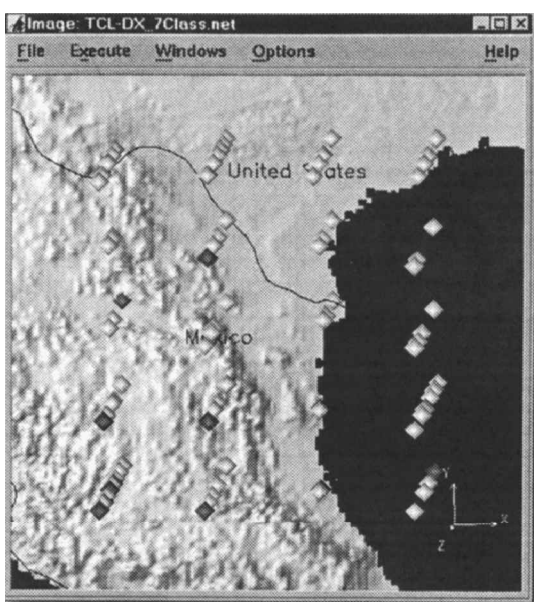

Figure 4. A geoview. In this case, the base of the view represents geographic space and political boundaries to help provide context. The z-axis represents time. Here, time is represented as linear and discrete starting with day 1 at the surface. Glyphs highlighted in red represent precipitation events at those times and places.

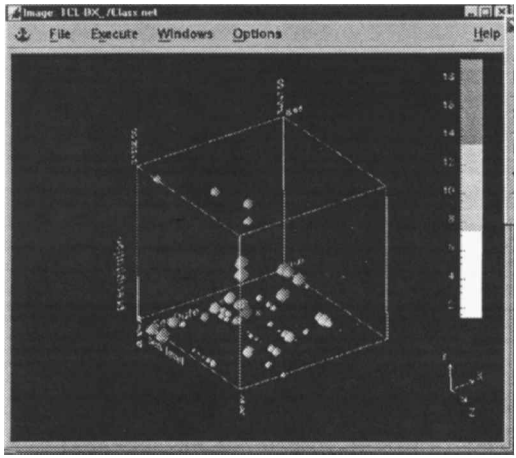

Figure 5. 3D scatterplot. In this scatterplot, samples of cases included in a data mining run are depicted in a space defined by three attributes: sequential date (number of days from the beginning of the data set), precipitation, and sea level pressure. Colour represents a fourth attribute (three classes of surface humidity) and size of 'glyphs' represent a fifth (humidity at 700 m).

axes, one for each variable in the data set (figure 6). The data are distributed along each axis and a line connects individual records from one axis to the next, producing a 'signature' for each data record. Parallel coordinate plots (PCPs) are particularly effective in uncovering relationships among many variables (Inselberg 1997). Most previous implementations, however, provided a limited perspective on data relationships because the assignment of variables

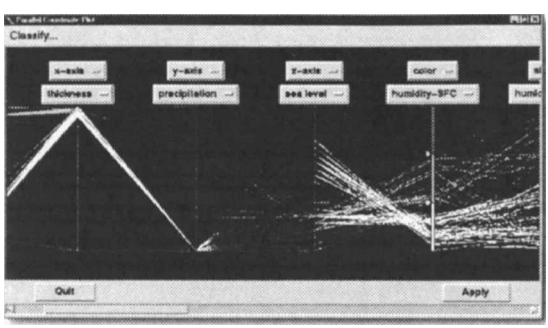

Figure 6. Parallel Coordinate Plot (PCP). This typical PCP depicts each case from a sample of data used in a data mining run by connecting the position of that case on each variable axis. The pattern of lines that results represents relationships among the classes generated in a data mining run (the initial axes of the PCP) and each of the variables included in the run (including spatial, temporal, and attribute variables). Our variation on standard parallel coordinate plots includes several additions related to user interaction (detailed below) as well as the use of colour to represent categories for one of the variable axes (in this case, the surface humidity axis is grouped into three equal range classes depicted with shades of green). Not surprisingly, the sample cases depicted that have low surface humidity (light green) all also have zero precipitation. Considering precipitation, we see that most of these zero precipitation events occur when surface atmosphere pressure is high and the high precipitation events occur when pressure is low.

to axes was fixed. In our implementation, the user can interactively adjust that assignment (see details below).

Each of our representation forms can be independently or simultaneously manipulated through applications of one or more interaction forms (figure 7). The latter can be considered the implementation of different visual analysis operations. The interaction forms we have implemented include assignment, brushing, focusing, colourmap manipulation, viewpoint perspective manipulation, and sequencing. In principle, linking of the three representation forms allows changes executed in one representation to be reflected in all of them, however, some of the interaction forms are logically applicable to only one representation form. Below we describe each of the interaction forms implemented and how they relate to each of the representation forms (see web supplement for manipulable illustrations).

— *Assignment*: Assignment allows the user to link data variables to graphic or dynamic variables of the display. An early version of this kind of assignment used to explore multi-dimensional data is Bertin's (1981) concept of matrix manipulation in which users can iteratively reassign variables to columns and rows of matrix in an effort to find relationships among variables. In our prototype, users can similarly assign variables to PCP axes (figure 8, Web only). By doing so, the user can put any two variables next to one another, or place multiple variables in a series to search for a particular signature. They can even repeat variables in order to group a key variable with multiple potentially related ones. Similarly, the 3D scatterplot allows a user to compare any three variables by assigning them to axes of the cube. In addition to assigning variables to location, users can assign variables to line colour (PCP) or to glyph size and colour (3D scatterplot). When assignment is used to match data variables to particular display positions (e.g. the axes of the scatterplots or of the PCPs), the action affects only the representation to which it is applied. When assignment is used to match data variables to other visual variables (e.g. line or glyph colour), the action is reflected in all appropriate linked representation forms.

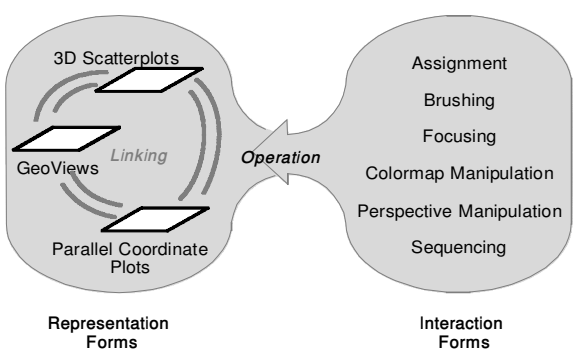

Figure 7. Integration of representation and interaction forms. Each of the three representation forms implemented (as well as others that we may implement in the future), is controlled through multiple interaction forms that allow users to manipulate various parameters to the data-to-display mapping. In most cases, linking among the representation forms results in an action applied by a user to one representation being reflected across the set of representations displayed.

— *Brushing*: Brushing is a technique that allows the user to highlight directly any set of entities in a representation that seem related to one another, appear to be outliers, or are otherwise of particular interest (figure 9). While brushing can be used in a single representation, it is particularly powerful when applied to one in a set of linked representations. For instance, by highlighting all or part of a cluster in a 3D scatterplot, a user can quickly determine where those particular records fall in geographic and/or temporal space (in the geoview) and can learn whether they have similar signatures in the PCP.

— *Focusing*: Focusing is a technique that allows a user to interactively 'focus' in on a particular range of values for a numerical variable (Buja *et al.* 1991). Rather than identifying display entities to be highlighted directly (as in brushing), focusing tools allow users to manipulate the number and definition of categories displayed (Haug *et al.* 1997). The simplest focusing method allows a user to manipulate a threshold or break point of a two class display depicted by a binary colour scheme—with all data entities having values above threshold represented by one colour or symbol and all values below the threshold represented by a contrasting colour or symbol (see MacEachren *et al.* 1993 for application to visualization of data reliability). Here we allow the user to focus on any data subset by adjusting a maximum and a minimum threshold—with data entities having values between these extremes being highlighted by a hue difference (figure 10, Web only) and allow users to manipulate any break point between adjacent categories on multi-class maps. Focusing, like brushing, tends to be particularly effective when used with linked representations.

— *Colourmap manipulation*: Colourmap manipulation involves replacing one colour scheme (to which data are mapped) with an alternative colour scheme, or replacing the colours assigned to specific data classes with alternative colours. By swapping schemes (figure 11), the user can emphasize different patterns in the data (Brewer 1997). For instance, a divergent colour scheme emphasizes deviation from a mean in two directions, while a sequential colour scheme emphasizes a singular trend. Changing the colours of a particular class can emphasize or de-emphasize that class. We have taken the approach of embedding several useful colour schemes in GKConstruct, while at the same time, allowing the user to customize those schemes. Colourmap manipulation is applied to all representations simultaneously through linking.

— *Viewpoint manipulation*: Viewpoint manipulation is implemented through tools such as panning and zooming in both two- and three-dimensional representations, or rotation in a 3D representation (figure 12, Web only). In

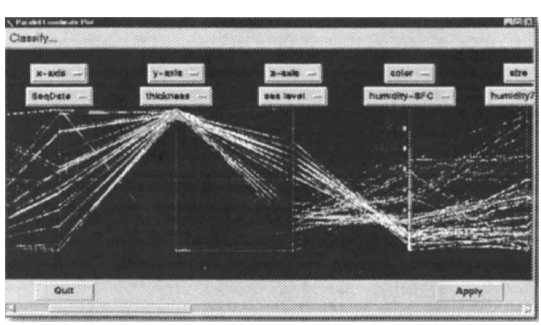

Figure 9. Brushing. Here, the PCP has been 'brushed' to highlight a group of adjacent days.

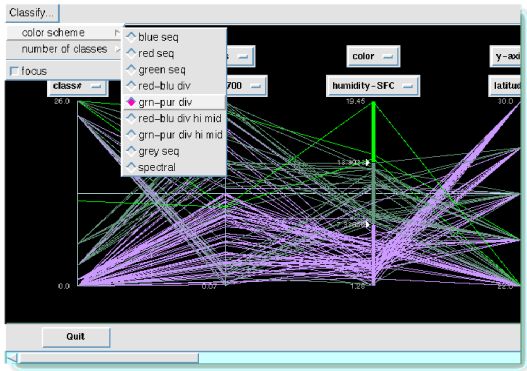

Figure 11. Colourmap manipulation. Here, the user can select among various colour maps to be applied in all views. In this case, the user has opted for a 3-step colourmap with purple representing high, grey representing medium, and green representing low.

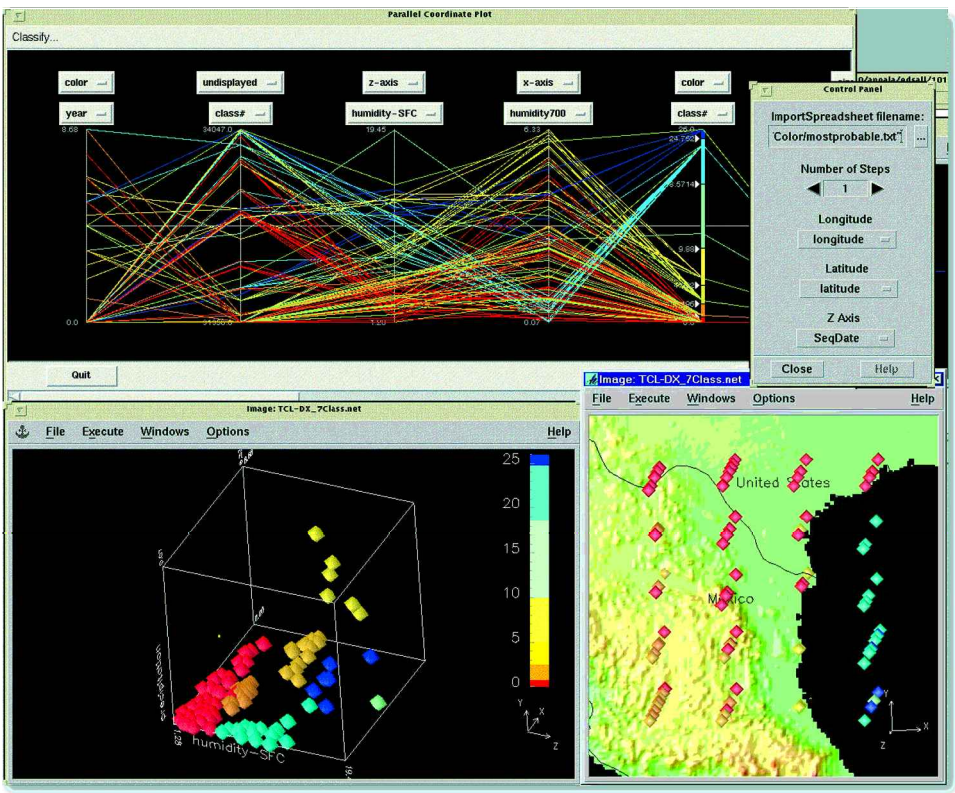

Figure 22. Here, colour hue is used to distinguish the 7 classes being explored. The distinct space, time, and attribute characteristics of the classes are clearly visible.

a 3D representation, user controlled rotation can provide depth cues that allow the user to perceive relationships that are not otherwise apparent (Kaiser and Profitt 1992). Viewpoint manipulation, as implemented here, does not effect change in linked representations.

— *Sequencing*: Sequencing, or use of the dynamic variable of *order* to display

one time slice after another, has an obvious use in facilitating a search for trends over time. Sequencing is also effective when applied to non-temporal orders (e.g. ordering views along geographical or attribute dimensions of data) (Slocum *et al.* 1990). As implemented in GKConstruct, sequencing allows the user to choose any variable dimension from the data set, sort on that variable, partition the data into equal interval bins along that dimension, then play a sequence of images depicting each bin (figure 13, Web only). Sequencing, when applied, is reflected in other linked representations as well.

Although each interaction form is described above separately, the tools are expected to be most effective when used together (figure 14, Web only). For example, a user might sequence through a variable, see an interesting pattern, and stop the sequence, sort the order in which the variables are presented in the PCP, manipulate the colourmap, and then restart the sequencer in order to see if a pattern is more apparent. In essence, these techniques all share the similar goal of facilitating pattern noticing. Whether this pattern noticing is used for feature ID, comparison, or interpretation or for understanding the data mining process, a key goal is to find relationships among features in attribute, temporal, or geographical space.

The representation and interaction forms described above have been implemented in a hybrid visualization tool building environment that consists of three parts: (1) an IBM Data Explorer (DX) program that performs data analysis and display; (2) a Tcl/Tk script that presents and manages the graphic user interface; and (3) a C program that defines the execution context and linkages between Tcl/Tk and DX (figure 15, Web only). [A more detailed explanation of the integration of these tools is provided on our web supplement: www.geovista.psu.edu/ijgis.htm.].

Data Explorer (http//www-i.almaden.ibm.com/dx/) is a general-purpose data visualization software package. It employs a dataflow client-server execution model allowing developers to author visualization applications by selecting modules of appropriate functionality from a large library, and then describing the flow of data through those modules. The functionality of the provided module library can be extended through a macro-program facility that allows new modules to be created by grouping specific combinations of library modules. Alternately, new modules can be written in C, following well-defined guidelines, and added to local DX libraries. Additionally, developers can extend DX functionality by authoring applications in C that link directly to either individual modules or entire applications in DX. In this project, a facility available in DX called DXLink has been used in conjunction with a custom C program and Tcl/Tk scripts (http//sunscript.sun.com/about/) to create a unique, appropriate, graphic user interface for controlling a visualization application authored in DX. Tcl/Tk provides a simple, pragmatic, yet elegant, development toolkit for building graphic user interfaces that can control complex processes in a visually intuitive fashion. Additionally, the rapid development cycle of graphic user interfaces using Tcl/Tk allows for easy experimentation and detailed development of the interface. Tcl (Tool Command Language) is a simple, yet powerful, platform independent scripting language (runs on Unix, Windows and Macintosh) that is easily embedded into other applications. Tk is a window system toolkit that adds the functionality of creating and manipulating very sophisticated graphical user interfaces. Tcl/Tk scripts can run standalone, be linked with C programs, or extended over the Web.

3.3.2. *Categories extraction and classification with AutoClass*

As noted above, AutoClass is a public domain software package that provides unsupervised classification based on Bayesian statistics. AutoClass III, the most

recently released version, combines real and discrete data, allows data to be missing, and automatically extracts the number of classes from a target data set. The program assumes that all attributes are relevant, that they are independent of each other within each class, and that classes are not mutually exclusive (resulting in fuzzy classes in which each case has a probability of being a member of each different class).

AutoClass is designed to search for the best class descriptions rather than just partitioning the data. In AutoClass, a *class* is defined as a particular set of parameter values and their associated model. A *classification* is a set of classes and the probabilities of each class. The classification process proceeds by first choosing an appropriate class model (or set of alternate models), then searching out good classifications based on these models. AutoClass automatically trades off the complexity of a model against its fit to the evidence. Background knowledge from an expert can be included in the input, and the output is a flexible mixture of several different 'answers' (i.e. the fuzzy classification). The main disadvantage of this approach already cited in the literature (Cheeseman 1990) is the need to be explicit about the space of models one is searching in. In our case we explicitly define the models in terms of what, where, and when.

AutoClass includes four simple models, independent and covariant versions of the multinomial model for discrete attributes and of the Gaussian normal model for real valued attributes (with minor variations). To apply a model, a set of discrete parameters (T) describing the general form of the model usually is used to specify the functional form for the likelihood function (i.e. a function giving the probability of the data conditioned on the hypothesized model and parameters). Second, free variables (V) constitute the remaining continuous parameters within a model, such as the magnitude of the correlation or relative sizes of the classes. A likelihood function, defined as $L(E|VT)$, embodies an agent's theory of how likely it would be to see each possible evidence combination E in each possible model H (an agent combines the posterior beliefs with prior expectations based on the evidence—for details, see Berger 1985). E will consist of a set of cases (i.e. an associated set of attributes, such as daily precipitation and temperature observations which can include 'unknown' values). H, denotes a hypothesis specifying that the real-world is in some particular state. Adding more parameters to the model can induce a computational cost due to the mathematical and algorithmical approximations needed for Bayesian analysis. AutoClass automatically trades off the complexity of a model against its fit to the evidence.

4. Prototype demonstration—finding features in spatiotemporal climate data

In this section, we illustrate the potential of our integrated GVis-KDD approach to knowledge construction with an application of methods to a sample gridded regional climate data set for northern Mexico and southern US. The target audience for our demonstration consists of environmental scientists (particularly climatologists). Sample data examined represent climate phenomena that are continuous in both space and time. At a conceptual level, the analysis goal is to find both individual features and classes of feature in spatiotemporal climate data sets. A secondary conceptual level goal is to explicate the data mining algorithm applied to the data (so that we can make more informed decisions about setting model parameters and so that scientists can better interpret the meaning of entity classes derived). At an operational level, these goals are instantiated as a series of operations or data processing tasks. Emphasis is on column two, cells one and two, of the meta-operations matrix introduced above

(figure 3)—on the application of feature identification and comparison methods to the KDD operation of categories extraction and classification.

As noted above, at this stage in our research we have completed only a partial integration of GVis and KDD methods, concentrating particularly on developing tools that facilitate iteration between visual exploration of data mining results and subsequent (re)application of data mining tools. The presentation of methods must, of course, be more linear than is the actual data exploration process.

4.1. *Data selection and preprocessing*

Data sets for this case study consist of daily winter climate data from 1985 to 1993 (data from Cavazos-Perez 1998). These data were selected because they are nearly exempt of noisy, incomplete, or contradictory information (due to extensive preprocessing using recognized climate data analysis methods). Hence they serve as a good test data set for demonstrating our initial steps toward the integration of GVis and KDD methods. The data were output from the Goddard Space Flight Center (GSFC) 4-D assimilation scheme based on both observational data and a Global Change Model (GCM) to produce daily gridded data at a resolution of 2° latitude by 2.5° longitude, covering an area bounded by latitudes 20–43 N, longitudes 110–90 W. The winter months (November through March) are the focus because most surface cold fronts that affect the study area occur during this season.

Data mining is concerned with drawing inferences from data, thus with understanding the patterns of correlation and casual links among the data values in a target data set. Reliable spatiotemporal data mining must include proper consideration of the fundamental spatiotemporal nature of the inference problem. Each data entity in our analysis, therefore, consists of an attribute at a specific time and place (see table 1).

4.2. *Data transformation*

The data set used for this case study consists of observational data from irregularly located sample sites collected at regular times (but with missing values) integrated with model derived values calculated for a regular grid at regular temporal spacing. While commonly used methods of climate data transformation and integration were applied, the implications for analysis using our knowledge construction system are mixed. We can be certain that there are no missing values and that the effect of any data coding blunders will be dampened considerably. On the other hand, some local anomalies or other meaningful real data aberrations could be filtered out in this process, thus potentially preventing an interesting feature from being uncovered. In subsequent research we intend to address this issue directly by comparing results for modeled data to results for less regular (less generalized) observational data.

Table 1. Defining the what, when, and where components in a target data set.

what:	precipitation, minimum temperature, maximum temperature, sea level pressure change in 24 hours, specific humidity at the surface and at 700 mb, the geopotential heights at 700 mb and 500 mb levels, and the 700–500 mb thickness.
when:	year, month, day
where:	geographical coordinates of a grid point

4.3. *Data mining*

Our main goal in this demonstration is to illustrate how our knowledge construction tools can be used to search for space-time-attribute patterns in climate data. While an initial data mining step must precede interpretation-evaluation, an effective knowledge construction session is likely to include an iterative *data mining—visual analysis—data mining—visual analysis* cycle. For this demonstration, the particular task involves extracting categories and classes from the spatiotemporal data. As noted above, we use the AutoClass software in this step. The results from AutoClass contain the following components:

— a set of classes, each of which is described by a set of parameters that specify how the class is distributed along various attributes;
— a heuristic measure of class strength, i.e. a measure of how well each class predicts its cases;
— a probabilistic assignment of cases in the data to these classes, i.e. for each case, the relative probability that it is a member of each class;
— a global influence measure, a heuristic measure that indicates overall classification quality—with values above 5.0 indicating problems such as overfitting (when the model is unjustifiably elaborate, with the models structure in part representing merely random noise in the data), underfitting (when the model is an oversimplification of reality with additional structure being needed to describe the patterns in the data) and inadequate (e.g. having the wrong structure).

4.4. *Interpretation/evaluation*

An initial AutoClass run with our sample data results in a classification with over 40 populated classes and a maximum global influence value below 5.0, thus the classification is a potentially reliable one. The integrated set of representation and interaction forms in §3 are applied here to output from each AutoClass run on our data. When measures of class strength are examined, we find that those classes with highest strength are ones associated with particular years, thus classes that exhibit temporal dominance (the *when* component is the most critical for defining the category). Typical of these classes are ones with winters revealing el Niño (very wet/raining winters) and la Niña (very dry winters) years. On the other hand, classes with intermediate strength are characterized by cyclic patterns, with spatial dominance. For example, classes in this group show the location of very wet and very dry areas in the region of the case study over the eight years. Finally, classes with a very low strength are (not unexpectedly) difficult to interpret.

Analysis of AutoClass output using GVis tools allows the user to quickly generate hundreds (if not thousands) of perspectives on the data. Thus, it is impractical (in a printed paper) to provide a comprehensive description of even a single data analysis session. Therefore, we present here a limited application of tools to one subset of data that our initial consideration of data mining results suggests may exhibit interesting patterns (the prototypical cases from the top seven classes in terms of class strength, cases from those classes having a 1.0 probability of being a member of the class to which they are assigned).

A particularly informative perspective on these data is created by focusing the PCP on a specific year. We can then use the linking among representation forms to investigate characteristics of each class represented in that year. In 1987, for example,

the eight sample observations identified are members of three classes: class 0, class 1, and class 24. By brushing (in the PCP) the lines representing each class, the query is narrowed and the prototypical 'signatures' of each class represented in 1987 can be traced. For example, the four observations in 1987 that were most likely to be in class 24 have nearly identical signatures. Each is not only a zero precipitation event but also is characterized by moderate-to-low surface humidity, low mid-level humidity, and relatively low sea-level pressure (figure 16, Web only). The 1987 observations belonging to class 0, in contrast, do not have a 'signature' that tracks as consistently through all climate variables: the lines diverge at the mid-level humidity axis, but reconverge to one particular position on the sea level pressure axis (relatively high) (figure 17, Web only). Focusing on 1992 shows the same 'trace' for class 0: no precipitation, low surface humidity, high sea level pressure, and a wide variation of 700 mb humidity values (figure 18, Web only). Class 0, then, is clearly more dependent upon sea level pressure than on mid-level humidity.

The variable to which focusing is applied can be moved from 'year' to 'class' to examine the characteristics of class 0 across all years (i.e. the full set of prototype class members). As was true for the specific years described above, observations in class 0 overall are characterized by zero precipitation, low surface humidity, high sea level pressure, and a wide range of mid-level humidity. This relationship can be confirmed using the 3D scatterplot, which shows the clustering of large glyphs (size scaled to sea level pressure) along the x-axis, which is, in this case, 700 mb humidity (figure 19, Web only).

A more dramatic result of the visualization of prototypical class 0 cases is the spatiality of this class, as displayed in the geoview (see figure 19, Web only); events in this class happen exclusively over land. This is not a surprising result, given the combination of climate variables characterizing class 0. Class 0, then, seems to be an exception to the general pattern noted in our initial overview of the data—it is a class with relatively high strength, but with a spatial rather than a temporal pattern.

Focusing on another class, class 24, we find it characterized by zero precipitation, relatively high surface humidity, and moderately low mid-level humidity and sea level pressure. This clustering is shown not only in the traces of the PCP but also in the 3D scatterplot. The spatiality of this class also shows a dramatic inverse of that of class 0: all of the instances of this class occur over the Gulf of Mexico. Again, a climatologist would expect this result: high surface humidity and lower mid-level humidity is more common over open water than over land (figure 20, Web only).

As noted above, the general pattern apparent in the full data set is that the classes with high strength are distinguished more by temporal than spatial characteristics, particularly by patterns with similar events that are proximal in time (e.g. several locations with similar attributes on the same day). Class 26, is a good example of such a class. In terms of climate variables, cases in class 26 have moderate surface humidity, low precipitation, and high mid-level humidity events. Events with this combination of characteristics are clustered temporally. The temporal clustering of class 26 is apparent if the geoview is rotated to emphasize the time axis: events belonging to class 26 occur only on certain days of the data set (figure 21, Web only).

The results of data mining can thus be visualized and interpreted effectively using our combined representation and interaction forms. As a confirmation of the above analysis, each of the seven classes in the sample data set (classes 0, 1, 3, 7, 12, 24, and 26) can be assigned a different colour (using the 7-class spectral choice in the PCP Classify menu). The clustering of these colours in the 3D scatterplot dramatically

illustrates that the observations are classified according to the values of a combination of climate variables, in tandem with spatial and temporal characteristics (figure 22—see page 325 of this article).

5. Discussion and future research

The objectives of this paper have been to make the case for integration of GVis and KDD methods, to propose a conceptual framework for that integration emphasizing a merger of meta operations fundamental to each set of methods, and to describe an initial prototype knowledge construction environment and its application to a test data set. At this stage in our long term strategy for GVis-KDD integration, we have applied our knowledge construction methods to an isolated data set stored as flat files and focused on integration of GVis in the final two KDD stages. We plan a subsequent coupling of GVis-KDD methods and temporal GIS. One goal of this coupling is to make the early stages in the knowledge construction process more flexible and facilitate interactive exploration of user selected subsets of data (choice of which is prompted by prior analysis steps). We expect that many of the GVis methods developed here will be useful for applications at the earlier KDD stages.

At the data selection stage, visualization tools might be used to quickly recognize common or mismatched spatial or temporal coverage in variables or develop an understanding of the potential extent, resolution, quality, cost, or other attributes of available data. At the preprocessing stage, visualization may be particularly important because visualization has proven to be useful in finding holes or errors in data sets. Data mining efforts seem to be quite sensitive to such holes or errors, thus it is important to locate and correct them at this stage. Visualization methods can also facilitate user understanding of parameter setting for various transformations that might be applied to reduce data complexity (e.g. parameters of interpolation algorithms that transform point samples to a regular grid) and for understanding the implications of those transformations—see Edsall *et al.* (in press) for an example related to fourier transformations.

At the data mining stage, visualization has several potential roles beyond those highlighted above. Visual display of raw data for each variable in the analysis can facilitate decisions on appropriate model representation. In addition, visualization in the form of 'process tracking' (visual displays that represent key aspects of a process as it unfolds) can help domain specialists (who are unlikely to be experts in the database and statistical techniques being used in the mining process) to understand these techniques and their limitations. There also is a potential to use visualization for 'process steering' (controlling parameters of the data mining process as it unfolds, thus changing outcomes on the fly). We are currently exploring the potential of parallel computing to achieve the processing speed needed for this application.

As Cheeseman and Stutz (1996) point out, 'It is the interaction between domain experts and the machine, searching over the model space, that generates new knowledge. Both bring unique information and abilities to the database analysis task, and each enhances the others' effectiveness'. Certainly, the value of human insight, intuition and imagination in this process cannot be overestimated. It is our contention that successful applications of KDD will be defined by the strength of the graphic user interface and the GVis methods it supports—thus by the ability of these tools to provide a gateway to both human information processing abilities and knowledge discovery software.

Acknowledgments

This research was supported by the US Environmental Protection Agency (EPA), under Grant R825195-01-0 (Donna J. Peuquet and Alan M. MacEachren, Co-PIs). Support has also been provided by the Penn State Center for Academic Computing where MacEachren is a Faculty Fellow. We thank Tereza Cavazos for providing the Mexico data used in our demonstration and for help in interpretation of the data mining output. Mark Harrower's work on both print graphics and on design and production of the paper's Web supplement is greatly appreciated.

References

ARMSTRONG, M. P., DENSHAM, P. J., LOLONIS, P., and RUSHTON, G., 1992, Cartographic displays to support locational decision making. *Cartography and Geographic Information Systems*, **19**, 154–164.

BERGER, J. O., 1985, *Statistical Decision Theory and Bayesian Analysis* (New York: Springer-Verlag).

BERTIN, J., 1981, *Graphics and Graphic Information Processing* (Berlin: Walter de Gruyter).

BRACHMAN, R. J., and ANAND, T., 1996, The process of knowledge discovery in databases. In *Knowledge Discovery and Data Mining*, edited by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Menlo Park, CA: AAAI Press/The MIT Press), pp. 37–57.

BREWER, C. A., 1997, Spectral schemes: controversial color use on maps. *Cartography and Geographic Information Systems*, **24**, 203–220.

BUJA, A., MCDONALD, J. A., MICHALAK, J., and STUETZLE, W., 1991, Interactive data visualization focusing and linking. In *Proceedings, Visualization '91, IEEE Conference on Visualization* (San Diego, CA: IEEE), pp. 156–163.

BUTTENFIELD, B. P., and MACKANESS, W. A., 1991, Visualization. In *Geographical Information Systems: Principles and Applications*, edited by D. Macguire, M. Goodchild and D. Rhind (New York: Wiley), pp. 427–444.

CARR, D. B., OLSEN, A. R., COURBOIS, J. P., PIERSON, S. M., and CARR, D. A., 1998, Linked micromap plots: names and described. *Statistical Computing and Graphics Newsletter*, **9**, 24–32.

CAVAZOS-PEREZ, T., 1998, Downscaling Large-Scale Circulation to Local Winter Climate Using Neural Network Techniques, Ph.D. Thesis, Department of Geography, The Pennsylvania State University, University Park, PA.

CHEESEMAN, P., 1990, On finding the most probable model. In *Computational Models of Discovery and Theory Formation*, edited by J. A. L. Shrager (Palo Alto, CA: Morgan Kaufmann), pp. 73–96.

CHEESEMAN, P., and STUTZ, J., 1996, Bayesian Classification: Theory and Results. In *Advances in Knowledge Discovery and Data Mining*, edited by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Menlo Park, CA: AAAI Press/The MIT Press), pp. 153–189.

CHEN, M., HAN, J., and YU, P. S., 1996, Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 866–883.

COOK, D., SYMANZIK, J., MAJURE, J. J., and CRESSIER, N., 1997, Dynamic graphics in a GIS: more examples using linked software. *Computers and Geosciences*, special issue on Exploratory Cartographic Visualization, **23**, 371–386.

DIBIASE, D., 1990, Visualization in the Earth sciences. *Earth and Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences, Penn State University*, **59**, 13–18.

DORLING, D., 1992, Stretching space and splicing time: From cartographic animation to interactive visualization. *Cartography and Geographic Information Systems*, **19**, 215–227.

DYKES, J., 1997, Exploring spatial data representation with dynamic graphics. *Computers and Geosciences*, special issue on Exploratory Cartographic Visualization, **23**, 345–370.

EDSALL, R. M., HARROWER, M., and MENNIS, J., In Press, Visualizing properties of temporal and spatial periodicity in geographic data. *Computers and Geosciences*.

ESTER, M., KRIEGEL, H. P., and XU, X., 1995, Knowledge discovery in large spatial databases:

Focusing techniques for efficient class identification. In *Advances in Spatial Databases: Proceedings International Symposium on Large Spatial Databases (SSD'95)*, edited by Max J. Egenhofer and John R. Herring (Portland Maine: Berlin: Springer-Verlag), pp. 67–82.

FAYYAD, U., PIATETSKY-SHAPIRO, G., and SMYTH, P., 1996, From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*, edited by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Menlo Park, CA: AAAI Press/The MIT Press), pp. 1–34.

FRANK, A. U., and TIMPF, S., 1994, Multiple representations for cartographic objects in a multi-scale tree—an intelligent graphical zoom. *Computers and Graphics; Special Issue on Modelling and Visualization of Spatial Data in GIS*, **18**, 823–829.

FRAWLEY, W. J., PIATETSKY-SHAPIRO, G., MATHEUS, C. J., and SMYTH, P., 1991, Knowledge discovery in databases: An overview. In *Knowledge Discovery and Data Mining*, edited by G. Piatetsky-Shapiro and B. Frawley (Cambridge, Mass: AAAI Press/The MIT Press), pp. 1–27.

GRAY, J., CHAUDHURI, S., BOSWORTH, A., LAYMAN, A., REICHART, D., and VENKATRAO, M., 1997, Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery*, **1**, 29–53.

HAN, J., 1995, Mining knowledge at multiple concept levels. In *Proceedings 4th International Conference on Information and Knowledge Management*, Baltimore, Maryland, November 1995, pp. 19–24.

HAN, J., and FU, Y., 1996, Exploration of the power of attribute-oriented induction in data mining. In *Advances in Knowledge Discovery and Data Mining*, edited by U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Menlo Park, CA: AAAI Press/ The MIT Press), pp. 399–421.

HAN, J., KOPERSKI, K., and STEFANOVIC, N., 1997, GeoMiner: A system prototype for spatial mining. In *Proceedings, 1997 ACM-SIGMOD International Conference on Management of Data (SIGMOD'97)*, Tucson, AZ, (New York: ACM Press), pp. 553–556.

HARINARAYAN, V., RAJARAMAN, A., and ULLMAN, J. D., 1996, Implementing data cubes efficiently. In *Proceedings of ACM-SIGMOD on Management of Data, Montreal, Canada*, New York: ACM Press, pp. 205–216.

HAUG, D. B., MACEACHREN, A. M., BOSCOE, F. P., BROWN, D., MARRARA, M., POLSKY, C. H., and BEEDASY, J., 1997, Implementing exploratory spatial data analysis methods for multivariate health statistics. In *Proceedings of GIS/LIS '97 Annual Conference and Exposition, 28–30 October, 1997*. Cincinnati, Ohio. Wash. DC: AAG, pp. 205–213.

HOSKING, J. R. M., PEDNAULT, E. P. D., and SUDAN, M., 1997, A statistical perspective on data mining. *Future Generation Computer Systems*, **13**, 117–134.

HOWARD, D., and MACEACHREN, A. M., 1996, Interface design for geographic visualization: Tools for representing reliability. *Cartographic and Geographic Information Systems*, **23**, 59–77.

INSELBERG, A., 1997, Multidimensional detective. *IEEE Visualization '97, 19–24 October, Phoenix, AZ*, Los Alamitos, CA: IEEE Computer Society (IEEE Computer Society Technical Committee on Computer Graphics), pp. 100–107.

JPL, 1987, *LA: the movie* (Pasadena: Jet Propulsion Laboratory).

KAISER, M. K., and PROFFITT, D. R., 1992, Using the stereokinetic effect to convey depth: Computationally efficient depth-from-motion displays. *Human Factors*, **34**, 571–581.

KAUFMANN, W. J., and SMARR, L. L., 1993, The emergence of a digital science. In *Supercomputing and the Transformation of Science* (New York, Scientific American Library), pp. 1–23.

KOPERSKI, K., HAN, J., and ADHIKARY, J., in press, Mining knowledge in geographical data. *Communications of ACM*.

KRAAK, M.-J., 1998, Exploratory cartography, maps as tools for discovery. *ITC Journal*, **1**, 46–54.

MACEACHREN, A. M., 1994, Visualization in modern cartography: Setting the Agenda. In *Visualization in Modern Cartography*, edited by A. M. MacEachren and D. R. F. Taylor (Oxford, UK: Pergamon), pp. 1–12.

MACEACHREN, A. M., 1995, *How Maps Work: Representation, Visualization and Design* (New York: Guilford Press).

MacEachren, A. M., (in collaboration with B. Buttenfield, J. Campbell, D. DiBiase and M. Monmonier), 1992, Visualization. In *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography*, edited by R. Abler, M. Marcus and J. Olson (New Brunswick, NJ: Rutgers University Press), pp. 99–137.

MacEachren, A. M., and Ganter, J. H., 1990, A pattern identification approach to cartographic visualization. *Cartographica*, **27**, 64–81.

MacEachren, A. M., Howard, D., von Wyss, M., Askov, D., and Taormino, T., 1993, Visualizing the health of Chesapeake Bay: An uncertain endeavor. In *Proceedings, GIS/LIS '93, Minneapolis, MN, 2–4 November, 1993* (Madison: ACSM & ASPRS), pp. 449–458.

MacEachren, A. M., and Kraak, M.-J., 1997, Exploratory cartographic visualization: Advancing the agenda. *Computers and Geosciences*, **23**, 335–343.

Mitas, L., Brown, W. M., and Mitasova, H., 1997, Role of dynamic cartography in simulations of landscape processes based on multivariate fields. *Computers & Geosciences*, **23**, 437–446.

Monmonier, M., 1992, Authoring graphics scripts: experiences and principles. *Cartography and Geographic Information Systems*, **19**, 247–260.

Ng, R., and Han, J., 1994, Efficient and effective clustering method for spatial data mining. *Proceedings International Conference on VLDB, Santiago, Chile*, September 12–15, Wash. DC: AAAI Press, pp. 144–155.

Openshaw, S., 1994, Computational human geography: Towards a research agenda. *Environment and Planning A*, **26**, 499–508.

Robertson, G. G., 1991, Cone trees: Animated 3D visualization of hierarchical information. In *Proceedings, ACM CHI'91* (New Orleans, LA), **April**, pp. 189–194.

Simoudis, E., Livezey, B., and Kerber, R., 1996, Integrating inductive and deductive reasoning for data mining. In *Advances in Knowledge Discovery and Data Mining*, edited by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Menlo Park, CA: AAAI Press/The MIT Press), pp. 353–374.

Skupin, A., and Buttenfield, B. P., 1996, Spatial metaphors for visualizing very large data archives. *GIS/LIS '96, Denver, Colorado*, November 19–21 (Madison: ACSM/ASPRS), pp. 607–617.

Slocum, T. A., Robertson, S. H., and Egbert, S. L., 1990, Traditional versus sequenced choropleth maps: An experimental investigation. *Cartographica*, **27**, 67–88.

Slortz, P., Nakamura, H., Mesrobian, E., Muntz, R. R., Shek, C., Santos, J. R., Yi, J., Ng, K., Chien, S., Mechoso, C. R., and Farrara, J. D., 1995, Fast spatiotemporal data mining of large geophysical data sets. In *Proceedings of the First Conference on Knowledge Discovery and Data Mining, Montreal, Canada*, Wash. DC: AAAI Press, pp. 300–305.

Treinish, L., 1993, Visualization techniques for correlative data analysis in the earth and space sciences. In *Animation and Scientific Visualization: Tools and Applications*, edited by R. A. Earnshaw and D. Watson (New York: Academic Press), pp. 193–204.

Uthurusamy, R., 1996, From data mining to knowledge discovery: Current challenges and future directions. In *Advances in Knowledge Discovery and Data Mining*, edited by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Menlo Park, CA: AAAI Press/The MIT Press), pp. 561–572.

Wang, W., Yang, J., and Muntz, R., 1997, STINGA: Statistical information grid approach to spatial data mining. In *Proceedings of the 23rd VLDB Conference, Athens, Greece, August*, New York: ACM Press, pp. 186–196.

Wilbanks, T., Adams, R., Church, M., Clark, W., deSouza, A., Gilmartin, P., Graf, W., Harrington, J., Horn, S., Kates, R., MacEachren, A., Murphey, A., Rushton, G., Sheppard, E., Turner, B., and Willmott, C., 1997, *Rediscovering Geography: New relevance for science and society* (Washington, DC: National Academy of Sciences).

Wood, M., 1994, Visualization in historical context. In *Visualization in Modern Cartography*, edited by A. M. MacEachren and D. R. F. Taylor (London: Pergamon), pp. 000–000.

Zhang, T., Ramakrishnan, R., and Linvy, M., 1996, BIRCH an efficient data clustering method for very large databases. *Proceedings ACM-SIGMOD on Management of Data, Montreal, Canada, June*, New York: ACM Press, pp. 103–114.