

Research Article

The Role of Knowledge Representation in Geographic Knowledge Discovery: A Case Study

Jeremy Mennis
Department of Geography
University of Colorado, Boulder

Donna J Peuquet
Department of Geography
The Pennsylvania State University

Abstract

With the advent of massive, heterogeneous geographic datasets, data mining and knowledge discovery in databases (KDD) have become important tools in deriving meaningful information from these data. In this paper, we discuss how knowledge representation can be employed to significantly enhance the power of the knowledge discovery process to uncover patterns and relationships. We suggest that geographic data models that support knowledge discovery must represent both observational data and derived knowledge. In addition, knowledge representation in the context of KDD must support the iterative and interactive nature of the knowledge discovery process to enable the analyst to iteratively apply, and revise the parameters of, specific analytical techniques. Our approach to knowledge representation and discovery is demonstrated through a case study that focuses on the identification and analysis of storms and other related climate phenomena embedded within a spatio-temporal data set of meteorological observations.

1 Introduction

Advances in data gathering and data sharing technologies have made massive geographic datasets readily available over the past decade. Because of the sheer size and complexity of these data, advanced analytical approaches, such as those associated with data mining and knowledge discovery in databases (KDD), are becoming essential tools in deriving useful information from these data for decision support, environmental modeling, and other applications (Openshaw 2000, Ramachandran et al. 2000).

Fayyad et al. (1996) define KDD as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data.” The

Address for correspondence: Jeremy Mennis, Department of Geography, University of Colorado, Boulder, CO 80309-0260. E-mail: jeremy@colorado.edu

knowledge discovery process involves a number of steps, including data preparation, user interaction, and iterative refinement of analytical methodologies. The term data mining refers to the specific step within this process where patterns and associations in the data are derived through the use of algorithmic procedures. Data mining techniques can be categorized into four general types based on the type of pattern they attempt to extract: association rules, data characterization, classification, and clustering (Chen et al. 1996). Data mining techniques can also be categorized along an axis extending between deductive (expert knowledge-driven) and inductive (data pattern-driven) poles (Gahegan 2000). Ideally, a variety of data mining methods (algorithms) are available within a KDD software environment.

Although most research in KDD and data mining has focused on non-spatial data, there has been recent progress in the development of geographic KDD environments (Han et al. 1997, Gahegan et al. 2002, Peuquet and Kraak 2002) and spatial data mining techniques (Guo et al. 2002, Roddick and Hornsby 2000, Miller and Han 2001). If efforts in applying KDD to geographic data and problems are to be successful, however, research in KDD must incorporate aspects of knowledge representation. The integration of geographic knowledge within KDD environments is necessitated by the complexity of geographic domains and the subsequent need for domain-specific knowledge as a means of guiding the knowledge discovery process (Yuan et al. 2001). In addition, because the knowledge discovery process is intended to be interactive (Fayyad et al. 1996), geographic knowledge representation within KDD must support interaction with, and the iterative refinement of, that knowledge. While there are a variety of approaches to knowledge representation that have been developed in the fields of database modeling and artificial intelligence (cf. Sowa 2000), the issue of how to store and integrate geographic knowledge into the knowledge discovery process remains a challenge.

The purpose of this paper is to discuss how geographic knowledge can be incorporated within an interactive KDD environment. We propose an approach for integrating geographic knowledge representation within the knowledge discovery process and present a prototype implementation of a KDD environment that demonstrates this approach. In order to demonstrate this implementation, we describe a case study that focuses on the identification and analysis of storms and other related climate phenomena from a spatio-temporal data set of meteorological observations.

2 Knowledge Representation in KDD

In the context of KDD, data are viewed as empirical measurements or observations. As such, they do not carry meaning, *per se*. Knowledge, on the other hand, consists of meaningful characteristics or generalized behavioral rules concerning the domain under investigation. This can also be called the semantics of the data. We consider two types of knowledge that may be represented within a KDD environment. The first type is knowledge derived from the knowledge discovery process; the second is a priori knowledge that is brought to bear on the knowledge discovery process by a domain expert/user. There has been little research on incorporating knowledge, whether derived from KDD or defined externally by experts, within the knowledge discovery process. Regarding the use of knowledge derived from KDD, a few researchers have proposed mining the results of previous data mining operations (Schoenauer and Sebag 1990, Spiliopoulou and Roddick 2000). For instance, Spiliopoulou and Roddick (2000) suggest that association

rules derived from temporal data can be further mined to produce 'meta-rules' that describe how the initial, data-derived association rules vary over time.

Externally defined expert knowledge may be used as rules within a data mining algorithm to tune parameters and improve the efficiency of data mining, or to ensure that the KDD results are meaningful by minimizing the derivation of coincidental groupings and associations. One area of spatial data mining research that has incorporated the use of externally defined expert knowledge is image segmentation. Studies in this area have sought to improve image classification accuracy and efficiency as well as to automate the extraction of semantic content from medical and satellite-based imagery (Ton et al. 1991, Ezquerro and Mullick 1996, Sonka et al. 1996, Zhang et al. 2002). In these studies, data mining typically takes place via an initial unsupervised classification of pixels, which is then refined through the application of expert knowledge of the domain. This knowledge may be input to the data mining algorithm via user interaction or through the use of pre-programmed (i.e. hardwired) production rules. This area of research is focused primarily on the automation of image interpretation using data mining techniques, and has generally not been extended to support the discovery of new knowledge from imagery.

The effective integration and use of knowledge within KDD demands explicit database support not only for data but also for knowledge, whether that knowledge is derived from the knowledge discovery process itself or is derived from experts and stored a priori in the database. We suggest that research in cognition may inform the development of database support for both data and knowledge representation in KDD. The high degree of interdependency between data and knowledge in KDD is analogous to the distinction between sensory input and conceptual knowledge in the context of cognitive knowledge acquisition. Cognitively, people use stored knowledge to interpret sensory information (e.g. sights and sounds) in everyday circumstances and gradually acquire new conceptual knowledge through this interactive process (Peuquet 2002). The process of discovering knowledge from observational data in a computational environment is also typically highly interactive, bringing previously derived knowledge to bear in interpreting and subsequently understanding new observations (Fayyad et al. 1996, Mennis et al. 2000). Existing knowledge can then be augmented or modified to reflect any new insights, and the process continued. Moreover, the interaction of the human and the computer in this process takes advantage of the computational speed, mechanical accuracy, objectivity and persistent memory of the computer and the human's intuitive powers and ability to bring knowledge from disparate knowledge domains to bear on a problem (Peuquet 2002).

Knowledge representation in KDD should support the iterative and interactive nature of the knowledge discovery process to enable the analyst to iteratively apply, and revise the parameters of, specific data mining techniques. Knowledge representation in the context of KDD should also be 'flexible' in the sense that the KDD system user should be able to revise and interpret stored and derived knowledge interactively. As an approach to this interactivity, Imielinski and Mannila (1996) have suggested that users of KDD environments should be able to access knowledge elements (e.g. classification rules or the results of a cluster analysis) directly as database objects. Such an approach assumes, of course, that there exist data structures to explicitly store these knowledge elements persistently in the database.

Over the past decade, researchers in geographic information science have focused significant attention on improving the representational power of geographic data models

beyond the conventional vector and raster models used in current geographic information systems (GIS) (Egenhofer et al. 1999). Many of these advances in geographic data modeling have focused on improving semantic representation in order to represent observational data, knowledge that may be derived from those data, and the mapping between data and knowledge (Leung et al. 1999, Mennis 2003). Here, we leverage this previous geographic data modeling research to develop database support for both data and knowledge and thus integrate geographic knowledge representation into KDD.

3 An Approach for Combined Data and Knowledge Representation

In previous research we have described the derivation and implementation of a semantic spatio-temporal data model based on the principles of human cognition (Mennis et al. 2000, Mennis 2003). This semantic data model incorporates many aspects of well-known knowledge representation strategies as discussed in the computer science literature (e.g. Minsky 1975, Booch 1994, Sowa 2000), including object-oriented modeling, frames, and rules. In the current research, this model is used as the basis for implementing the prototype KDD environment used in the case study. The semantic data model and KDD environment are associated with a larger software development project called Apoala (see www.geovista.psu.edu/grants/apoala for additional details).

The semantic spatio-temporal data model is implemented using the object-oriented database platform *Poet* (Poet, Inc.) and the Java programming language, and is divided into two primary parts: the Data Component and Knowledge Component (Figure 1). The Data Component is intended to store observational data, represented by the class *ATTVALUE* that stores an observed thematic value and a spatio-temporal reference for that measurement. The Knowledge Component is intended to store information on semantic categories and entities, represented by the classes *CATEGORY* and *THING*, respectively.

The data mining technique that is supported by the data model implementation is a deductive, classification-based algorithm that performs feature extraction. It is similar to the previously discussed use of expert knowledge for image segmentation (Ton et al. 1991), and to what has been described as ‘template-based mining’ because one uses a template of a pattern to find instantiations of that pattern in the data (Roddick and Spiliopoulou 2002). In our approach, expert knowledge is encoded as a set of rules that are stored in a *CATEGORY* object. These rules are intended as generic characterizations of objects in a particular domain and are not limited to a specific image or data set. These rules are used to construct a query on the Data Component using Poet’s implementation of the Object Query Language (OQL). This process distinguishes between first- and second-order properties. First-order properties address the composition of an entity while second-order properties address the descriptive attributes of an entity.

Two types of properties, called TProp and CProp properties, are predefined to allow storage of general properties of any given entity and category, respectively. While these properties are predefined within the stored knowledge base, they may or may not be used, depending on the specific observational data being examined. Examples of TProp properties include *BIRTH*, when a *THING* object begins its existence, and *LIFESPAN*, the duration of a *THING* object’s existence. CProp properties include those that describe the criteria for membership within a *CATEGORY*. For instance, while a *THING* object may have an observed *LIFESPAN*, a *CATEGORY* has a

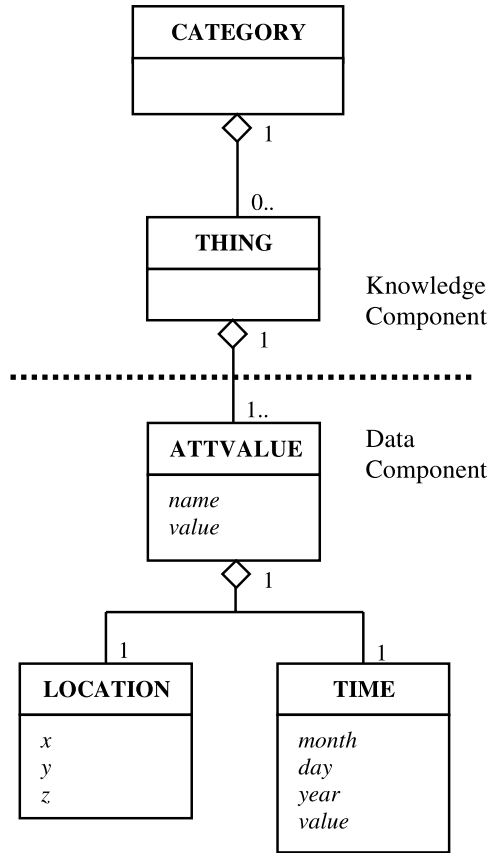


Figure 1 Unified Modeling Language (UML) diagram of the implementation of the semantic data model in Java (Mennis 2003). Note that class methods, and certain class attributes, are not shown for simplicity

LIFESPANRANGE (a maximum and minimum lifespan). These predefined properties may be extended through specialization in order to customize properties in a way that is most suitable for a given application domain.

4 Case Study: Representing and Extracting Storms from a Meteorological Data Set

4.1 Background

The application domain used for this case study concerns the representation and analysis of storms and related meteorological phenomena in the Susquehanna River Basin (SRB), which extends 70,448 km² throughout central Pennsylvania, southern New York, and a small portion of northern Maryland where the Susquehanna River empties into northern Chesapeake Bay. For the purpose of representing these meteorological phenomena, an observational data set consisting of daily maximum temperature, daily

minimum temperature, and daily total precipitation for the SRB was acquired from the Environment Institute at the Pennsylvania State University. Each day's observations were stored as a 97×91 cell, four kilometer resolution grid, generated from an inverse distance weighted interpolation of National Climate Data Center (1995) weather station data. The author of the data set checked for errors and logical inconsistencies in the data (e.g. negative precipitation values, missing values, etc.) and adjusted the two temperature variables for changes in elevation.

A time period of three months (June, July, and August 1969) was used for this case study. Summertime data were chosen in order to capture the presence of local, convective thunderstorms that typically do not occur during winter. These thunderstorms may be contrasted with larger, longer-lasting storms that occur year round in order to represent a variety of storm types. While a number of different years of summertime data were available, a visual review of the data showed that the summer of 1969 included a multiplicity of spatio-temporal patterns of precipitation that could be categorized as different storm types.

We used expert knowledge derived from the meteorological literature (Weisman and Klemp 1986, Carleton 1991) to create a typology of storm types that may be identified from the observational data. The typology distinguishes primarily between local storms, those brief, isolated storms that occur in small areas, and regional storms, those larger storms that covered a significant portion of the SRB. A secondary distinction is made between severe storms, those that are distinguished by a comparatively large amount of precipitation, and mild storms, those with a lesser amount of precipitation.

In addition, the distinctions between different types of storms and the representations of individual storms are necessarily constrained by the spatial and temporal resolution of the observational data. For this reason, no individual storm can be represented as being less than one day in duration, since the observational data have a daily temporal resolution. While we acknowledge that this data resolution may not capture small or brief storms, we have developed a storm typology that takes the spatial and temporal resolution constraints of the observational data into account. While it would certainly be useful to have observational data with a finer temporal resolution (in grid form), and thus be able to support the extraction of other mesoscale storm types that are commonly addressed in climate analysis (e.g. squall lines), the available data are sufficient for our purpose here of demonstrating the role of knowledge representation in geographic knowledge discovery.

4.2 Storm Representation

The typology of storms is represented in the database by extending the CATEGORY class to create a hierarchy of storm types (Figure 2). We first define a generic category of storm, STORMCATEGORY, by extending the CATEGORY class. We then distinguish between local and regional storms (LOCAL and REGIONAL classes, respectively). Each local or regional storm may also be categorized as either mild or severe (e.g. LOCALMILD and LOCALSEVERE, respectively). Thus, each storm that is extracted from the observational data is categorized according to one of the storm types contained within this hierarchy.

The STORMCATEGORY class includes those properties that are deemed essential for representing a type of storm:

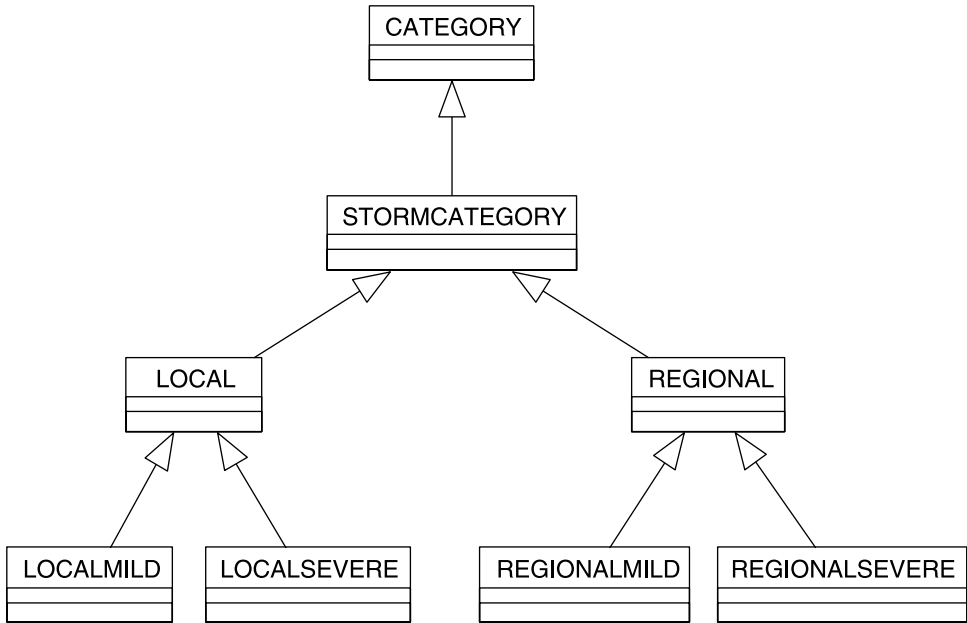


Figure 2 UML overview diagram of the classes that represent the storm taxonomy, derived from the CATEGORY class

- LIFESPANRANGE range in lifespan
- SIZERANGE range in size
- PRCPVALRANGE range in precipitation value
- AVGPRCPRANGE range in average precipitation value of a storm
- MAXPRCPRANGE range in maximum precipitation value of a storm

Note that some of these properties, such as LIFESPANRANGE, are taken directly from the CProp library. Other properties, such as PRCPVALRANGE, are extended from more ‘generic’ CProp properties to capture a particular aspect of the application domain. All the classes that are subcategories of the STORMCATEGORY class inherit these properties, although the properties are assigned different values when, say, the LOCALMILD class is instantiated versus the REGIONALSEVERE class.

This storm typology bears some resemblance to the concept hierarchy idea associated with data characterization-based, data mining techniques implemented using on-line analytical processing (OLAP) data cubes (Shoshani 1997) and attribute-oriented induction (Han et al. 1993). The storm typology is similar to a concept hierarchy in that both are used for classification. However, the concept hierarchy is a classification of data and consists of simply setting the upper and lower bounds of a range of data values. The storm typology, on the other hand, is a classification of semantic entities (i.e. storms) that are the results of a previous feature extraction, data mining operation.

Analogous to how the CATEGORY class is extended to represent different types of storms, the THING class is extended to the STORMTHING class to represent an observed, actual storm. The STORMTHING class has a number of property classes that

are either added as a component of the class directly from the TProp library or extended from one of those classes. These properties include:

- BIRTH when a storm begins its existence
- DEATH when a storm ends its existence
- LIFESPAN the duration of a storm's existence
- SIZEMIN the minimum size of a storm during its existence
- SIZEMAX the maximum size of a storm during its existence
- SIZEAVG the average size of a storm during its existence
- MAXPRCP the maximum precipitation value observed within a storm during its existence
- AVGPRCP the average precipitation value observed within a storm during its existence

Note that characteristics of size, duration, and severity are deemed important for all observed storms, regardless of type, even though the observed values for each of the properties obviously differ among individual storms. The STORMTHING class also contains a method to calculate the value for each of its own properties through an analysis of its instantiation in the Data Component. For example, the value for the property BIRTH is calculated by iterating through a given STORMTHING object's collection of ATTVALUE objects to find the earliest stored time reference.

4.3 A Demonstration of Storm Extraction and Analysis

Here, we present an example of how the user may interact with the KDD environment in order to extract and analyze storms from the meteorological data. The nature of this interaction is presented schematically in Figure 3. At the top of the diagram is the

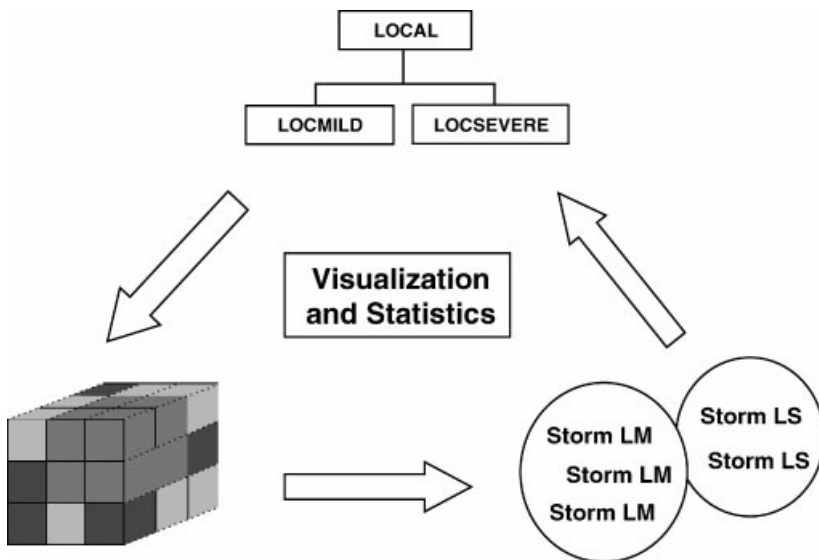


Figure 3 Cycle of knowledge discovery for knowledge representation and feature extraction demonstrated in the case study

categorical hierarchy (i.e. taxonomy) of different storm types and their stored properties. This expert knowledge (i.e. part of the Knowledge Component of the database) is used to interpret the observational meteorological data stored in the Data Component, represented by the cube in the lower left of the diagram. The result of this interpretation (through an automated feature extraction algorithm) is the identification of a set of storm entities within the observational data, i.e. specific storms of a given type, that can then be subsequently added to the Knowledge Component of the database as derived knowledge. Visualization, as well as other algorithmic/statistical methods, may be used to verify the results of this analysis and the categorical properties may be revised, leading to a new interpretation of the observational data, new verification procedures, and so on.

Note that this process of user interaction with the KDD environment is intended to support the analysis of knowledge derived from the knowledge discovery process itself, as is suggested by Spiliopoulou and Roddick (2000). As a first step, the user mines the precipitation data to extract individual storms. As a second step, the user may apply further mining tools, such as visualization and statistical techniques, to detect patterns among those extracted storms. This approach thus supports the investigation of knowledge not just about patterns embedded in the precipitation data, but about patterns evident among the results of previous data mining operations, i.e. among the storms.

The feature extraction process is a three step procedure implemented through a set of methods contained in the CATEGORY class and then customized for storm feature extraction in the STORMCATEGORY class. The first step retrieves those ATTVALUE objects that meet the first-order property criteria of a STORMCATEGORY object, the presence of precipitation, which is specified in the PRCPVALRANGE property. The second step organizes the collection of ATTVALUE objects retrieved from the first step into spatially and temporally contiguous groups which serve as temporary, 'candidate' STORMTHING objects. The user can interactively specify the nature of how spatially and temporally contiguous ATTVALUE objects are grouped into candidate STORMTHING objects by setting the contiguity constraint, a parameter that controls the distance that the algorithm searches in the three spatial dimensions (x , y , and z) and one temporal dimension (t) in order to define contiguity between two ATTVALUE objects. Note that in this case study the observational data do not extend in the z dimension (they do not have an elevation or altitude reference); thus the contiguity constraint settings apply only to the x , y , and t dimensions. Once these candidate STORMTHING objects are identified, their second-order properties of BIRTH, LIFESPAN, etc. are calculated. The third step in the feature extraction process compares the second-order properties of the candidate STORMTHING objects (e.g. LIFESPAN) to the second-order properties of the STORMCATEGORY class (e.g. LIFESPAN-RANGE) to determine whether a particular candidate STORMTHING object may be considered a member of a particular type of storm.

As an example, consider first the extraction of local, severe storms from the observational data. Table 1 shows the instantiated properties, called a 'property template,' for the LOCALSEVERE storm type. A 'null' value listed in Table 1 indicates there is no threshold maximum or minimum value specified for that particular property. The values stored within the PRCPVALRANGE property are used to identify those observations in the meteorological data where there is the (first-order property) presence of precipitation (precipitation ≥ 0.01 inches). Candidate storms, i.e. regions of precipitation that are contiguous in space and time, are then identified using the contiguity constraints

Table 1 Property template for the LOCALSEVERE storm type

	Maximum Value	Minimum Value
LIFESPANRANGE (Days)	1	1
SIZERANGE (km ²)	3,200	80
PRCPVALRANGE (inches)	null	0.01
MAXPRCPRANGE (inches)	null	0.5
AVGPCRPRANGE (inches)	null	null
Contiguity constraints (x, y, z, t) = (1, 1, 0, 0)		

Table 2 LOCALSEVERE storms extracted using the property template shown in Table 1

Month	Day	Size (km ²)	MinPrpc (in)	MaxPrpc (in)	AvgPrpc (in)
6	8	3,136	0.01	0.51	0.14
7	9	2,992	0.01	0.79	0.17
7	17	448	0.01	0.65	0.26
7	31	2,720	0.01	0.64	0.13
8	15	2,048	0.01	2.38	0.44
Average:		2,269	0.01	0.99	0.23

specified for the spatial (x and y) and temporal (t) dimensions. The second-order property values of each candidate storm (e.g. SIZERANGE) are then calculated and compared to the values found in the LOCALSEVERE property template given in Table 1 to determine whether a particular candidate may be recognized as an instance of a LOCALSEVERE storm.

When the property template listed in Table 1 is used to extract LOCALSEVERE storms, 398 individual candidate storms are identified that meet the first-order property that specifies spatio-temporal contiguity of the presence of precipitation. Of these 398, only five candidates meet the second order properties specified by the SIZERANGE and MAXPRCPRANGE properties and were thus recognized as LOCALSEVERE storms. Table 2 lists these five storms, their observed properties, and the group mean for each property.

It is important to note that nothing was 'set-in-stone' about the criteria listed in the LOCALSEVERE property template as shown in Table 1, and the property values themselves were set somewhat arbitrarily. In fact, it is intended that the criteria listed in the property templates are gradually modified and refined to support the iterative nature of the knowledge discovery process. For instance, a researcher may be interested in how the recognition of LOCALSEVERE storms may be altered if certain criteria are relaxed. If the minimum value for the MAXPRCPRANGE property is changed from 0.5 inches to 0.25 inches, nine additional LOCALSEVERE storms may be identified. One might also wish to investigate the effect of relaxing the criteria regarding the size of a storm. If the property template listed in Table 1 is further altered so that the maximum value of the SIZERANGE criteria is changed from 3,200 km² to 8,000 km², a total of 23 storms may be identified.

Table 3 LOCALSEVERE storms extracted using the property template shown in Table 1 altered by changing the minimum value for the MAXPRCPRANGE property to 0.25 inches, the maximum value of the SIZERANGE criteria to 8,000 km², and both the *x* and *y* contiguity constraints to 10

Month	Day	Size (km ²)	MinPrpc (in)	MaxPrpc (in)	AvgPrpc (in)
6	8	3,136	0.01	0.51	0.14
6	12	3,408	0.01	0.38	0.07
7	9	6,144	0.01	0.79	0.09
7	14	6,112	0.01	0.30	0.05
7	17	976	0.01	0.65	0.15
8	12	2,832	0.01	0.26	0.04
8	26	5,888	0.01	0.29	0.07
Average:		4,390	0.01	0.46	0.08

Table 4 Property template for the LOCALSEVERE storm type

	Maximum Value	Minimum Value
LIFESPANRANGE (Days)	1	1
SIZERANGE (km ²)	10,000	160
PRCPVALRANGE (inches)	null	0.01
MAXPRCPRANGE (inches)	null	2.00
AVGPCRPRANGE (inches)	null	null
Contiguity constraints (<i>x</i> , <i>y</i> , <i>z</i> , <i>t</i>) = (1, 1, 0, 0)		

Throughout the alterations to the LOCALSEVERE property template thus far the *x* and *y* contiguity constraints have both remained at a value of '1,' thus ensuring that each individual storm is composed of a set of ATTVALUE objects with precipitation values ≥ 0.01 inches that are directly spatially adjacent to one another. If we continue to alter the property template shown in Table 1 by relaxing the contiguity constraint to '10' in both the *x* and *y* dimensions, a total of seven individual storms may be identified (Table 3). Note that these storms may be thought of as 'regions' in which there may be spatial 'gaps' up to nine cells wide between areas of precipitation. A comparison of Tables 2 and 3 clearly shows that by iteratively revising the definition of a category and rerunning the feature extraction algorithm, very different sets of entities may be extracted.

Consider another example which seeks to extract both LOCALSEVERE and LOCALMILD storm types with the property templates shown in Tables 4 and 5, respectively. When these LOCALSEVERE and LOCALMILD templates were used to extract storms from the observational data, a total of 202 individual storms were identified. Figure 4 shows a map depicting six of these storms that occurred on 15 August 1969. Note that some contiguous regions of precipitation were too small to meet the size criteria for recognition as a local storm, such as those isolated cells of precipitation at the extreme southern end of the SRB.

Table 5 Property template for the LOCALMILD storm type

	Maximum Value	Minimum Value
LIFESPANRANGE (Days)	1	1
SIZERANGE (km ²)	10,000	160
PRCPVALRANGE (inches)	null	0.01
MAXPRCPRANGE (inches)	1.99	null
AVGPCRPRANGE (inches)	null	null
Contiguity constraints (x, y, z, t) = (1, 1, 0, 0)		

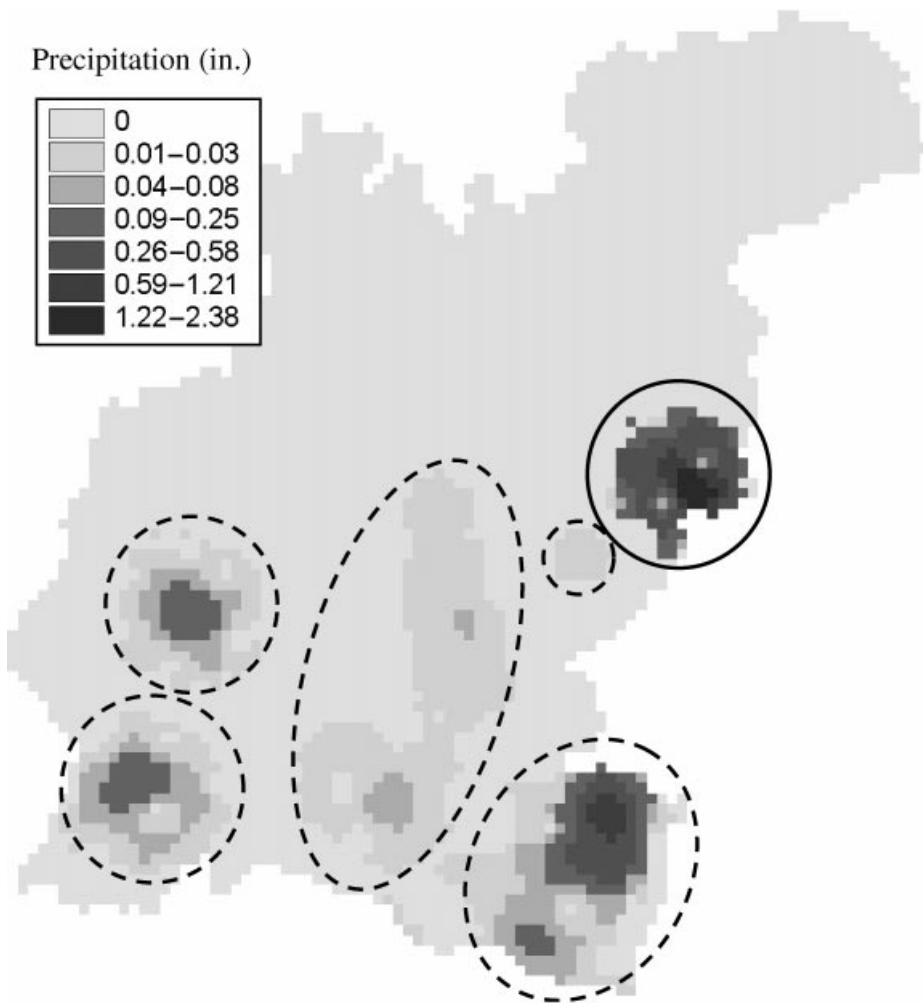


Figure 4 A LOCALSEVERE storm (solid circle) and five LOCALMILD storms (dashed circles) that occurred on 15 August 1969, generated according to the storm type properties listed in Tables 4 and 5

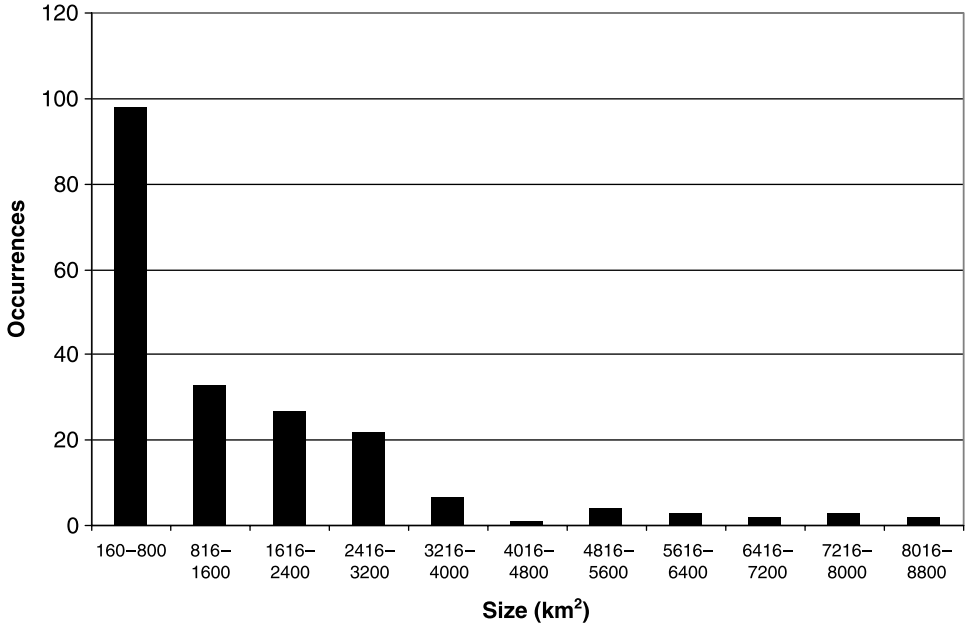


Figure 5 Histogram of size of the 202 LOCALSEVERE and LOCALMILD storms generated according to the storm type properties listed in Tables 4 and 5, respectively

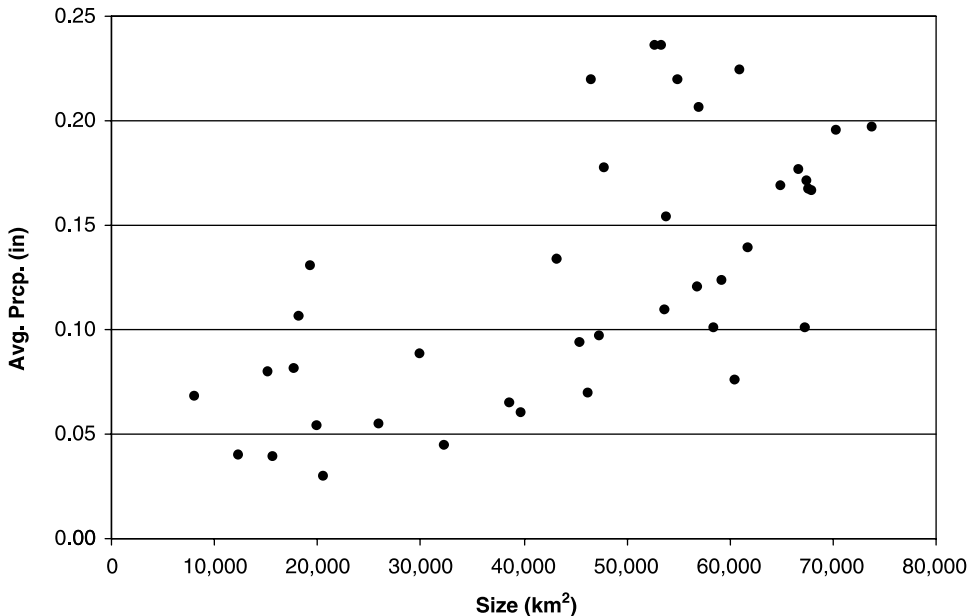
In order to investigate the character of these local storms a histogram of storm size was generated (Figure 5). The histogram clearly shows that most of the storms are relatively small, with a pronounced peak in concentration at less than or equal to 800 km², even though local storms may range in size up to 10,000 km² as specified in the LOCALSEVERE and LOCALMILD property templates (Tables 4 and 5, respectively). A visual display of precipitation (Figure 4) provides insight into the distribution of the size of storms. On 15 August 1969, six individual local storms can be recognized, many of which are very close to one another spatially. However, because the contiguity constraint was specified as '1' in the *x* and *y* dimensions, each individual 'cluster' of precipitation was recognized by the clustering algorithm as an individual storm, even if it was only separated by one 'non-precipitation' cell from another precipitation cluster.

Of course, these individual storms that all took place on 15 August 1969 in the SRB are most likely dynamically related to one another. A user may therefore wish to consider all of these individual local storms as one larger storm system and investigate the properties of this type of storm and its distribution throughout the observational data. As an example of this approach, a new property template was interactively defined for a REGIONALMILD storm type with the properties listed in Table 6. Storm feature extraction using this template resulted in the identification of 40 REGIONALMILD storms. Because all precipitation values on 15 August 1969 are spatially within ten cells of another precipitation value, all of the local storms together are recognized as one REGIONALMILD storm with the following properties:

- SIZE: 18,288 km²
- LIFESPAN: 1 day

Table 6 Property template for the REGIONALMILD storm type

	Maximum Value	Minimum Value
LIFESPANRANGE (Days)	1	1
SIZERANGE (km ²)	null	10,000
PRCPVALRANGE (inches)	null	0.01
MAXPRCPRANGE (inches)	null	null
AVGPRCPRANGE (inches)	0.24	null
Contiguity constraints (x, y, z, t) = (10, 10, 0, 0)		

**Figure 6** Scatterplot of size versus average precipitation of REGIONALMILD storms, generated according to the storm type properties listed in Table 6

- AVGPRCP: 0.11 inches
- MAXPRCP: 2.38 inches

In order to investigate the character of these REGIONALMILD storms, a scatterplot of the size versus the average precipitation associated with each storm is generated (Figure 6). Note that as the size of a storm increases, so does the average precipitation. This relationship is also demonstrated through further statistical analysis. A regression of storm size on average precipitation results in a standardized $\beta = 0.66$ and an $R^2 = 0.44$ (significance level < 0.0005).

A set of 22 REGIONALSEVERE storms are also identified using a property template similar to Table 6, but in which the minimum and maximum AVGPRCPRANGE values were set to 0.25 and null, respectively. Interestingly, there is not a statistically significant relationship between storm size and average precipitation among these

REGIONALSEVERE storms. These results appear to indicate a difference in the dynamics between large mild and large severe storms in the SRB that is of climatological significance. The positive relationship between storm size and average precipitation in REGIONALMILD storms may be due to the ability of large storms in the SRB to tap into Atlantic Ocean moisture. The lack of such a relationship in the REGIONALSEVERE storms may indicate that the genesis of large and severe, as opposed to mild, summer storms in the SRB are influenced primarily by local meteorological factors.

The motivation for presenting this information here, however, is not necessarily to demonstrate any specific climatologic principle, but rather to demonstrate how knowledge representation, and its iterative refinement, may aid in the knowledge discovery process. Note that the relationship between REGIONALMILD storm size and average precipitation may be recognized using the knowledge representation and feature extraction techniques described here, but may otherwise have gone undetected using more standard data mining techniques. Consider, for example, that attribute-oriented induction would allow for the characterization of precipitation observations summarized according to various abstraction levels of some predefined concept hierarchy. However, because attribute-oriented induction focuses strictly on identifying patterns evident within the observational data, and not on patterns that may be observed among the mined features extracted from those data, it would not reveal the relationship between REGIONALMILD storm size and average precipitation that is demonstrated here. Simply put, there would be no such thing as a ‘storm’ represented, and thus capable of being analyzed, in the attribute-oriented induction of the precipitation data.

4.4 Extending the Demonstration

Thus far in the demonstration, the representation of entities was limited to those entities with a lifespan of one day, essentially one ‘time-step’ in the observational data. In order to demonstrate the representation of spatio-temporal entities that last for longer time periods (more than one ‘time-step,’ within our example data set), we focus here on the representation of temperature regions: regions of heat or cold that extend across space and through time.

Just as the CATEGORY class was extended to create the STORMCATEGORY class for the representation of storm phenomena, the CATEGORY class was also extended to create the TREGIONCATEGORY class for the representation of temperature regions. The TREGIONCATEGORY class may be extended to represent a particular type of temperature region entity, such as a heat wave or cold wave. The TREGIONCATEGORY class contains the following properties:

- SIZERANGE range in size
- LIFESPANRANGE range in lifespan
- TMAXVALRANGE range in maximum temperature value
- TMINVALRANGE range in minimum temperature value

Analogously, the THING class was extended to create the TREGIONTHING class, which represents an individual temperature region that may be observed in the observational data. The TREGIONCATEGORY class was then extended to create a class called HOTREGION, which is intended to represent spatio-temporally contiguous regions of high temperature, i.e. a heat wave. The properties associated with the HOTREGION class are listed in Table 7.

Table 7 Property template for the HOTREGION temperature region type

	Maximum Value	Minimum Value
LIFESPANRANGE (Days)	null	null
SIZERANGE (km ²)	null	160
TMAXVALRANGE (°F)	null	95
TMINVALRANGE (°F)	null	null
Contiguity constraints (x, y, z, t) = (10, 10, 0, 1)		

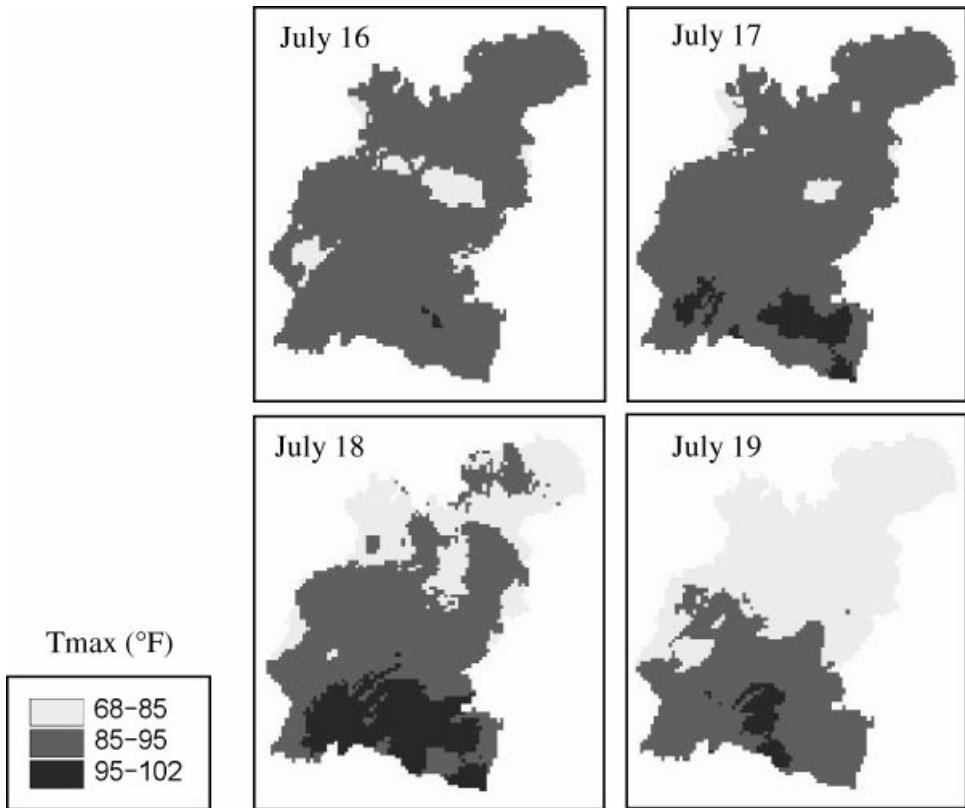


Figure 7 A HOTREGION temperature region that began on 16 July 1969 and ended 19 July 1969, generated according to the properties listed in Table 7. Areas of the darkest shade of gray denote membership within the HOTREGION object

The properties for the HOTREGION class specify that there be a contiguous region of greater than 160 km² in which each observation has a maximum daily temperature greater than or equal to 95°F. Note that the temporal contiguity constraint ' t ' is set to '1' so that regions may be extended in the temporal dimension. Figure 7 shows, via a series of four small maps, a HOTREGION temperature region that began on 16 July

Table 8 Property template for the HOTREGION temperature region component of the PTEVENTCATEGORY

	Maximum Value	Minimum Value
LIFESPANRANGE (Days)	1	1
SIZERANGE (km ²)	null	160
TMAXVALRANGE (°F)	null	90
TMINVALRANGE (°F)	null	null
Contiguity constraints (x, y, z, t) = (10, 10, 0, 1)		

1969 and ended on 19 July 1969. In addition to its BIRTH and DEATH, other properties of the HOTREGION include:

- SIZEMIN: 176 km² (16 July 1969)
- SIZEMAX: 14,848 km² (18 July 1969)
- LIFESPAN: 4 days
- AVGTEMP: 96°F
- MAXTEMP: 102°F

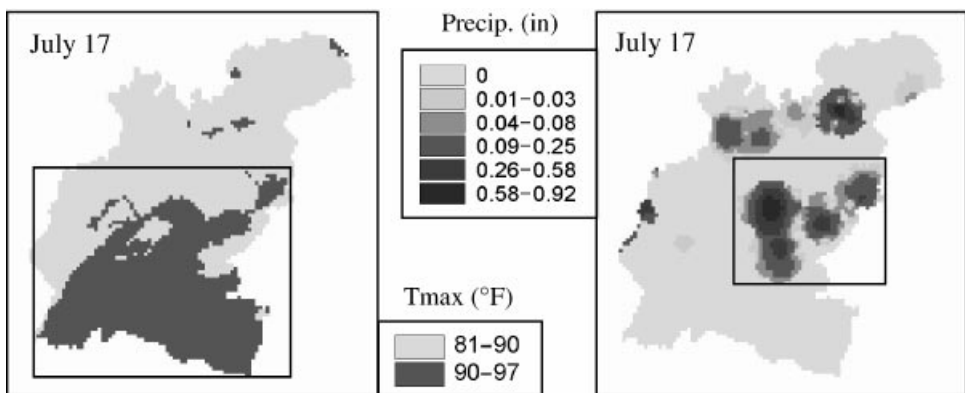
This example of temperature regions may be extended for the representation of complex spatio-temporal phenomena, those phenomena that are composed of multiple spatio-temporal components. As an example of the representation of complex phenomena, we define and extract ‘precipitation-temperature events’ from the observational data, each of which is composed of a TREGIONTHING object and a STORMTHING object that are coincident in space and time. The motivation for this example is to examine the relationships between temperature and precipitation patterns by identifying when, and how often, individual types of storms and temperature regions co-exist. The category and entity aspects of this complex phenomenon are represented in the database by the PTEVENTCATEGORY and PTEVENTTHING classes, respectively.

To extract PTEVENTTHING objects from the observational data, the PTEVENTCATEGORY class calls a method that iterates through the collections of STORMTHING and TREGIONTHING objects and finds entities of each type that spatially and temporally intersect. When such an intersection is found, the coincident objects are stored as components of a new PTEVENTTHING object. For this example, a PTEVENTTHING was defined as the spatio-temporal intersection of a HOTREGION object and a REGIONMILD object. By altering the properties of the HOTREGION and REGIONALMILD classes, one can identify a different set of temperature regions and storms and therefore a different set of PTEVENTTHING objects. The properties of the HOTREGION and REGIONALMILD classes used for this demonstration are listed in Tables 8 and 9, respectively.

Figure 8 shows a PTEVENTTHING object that was found for 17 July 1969. This PTEVENTTHING object is composed of two component objects: one HOTREGION object and one REGIONALMILD object. Note that these two objects partially overlap in space. While there are other storm entities present on 17 July 1969, they are not considered part of the PTEVENTTHING object because they are not spatially coincident with the HOTREGION object.

Table 9 Property template for the REGIONALMILD storm type component of the PTEVENTCATEGORY

	Maximum Value	Minimum Value
LIFESPANRANGE (Days)	1	1
SIZERANGE (km ²)	null	10,000
PRCPVALRANGE (inches)	null	0.01
MAXPRCPRANGE (inches)	2.00	null
AVGPCRPRANGE (inches)	null	null
Contiguity constraints (x, y, z, t) = (1, 1, 0, 0)		

**Figure 8** A PTEVENTTHING object that took place on 17 July 1969, according to the properties listed in Tables 8 and 9. This object is composed of one HOTREGION object, shown boxed in on the left side of the diagram, and one REGIONALMILD object, shown boxed in on the right side of the diagram

5 Discussion and Conclusions

This case study demonstrates how geographic knowledge can be represented and applied within the context of KDD to extract semantic entities from an observational data set. The example concerning the representation of storms in the Susquehanna River Basin shows how a hierarchy of storm types defined from expert knowledge may be formally represented within a database context and used to extract instances of those storm types from the observational data. The example further demonstrates how the user may interact with the Knowledge Component in order to build and refine that knowledge in a flexible manner using both his/her own a priori knowledge and new information derived from the data mining procedures. The example concerning temperature regions demonstrates the representation and extraction of features that exist across space and through time, and the example concerning precipitation-temperature entities demonstrates the representation and extraction of complex entities and categories.

Note that because KDD is an iterative and interactive process, the knowledge representation and discovery environment described here is intended to be flexible, and to

allow a researcher to revise the database representation of the categories and entities that compose a given domain. The user is thus free to create a property template for a particular semantic category, extract the appropriate features from the observational data, analyze the results using visualization or statistics, revise the property template, extract a new set of features, and so on. The representation of knowledge in the database is therefore not simply the 'end-product' of the knowledge discovery process, but also acts to guide the researcher through the process by providing multiple semantic interpretations of the observational data.

Consider, for instance, the example of storm extraction and the cycle of knowledge discovery illustrated in Figure 3. As a first step we defined the LOCALSEVERE storm type using the property template described in Table 1 to extract individual storms. Note that while this first step demonstrates the ability to represent geographic knowledge in a database, it does not, by itself, demonstrate the role of knowledge representation in KDD. It is only in conjunction with the following steps in the process, whereby the definition of the LOCALSEVERE storm type is revised and the results explored using map and statistical techniques, that the iterative process of knowledge discovery is supported. While we used mapping and regression to investigate the patterns associated with the extracted storms, other statistical and graphical data mining techniques may be easily incorporated within this cycle of knowledge discovery as part of a KDD process. For example, within the demonstration domain used here, attribute-oriented induction or association rule mining could be used to extract relationships among the set of extracted storms, and this information could then be used to revise the storm feature extraction rules.

While we acknowledge that the case study presented here uses a well-established technique, rule-based reasoning, for purposes of knowledge representation and feature extraction, we use this technique to extend KDD research in a variety of ways. First, we have provided a flexible and general method for representing stored knowledge within the context of KDD. Second, we have demonstrated a prototype KDD environment that allows the user to interact and iteratively revise that knowledge. Third, we have demonstrated how KDD can be applied to a geographic domain through database support for geographic knowledge representation and geographic feature extraction algorithms that utilize that knowledge. In future research, we plan to extend the general approach we have established here by exploring more complex geographic domains and incorporating more sophisticated data mining techniques.

Acknowledgements

The authors would like to thank Diansheng Guo for assistance with implementation of the KDD environment and Brent Yarnal for providing the meteorological data under EPA grant #824807-010. This research was supported in part by a NASA Earth System Science Graduate Fellowship and by EPA grant #R825195-01-0.

References

- Booch G 1994 *Object-Oriented Analysis and Design*. New York, Benjamin/Cummings Publishing Co
 Carleton A 1991 *Satellite Remote Sensing in Climatology*. London, Bellhaven Press

- Chen M-S, Han J, Yu P S 1996 Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* 8: 866–83
- Egenhofer M J, Glasgow J, Gunther O, Herring J R, and Peuquet D J 1999 Progress in computational methods for representing geographical concepts. *International Journal of Geographical Information Science* 13: 775–96
- Ezquerria N, Mullick R 1996 Knowledge-guided segmentation of 3D imagery. *CVGIP: Graphical Models and Image Processing* 58: 512–23
- Fayyad U, Piatetsky-Shapiro G, and Smyth P 1996 From data mining to knowledge discovery: An overview. In Fayyad U, Piatetsky-Shapiro G, Smyth P, and Uthurusamy R (eds) *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA, AAAI/MIT Press: 1–34
- Gahegan M 2000 On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis* 32: 113–39
- Gahegan M, Takatsuka M, Wheeler M, and Hardisty F 2002 Introducing GeoVISTA Studio: An integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems* 26: 267–92
- Guo D, Peuquet D, and Gahegan M 2002 Interactive subspace clustering for mining high-dimensional spatial patterns. In Zavala G (ed) *GIScience 2002: The Second International Conference on Geographic Information Science*. Oakland, CA, University of California Regents: 60–3
- Han J, Cai Y, and Cercone N 1993 Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Data and Knowledge Engineering* 5: 29–40
- Han J, Koperski K, and Stefanovic N 1997 GeoMiner: A system prototype for spatial data mining. *ACM SIGMOD '97*: 553–56
- Imieliński T and Mannila H 1996 A database perspective on knowledge discovery. *Communications of the ACM* 39: 58–64
- Leung Y, Leung K S, and He J Z 1999 A generic concept-based object-oriented geographical information system. *International Journal of Geographical Information Science* 13: 475–98
- Mennis J L 2003 Derivation and implementation of a semantic GIS data model informed by principles of cognition. *Computers, Environment, and Urban Systems* 27: in press
- Mennis J L, Peuquet D J, and Qian L 2000 A conceptual framework for incorporating cognitive principles into geographical database representation. *International Journal of Geographical Information Science* 14: 501–20
- Miller H and Han J (eds) 2001 *Geographic Data Mining and Knowledge Discovery*. London, Taylor and Francis
- Minsky M 1975 A framework for representing knowledge. In Winston P H (ed) *The Psychology of Computer Vision*. New York, McGraw-Hill: 211–77
- National Climate Data Center 1995 Surface Land Daily Cooperative: Summary of the Day. Washington, D.C., U.S. Department of Commerce NCDC Numeric Data Collection No TD-3200
- Openshaw S 2000 GeoComputation. In Openshaw S and Abraham R J (eds) *GeoComputation*. London, Taylor and Francis: 1–31
- Peuquet D J 2002 *Representations of Space and Time*. New York, Guilford
- Peuquet D J and Kraak M-J 2002 Geobrowsing: Creative thinking and knowledge discovery using geographic visualization. *Information Visualization* 1: 80–91
- Ramachandran R, Conover H, Graves S, and Keiser K 2000 Challenges and solutions to mining earth science data. In Dasarathy B V (ed) *Data Mining and Knowledge Discovery: Theory, Tools, and Technology II*. Bellingham, Washington, SPIE – The International Society for Optical Engineering: 259–64
- Roddick J F and Hornsby K (eds) 2000 *Temporal, Spatial, and Spatio-Temporal Data Mining*. Berlin, Springer
- Roddick J F and Spiliopoulou M 2002 A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering* 14: 750–67
- Schoenauer M and Sebag M 1990 Incremental learning of rules and meta-rules. In Porter B W and Mooney R J (eds) *Machine Learning: Proceedings of the Seventh International Conference on Machine Learning*. San Francisco, CA, Morgan Kaufman: 49–57
- Shoshani A 1997 OLAP and statistical databases: similarities and differences. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Washington D.C., ACM: 185–96

- Sonka M, Tadikonda S, and Collins S 1996 Knowledge-based interpretation of MR brain images. *IEEE Transactions on Medical Imaging* 15: 443–52
- Sowa J F 2000 *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA, Brooks/Cole Thomson Learning
- Spiliopoulou M and Roddick J F 2000 Higher order mining: Modeling and mining the results of knowledge discovery. In Ebecken N and Brebbia C A (eds) *Data Mining 2000: Proceedings of the Second International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields*. Southampton, WIT Press: 309–20
- Ton J, Sticklen J, and Jain A K 1991 Knowledge-based segmentation of Landsat images. *IEEE Transactions on Geoscience and Remote Sensing* 29: 222–32
- Weisman M L and Klemp J B 1986 Characteristics of isolated convective storms. In Ray P S (ed) *Mesoscale Meteorology and Forecasting*. Boston, MA, American Meteorological Society: 331–58
- Yuan M, Battenfield B, Gahegan M N, and Miller H 2001 *Geospatial Data Mining and Knowledge Discovery*. Washington, D.C., University Consortium for Geographic Information Science White Paper on Emerging Research Themes (available at <http://www.ucgis.org/emerging/>)
- Zhang M, Hall L O, and Goldof D B 2002 A generic knowledge-guided image segmentation and labeling system using fuzzy clustering algorithms. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics* 32: 571–82