UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ASSESSING SIMILARITY OF DYNAMIC GEOGRAPHIC PHENOMENA IN SPATIOTEMPORAL DATABASES

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

John McIntosh
Norman, Oklahoma
2003

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/614

UMI Number: 3093582

UMI®

UMI Microform 3093582

Copyright 2003 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

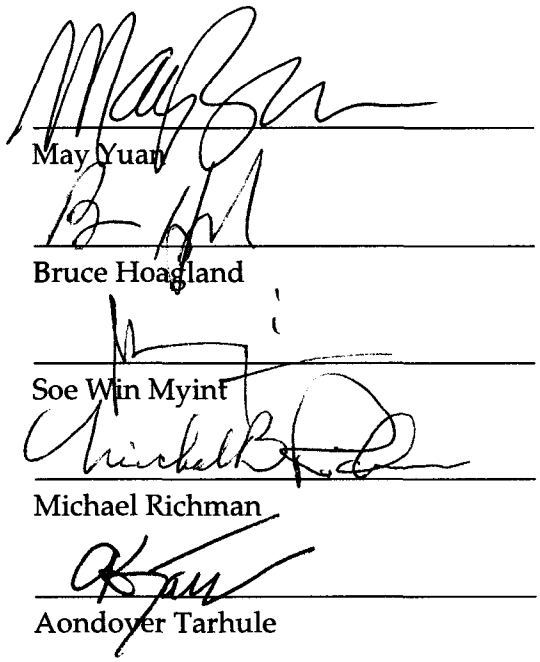
ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© Copyright by John McIntosh 2003
All Rights Reserved.

ASSESSING SIMILARITY OF DYNAMIC GEOGRPAHIC PHENOMENA IN SPATIOTEMPORAL DATABASES

A Dissertation APPROVED FOR THE
DEPARTMENT OF GEOGRAPHY

BY



May Yuan

Bruce Hoagland

Soe Win Myint

Michael Richman

Aondover Tarhule

Acknowledgements

My deepest appreciation goes to all who helped me in this educational adventure. To my Committee, Dr. May Yuan, Dr. Bruce Hoagland, Dr. Soe Win Myint, Dr. Michael Richman, and Dr. Aondover Tarhule, I value the insights and knowledge gained through activities ranging from coursework to casual conversation. I appreciate their time, advice, and patience as I have gone through this process. Special thanks to my Chair, Dr. May Yuan, whose generous support made my studies at OU possible and whose guidance was essential for me to complete this dissertation. Finally, I would like to acknowledge the support of my wife Lesa, who was willing to sacrifice a comfortable existence for a student budget, and my children Emma and Clayton, who have provided a much needed refuge from the stresses of graduate school.

Table of Contents

Chapter 1: The Dissertation Research 1

1. Introduction 2

2. Statement of Research Problems 3

3. Goals and Objectives 6

4. Contribution and Limitations 8

5. Dissertation Organization 9

6. References 10

Chapter 2: A framework to enhance semantic flexibility for analysis of distributed phenomena 11

Abstract 12

1. Introduction 14

2. Representation needs for distributed phenomena 16

3. The Proposed Framework 22

4. Implementation 29

5. Case Study 31

6. Conclusions 39

7. References 43

Chapter 3: Object-Based Characterization of Surface Changes to Enhance Querying and Analysis in Geographic Information Systems 46

Abstract 47

1. Introduction 49

2. Background 51

2.1 Characterizing Surfaces 51

2.2 Characterizing Conceptual Objects in Surfaces 52

2.3 Characterizing change in Distributed Phenomena 54

3. Proposed Indices for Object-Based Characterization of Surface Change 57

3.1 Design Basis 59

4. Case Study 69

5. Conclusions 76

6. References **82**

Chapter 4: Assessing Similarity of Geographic Processes and Events **86**

Abstract **87**

1. Introduction **89**

2. Background **90**

3. Assessing Spatiotemporal Similarity **92**
3.1 Characterizing Processes and Events in Surfaces Using Conceptual Objects 93
3.2 Assessing Similarity of Events Using Temporal Sequences 94

4. Case Study **105**
4.1 Methods 107
4.2 Evaluation 109

5. Conclusions **114**

6. References **117**

Chapter 5: Summary and Conclusions **119**

1. Introduction **120**

2. Summary of Results **121**

3. Discussion and Future Research **124**

Abstract

The growing availability of routine observations from satellite imagery and other remote sensors holds great promise for improved understanding of processes that act in the landscape. However, geographers' ability to effectively use such spatiotemporal data is challenged by large data volume and limitations of conventional data models in geographic information systems (GIS), which provide limited support for querying and exploration of spatiotemporal data other than simple comparisons of temporally referenced snapshots. Current GIS representations allow measurement of change but do not address coherent patterns of change that reflects the working of geographic events and processes. This dissertation presents a representational and query framework to overcome the limitations and enable assessing similarity of dynamic phenomena. The research includes three self contained but related studies: (1) development of a representational framework that incorporates spatiotemporal properties of geographic phenomena, (2) development of a framework to characterize events and processes that can be inferred from GIS databases, and (3) development of a method to assess similarity of events and processes based on the temporal sequences of spatiotemporal properties. Collectively the studies contribute to scientific understanding of spatiotemporal components of geographic processes and technological advances in representation and analysis.

Chapter 1

The Dissertation Research

1. Introduction

The discipline of geography seeks to explain and understand the world using space and spatial relations as the unifying theme. Haggett (1990) points out that "...unlike microscopic sciences where small features have to be brought up by magnification to a scale where out eyes and minds can understand them, geography is macroscopic in that it has to shrink very large features to make them more comprehensible." Analysis of distributions, spatial relationships and other characteristics, such as shape, from maps and charts allow geographers to infer information about processes that shape the landscape. In the past, geographers often worked in data poor environments, but increased use of remote sensing and other innovations of geographic data acquisition are beginning to provide data rich environments for investigation of many phenomena. Extensive data that were previously unobtainable due to costs of data collection are now readily available in digital forms. Sources such as earth observation satellites and other remote sensors not only provide means of collecting new data, but also provide regularly scheduled measurements that document change in space and time.

While many processes occur continuously, regularly scheduled measurements made at discrete points or intervals in time may capture data useful for change detection. Geographical understanding of the processes that shape the world is enhanced by information relating to how change occurs. Regularly scheduled measurements have proven valuable in a wide range of change analysis. For example in

meteorology, temporal loops of remotely sensed information from satellites or weather radars are viewed to monitor and predict weather patterns. Regularly scheduled measurements have also been used for other studies such as the impacts of urbanization, pollution, and drought.

2. Statement of Research Problems

While the growing availability of spatiotemporal data from satellite imagery and other remote sensors holds great promise for improved understanding of geographic processes, geographers ability to effectively use this data is challenged by large data volume and limitations of conventional data models. Data organization and tools available in geographic information systems are based largely on the map metaphor and provide limited support for querying and exploring spatiotemporal data. Techniques such as viewing time series snapshots to observe how events or processes evolve have been used in fields such as meteorology. However, searching spatiotemporal datasets for events or processes based on such patterns typically requires intensive human interaction. For many types of data, the volume produced exceeds the ability for analysts to manually explore all of the available data (MacEachren et al., 1999).

Difficulties of spatiotemporal representation arise for four principal reasons. The first involves the conceptual views of geographic phenomena and the implementation of these views in GIS. Much of the newly available data mentioned above is from observations of geographic phenomena that possess distributed properties yet also contain areas of significance that are perceived as discrete features. For example,

elevation varies continuously yet within the elevation surface there are mountains or plains. In both static and temporal geographic information systems, spatial representation is dominated by two distinct views of the world: continuous fields or discrete objects (Erwig and Schneider 1997, Burrough and McDonnell 1998). Data models that adhere to just one of the world views are unable to provide a complete representation of phenomena that exhibits properties of both continuous fields and discrete objects. For example air pressure varies continuously over the surface of the earth but individual features within the pressure fields such as "ridges" and "troughs" are well recognized and used for analysis, interpretation and forecasting. If the pressure ridges are modeled as objects, variation of pressure within the ridge will be lost. In contrast, if a field model is used, boundaries and dimensions of the "ridge" will not be explicitly represented. Each approach leads to an incomplete representation. While techniques, such as viewing movies of snapshots, allow GIS users to reason about the conceptual objects within the fields, the fact that these objects are not explicitly modeled in a GIS database means that they can not easily be used as a basis for spatiotemporal queries and automated spatiotemporal analysis.

The second issue relates to how time and change are modeled in GIS databases. Time in GIS is often addressed by modeling phenomena using temporally indexed snapshots representing the state of the phenomena at given times or time intervals. Representations are typically indexed on time and object identity, or location in the case of field representations. In such static representations, the state of a surface or object at given times can be easily retrieved and change in the state of a surface or an object can

be determined by overlays. However, change within the landscape often occurs as series of stimuli and transitions that span multiple snapshots and are perceived as processes or events. It is often these series of related changes associated with events that are of most interest to researchers using spatiotemporal datasets.

The third representational issue involves absolute versus relative frames of reference. GIS representations are usually based on absolute spatial and temporal frames of reference. For interpreting spatiotemporal patterns in dynamic phenomena, relative spatiotemporal patterns are often of more importance than absolute time and location. Meteorological phenomena provide a good example. Storms are typically analyzed and classified based on relative properties relating to how the event evolves. Storms may be perceived as similar even if they are separated in place and time. While patterns of change may be observed by looping snapshots, these patterns are not explicitly stored. Representations based on absolute time and location hinders support and analysis based on relative spatiotemporal patterns.

Finally, most GIS support querying by retrieving records that match an explicit set of abstract conditions imbedded in the query. With queries based on Boolean logic, database objects either meet, or do not meet, the conditions specified in the query. In many spatiotemporal applications, this approach is too restrictive. Often there is a continuum of geographic properties and database objects that meet most of the conditions in the query may still be of interest to the user and should not be discounted. While query languages, such as structured query language (SQL), allow some flexibility for specifying ranges, these queries can be quite complex and lack ability to rank objects

based on applicability. There is a need for a straightforward and convenient method to specify complex spatiotemporal queries that are capable of retrieving data based on measures of similarity in spatiotemporal properties.

3. Goals and Objectives

The goal of this study is to develop a conceptual framework to facilitate analysis of spatiotemporal datasets representing dynamic geographic phenomena and to implement a prototype using rainfall events in the southern plains as a case study. The research includes:

- developing conceptual and logical frameworks to organize gridded hydrometeorological data that conforms to the related physical processes;
- extending GIS query and analysis capabilities to represent, search, and assess similarity of dynamic geographic phenomena in terms of processes and events;
- implementing a prototype application to test the framework using rainfall events in the southern plains as a case study.

As described in the previous section conceptual models and temporal representation in most geographic information systems hinder support for querying and analysis based on how events evolve. Most GIS data models represent fields as two-dimensional gridded surfaces of attribute values. In contrast, human analysis and reasoning about fields often centers on the state and behavior of significant features (i.e. zones of relatively high or low values) perceived as objects. With an emphasis on raster data, the first research objective of this dissertation is to develop a temporal representational framework to incorporate concepts of events and processes in GIS data models. In

doing, so GIS queries and analysis can be based on events and processes in addition to objects and states.

The second research objective is to develop a generic framework to characterize how events evolve and to propose a method to assess similarity of the processes and events. The representational framework developed in objective one provides a structure to compile data objects for events and processes from individual snapshots while event characterization in this objective provides a platform to characterize processes and events. The proposed framework uses temporal sequences of spatiotemporal properties that characterize the distribution, state, and behavior of processes and events. Furthermore, these sequences offer a basis for similarity assessments.

The final research objective is to implement the representational and query framework in a prototype application. The prototype is implemented and tested using rainstorm events derived from hourly digital precipitation arrays and National Climatic Data Center storm data reports within the Arkansas-Red River basin located in Oklahoma, and in portions of New Mexico, Texas, Colorado, Kansas, Missouri and Arkansas (Figure 1). The prototype is intended to test the representational framework on spatiotemporal query support beyond simple questions of "what", "where" and "when". The representational and query frameworks identified in the first two research objectives are more complex than standard GIS models and require a new means of formulating queries. The prototype application includes a "query by example" interface to formulate queries and facilitate analysis.

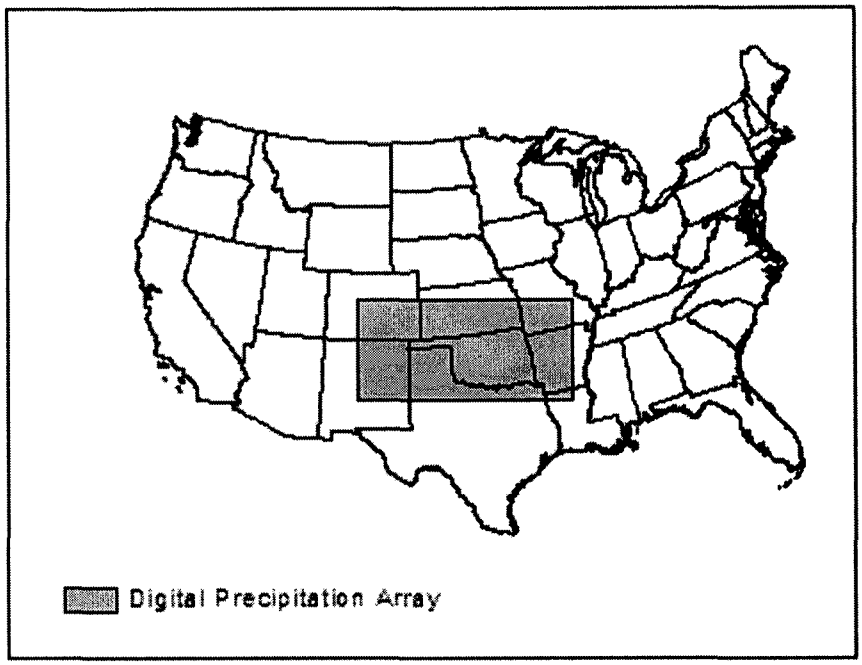


Figure 1. Arkansas-Red Basin River Forecast Center Digital Precipitation Array (DPA).

4. Contribution and Limitations

The frameworks proposed in this dissertation are intended to improve exploration and analysis of dynamic geographic phenomena with properties varying across space. The approach assumes that change within continuous fields occurs in discrete processes and events such as thunderstorms, fires, or floods. The proposed frameworks however, are not suitable for phenomena with change occurring continuously without logical divisions. It is assumed that the distribution and behavior of zones of relatively high values within surfaces provide an adequate basis for queries based on how events and processes evolve. The query framework can only be implemented if the spatial and temporal granularity is adequate to capture relevant

distribution and behavior of the process of interest. With too coarse of spatial or temporal granularity the behavior of processes may be inadequately described. Such inadequacy will invalidate the similarity assessment on which the queries are based.

The framework proposed in this dissertation focuses on representation and querying of a single geographic theme. Most events involve more than a single data layer and event descriptions and comparisons would be more comprehensive to involve other correlated themes such as precipitation and temperature. Finally, the framework is based on two-dimensional fields. Most GIS data models represent fields in two dimensions with the third dimension representing the attribute values, but this is not adequate for some types of phenomena. For example, air pollution fields vary vertically as well has horizontally, and events modeled using two spatial dimensions will not capture this vertical variation. Despite these limitations, the proposed framework should attempt to provide benefit to analysis of spatiotemporal data sets representing a broad range of geographic phenomena.

5. Dissertation Organization

This dissertation is organized into five chapters. Chapters 2, 3, and 4 focus on representing events and processes, characterizing events and processes and finally assessing similarity of events and processes, respectively. The rainfall data is used to test the each of the proposed frameworks. Chapters 2, 3, and 4 are written as stand-alone papers, and are formatted for publication in the *International Journal of Geographical Information Science*. The final chapter provides a summary and conclusions.

6. References

Burrough, P.A. and McDonnell, R., 1998, *Principles of Geographical Information Systems*, (Oxford: Oxford University Press).

Erwig,M. and Schneider,M., 1997, Vague Regions, *5th International Symposium on Advances in Spatial Databases (SSD'97)*, edited by M. Scholl and A. Voisard (Berlin: Springer), pp. 298-320.

Haggett, P., 1990, *The Geographer's Art*, (Oxford: Blackwell).

MacEachren et al, 1999, Constructing knowledge from multivariate spatiotemporal data: integrating geographic visualization with knowledge discovery in database methods, *International Journal of Geographical Information Science*, 13(4), pp. 311-334.

Chapter 2

A framework to enhance semantic flexibility for analysis of distributed phenomena

A framework to enhance semantic flexibility for analysis of distributed phenomena

John McIntosh and May Yuan
Department of Geography
University of Oklahoma

Abstract

Many geographic phenomena are distributed in that their properties vary across an extended area. While such distributed phenomena are best represented as continuous surfaces, individual objects (or features) often emerge among clusters of high or low values in a field. For example, areas of relatively high elevation may be viewed as hills, while flat low-lying areas are perceived as plains in a terrain. A comprehensive spatial analysis of distributed phenomena should examine both the spatial variance of its attribute surfaces and the characteristics of individual objects embedded in the field. An immediate research challenge presents as these emerging features often have vague boundaries that vary according to the use and the user. The nature, and even existence, of these objects depend upon the range of values, or thresholds, used to define them. We propose a representation framework that takes a dual raster-vector approach to capture both field- and object-like characteristics of distributed phenomena and maintain multiple representations of embedded features delineated by boundaries that are likely to be relevant for the expected uses of the data. We demonstrate how boundaries influence the analysis and understanding of spatiotemporal characteristics of distributed phenomena. Using precipitation as a proof of concept, we show how the proposed

framework enhances semantic flexibility in spatiotemporal query and analysis of distributed phenomena in geographic information systems.

Key Words: field-object representation, spatiotemporal data modeling, spatiotemporal query

1. Introduction

Over the past 20 years, the production of geospatial data has increased exponentially. Most significantly is the increase in data for geographic phenomena in which properties are distributed across a wide area and are constantly monitored through remote or in-situ sensors. Examples include terrain, temperature, precipitation, and soil moisture. Accompanying the increase in data has been a shift in its availability for a wide range of users with diverse applications. This poses representational challenges because different users may have distinct views to analyze and interpret these phenomena. Conventionally geographic information systems (GIS) only represent and analyze distributed phenomena as raster fields, but individual features identifiable from distributed phenomena can be the primary indicators for how the phenomena behave in space and time. Weather forecasting presents a typical case to the representation needs for both fields and object-like features embedded in distributed phenomena. The characteristics of "jet streams," "lows," and "highs" in space and time are key indicators of weather progression.

Several approaches have been proposed to represent the field-object dual characteristics by a combination of fields and objects in GIS databases (Winter 1998, Blaschke et al 2000, Yuan 2001). The field representations capture the continuous variation in the surface and the object representations explicitly represent areas of high or low values that are perceived and reasoned about as objects. The combined strengths of the object and field representations enhance the ability to summarize and reason

about overall patterns within distributed phenomena. However, none of these representations has fully addressed issues on identification of individual features. Being embedded in continuous fields, these features do not have clear boundaries. Nevertheless, boundaries play a critical role in determining their spatiotemporal behaviors (e.g. how does a low move) and interactions (e.g. how do areas of high soil moisture relate to the initiation of convective storms). Clearly, boundaries should be set according to theoretical concerns or application needs, and a GIS representation for distributed phenomena should provide semantic flexibility to accommodate the needs for different boundaries.

Expanding upon the dual representation approach, we have developed a framework to represent distributed phenomena with semantic flexibility. Recognizing that boundaries of conceptual objects in fields are inexact and context specific (Egenhofer and Mark 1995, Burrough 1996), the proposed framework extends the dual object/field representation by explicitly storing multiple boundaries in an efficient way so that computational benefits of multiple representations outweigh the disadvantage of the need for extra storage space. The proposed framework also maintains related object-like characteristics and relationships for spatiotemporal analysis. It enhances GIS analytical capabilities by providing a means to: (1) investigate the sensitivity of object-like spatiotemporal characteristics to boundary definitions, and (2) capture hierarchies of geographic information within such phenomena. Furthermore, the proposed framework maintains object identity, necessary for many types of temporal analysis. A prototype for rainfall has been developed as a proof of concept. The prototype uses a

data set of hourly radar derived precipitation estimates over the state of Oklahoma from March 15, 2000 to June 15, 2000, a period with numerous rainstorms in the study area.

In the remaining paper, we elaborate further the needs and challenges of representing distributed phenomena and boundaries issues on identifying their object-like features. We then propose a representational framework to meet the challenges and present a prototype to demonstrate its enhancements to GIS representation and analysis of distributed phenomena. The next section overviews representation of geographic phenomena and establishes a conceptual basis for the proposed framework. The third section presents the proposed framework, followed by sections that present implementations and results of the framework's prototype for rainfall. The final section identifies strengths and weaknesses of the proposed framework and discusses areas for future work.

2. Representation needs for distributed phenomena

Two conceptual models of geographic phenomena dominate GIS views of the world: the exact object model and the continuous field model (Erwig and Schneider 1997, Burrough and McDonnell 1998). In the exact object model, the world is populated with discrete entities with an emphasis on the location of boundaries. Confined by a boundary, an entity acts as a container for attributes that apply uniformly to the space within. Because of the assumption that entities are discrete uniform objects, the exact object model approach does not address any variation that may occur within an entity. In contrast, the world, from the continuous field view, is filled with attributes that vary

continuously over space. Because fields are continuous, the concept of boundaries is not a basis of this model.

These two conceptual models work well for some types of geographic applications. For example, parcels of land, which have exact boundaries with uniform attributes, such as value or ownership, fit neatly into the exact object model. On the other hand, air pressure varies continuously over the Earth surface, and the field model is able to capture such spatial variation. The two conceptual views result in different approaches to representing and analyzing geographic data, although they share the underlying basis of absolute Cartesian space (Peuquet 1988, Couclelis 1992).
In practice, many geographic phenomena, which possess distributed properties yet exhibit discrete features, do not fit well into either of these conceptual models. Data models that adhere to just one of the world views are unable to provide a complete representation of such phenomena. While air pressure, for example, varies continuously over the surface of the earth, individual features within the pressure fields such as "ridges" are well recognized. If the pressure ridges are modeled as objects, variation of pressure within the ridge will be lost. In contrast, if a field model is used boundaries and dimensions of the "ridge" will not be explicitly described. Each approach leads to an incomplete representation.

Nevertheless, representation of object-like features embedded in distributed phenomena is critical to the analysis and understanding of distributed phenomena. For example, in weather forecasting the position of a pressure ridge may be used by a forecaster to identify areas that are unlikely to experience rainfall. Modeling the ridge as

an object also allows the topological relationships of the conceptual object with other objects to be established, such that a ridge may be over a city or be approaching the city. The general shape and orientation of object-like features can provide insight into physical processes or be related to conceptual models used by domain experts. Finally, object-like features enable associations of object identity over time, which can be used as a basis for detecting, characterizing, and tracking changes in the exact object model.

Likewise, it is inadequate to model distributed phenomena that possess both field- and object-like characteristics simply as exact objects because information on the distributed nature of the phenomena will be lost. In an effort to capture both field and object-like characteristics, dual, hybrid and object-oriented approaches have been proposed. Hybrid approaches allow vector and raster representations to be converted to each other and stored in an equivalent form based on grids with a skeleton of edges and nodes for the vector representation (Winter 1998). Dual representations combine vector and raster approaches by identifying vector objects or zones in raster layers to model the object-like characteristics, and use rasters, lattices, or triangular irregular networks (TIN) to model field-like properties (Yuan 2001). Object oriented approaches, on the other hand, store the geometry of object-like features using the raster model, TIN model, vector model or some combination (Blaschke et al 2000).

Dual, hybrid or object oriented representational approaches that model both field and object-like characteristics require the boundaries of object-like features imbedded in the fields, to be defined a priori. Being embedded in a continuous field, object-like features have undetermined boundaries, depending on the use and the user. The

nature, or even existence, of these conceptual objects depends upon the range of values, or thresholds, used to define the object. Even though there is no universally appropriate boundary, the issue of what boundary to use is important to consider because many useful object-like characteristics such as topological relationships or shape descriptions can vary as how the boundary is defined.

To illustrate the needs for representing both field- and object-like characteristics and multiple boundaries for object-like features in distributed phenomena, we used radar derived gridded hourly rainfall accumulations as a case study, and developed a representation framework. Rainfall is a distributed phenomenon that possesses both field- and object-like characteristics. It is commonly represented as raster layers either derived by interpolating point measurements or from remote sensing such as radar or satellite imagery. The field representation of rainfall is valuable for many uses such as water balance calculations and flood forecasting where it is necessary to have specific estimates of the rainfall on a cell-by-cell basis. Object-like features emerge from the presence of rainfall or relative rainfall extremes. These object-like features provide a summary, or interpretation of the overall patterns in a given rain field. The ways in which these features move and evolve suggest the underlying processes that govern the precipitation event. As boundary thresholds change, so do the identification of object-like features, their behaviors, and spatial relationships. Figure 1 illustrates boundaries for an area of rainfall based on three thresholds: >0 mm/hour, >20 mm/hour, and >40 mm/hour. Each of the thresholds determines the zones of precipitation areas of interest. These zones are identified as object-like features in the precipitation field. Figure 1

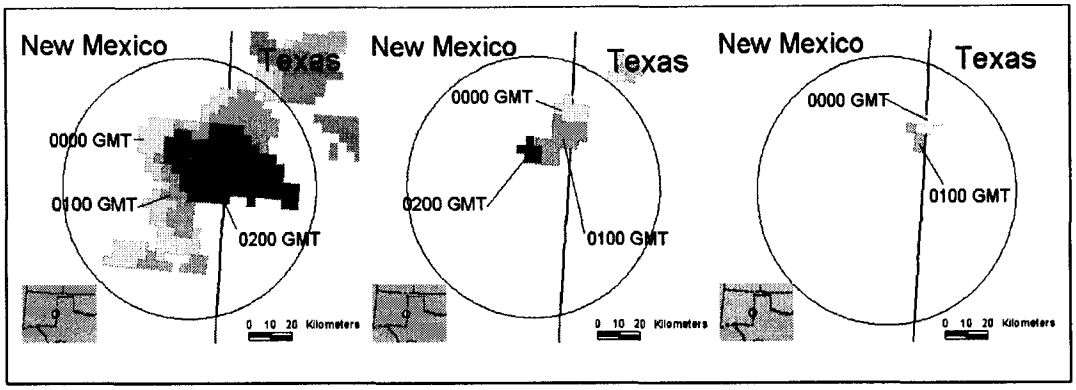


Figure 1. Boundary effects on object characteristics. The three frames show rainfall areas defined using 0, 2, and 4 mm thresholds. Note that the size shape and direction of movement varies by threshold.

illustrates the potential variation in commonly used object characteristics such as movement and spatial relationships associated with different boundaries.

However, there is no universal definition of how zones of rainfall should be defined for the identification of such object-like features. For example, zones of high rainfall may be defined based on a very high hourly rainfall threshold for flood analysis. For characterizing the structure of a rainstorm, zones might be based on relative rainfall amounts to capture characteristic patterns such as areas of low intensity stratiform precipitation and high intensity precipitation associated with convective cells. It is clear that the boundaries used to represent object-like characteristics of accumulated rainfall vary by use. Currently, most GIS store a single raster layer and provide functionality for a different uses by allowing the data to be viewed in a variety of formats generated on demand from the original database (Parent, Spaccapietra and Zimanyi 2000). Although conceptually elegant, this approach is not always practical or even possible. For example with temporal GIS, the geometry and attributes of modeled objects may change

over time requiring multiple representations that are valid for specific intervals or points in time. In addition, some phenomena cannot be adequately modeled by a field or an object-based representation alone requiring a dual or hybrid approach.

In light of the above discussion, distributed geographic phenomena present three major challenges to GIS representation: (1) to capture both field- and object-like characteristics; (2) to provide semantic flexibility in support of application-specific boundary requirements; and (3) to calculate and maintain geometry and spatiotemporal relationships among identified object-like features.

To meet the representation challenges for distributed phenomena, we take a dual approach and propose a framework to explicitly represent both field and object-like characteristics. Furthermore, unlike other dual approaches that identify object-like features according to one threshold, the proposed framework employs multiple representations of these features with boundaries based on a range of threshold values that are likely to be of utility to the users. Although the inherent fuzziness of object boundaries can be accommodated using a fuzzy logic approach, topological and metric properties such as shape or area based on crisp boundaries are often critical for reasoning about these regions. Alternatively, boundaries of object-like features can be delineated automatically for various thresholds on an ad hoc basis. However, the complexity of deriving the spatiotemporal relationships between objects on the fly make this approach impractical given the current technology so a multiple representational framework is taken in our proposed framework.

Maintaining multiple representations indeed increases storage requirements, but this approach is commonly used to support representations that are difficult or impractical to derive on demand. Numerous proposals use multiple representations to incorporate scaling effects such as change in geometry and semantics due to changes in cartographic scale (Rigaux and Scholl 1994, Buttenfield 1995, Jones et al 1996, Timpf 1997, Davis and Laender 1999, Vangenot et al 1999, Mountrakis et al 2000, Parent and Spaccapietra 2000). Multiple representations have also been proposed as a means to enhance interoperability. By storing multiple representations at different resolutions or in different data models, systems can better support complex analysis on the fly. However, there must be a balance between the added storage costs and computational costs. For example, Winter (1998) notes that physical storage of the space would quadruple over a simple raster model in his hybrid framework. With the multiple representations of object-like features, the proposed framework tracks objects over time and generates the complex set of spatiotemporal relationships between the modeled objects. Hence, it is impractical to implement the multiple boundaries of object- like features on demand. By storing multiple representations of object-like features and their relationships explicitly, all the related information can be accessed efficiently.

3. The Proposed Framework

The proposed framework is designed based on the following three strategies: (1) take a dual representation approach to capture both fields and object-like features in distributed phenomena with data representing fields and objects are explicitly stored in

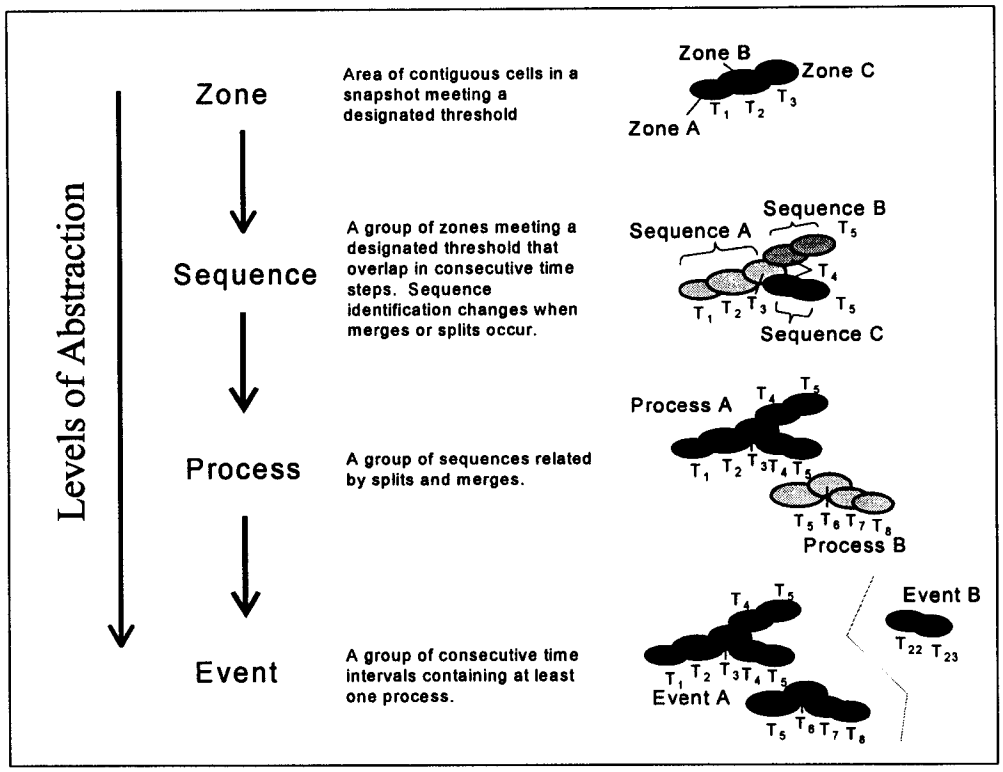


Figure 2. Organization of the object-based representational framework.

a GIS database; (2) take a multiple representation approach to meet the needs of multiple boundary thresholds for different uses; and (3) explicitly model spatiotemporal relationships among identified objects and maintain these relationships for spatial and temporal analysis.

In reference to Yuan's (2001) approach, the proposed framework organizes data into zones, sequences, processes, and events (Figure 2) for raster data collected as snapshots of distributed phenomena that are monitored at regular time intervals. Zones are defined as areas of contiguous cells within a snapshot that meet or exceed a specified threshold value. Because the proposed framework uses a variety of boundary thresholds to delineate object-like features, a zone may contain, or be part of other zones.

The framework explicitly stores *contains* and *part-of* relations for zones. With dynamic distributed phenomena, areas of relatively high values may appear, move, evolve and disappear over time. The framework tracks zones that overlap (or partially overlap) spatially with zones defined by the same threshold in the previous or next time interval as sequences. Under the assumption that the phenomenon of interest is monitored at an interval that is sufficiently fine to capture its temporal variability, a sequence (of zones) represents a continuum of an object-like feature over time. It is a temporal object that may exist over multiple snapshots, or in a single snapshot in the case when a zone does not overlap with another zone in the previous or subsequent snapshot.

With many phenomena, object-like features may split or merge over time. If a sequence splits into two branching sequences (i.e. when a zone at T_3 transits to two zones at T_4 in Figure 2), the original sequence ends, and each of the resulting zones becomes the first zone in its corresponding new sequence. Likewise, if two sequences merge (i.e. when two disjoint zones at T_1 transition to one zone at T_2), the original sequences end, and the resulting zone becomes the first zone in the new merged sequence. The strategy ensures that branching areas always have a distinct sequence identification number. The framework explicitly stores *contains* and *part-of* relations for sequences. Sequences at a higher threshold may be contained within a single lower threshold sequence. The *contains* and *part-of* relations provide a means to link semantically related sequences to corresponding sequences at a higher or the lower threshold. *Future* and *previous* relations are also stored to allow easy tracking of sequences that are involved in splits or merges.

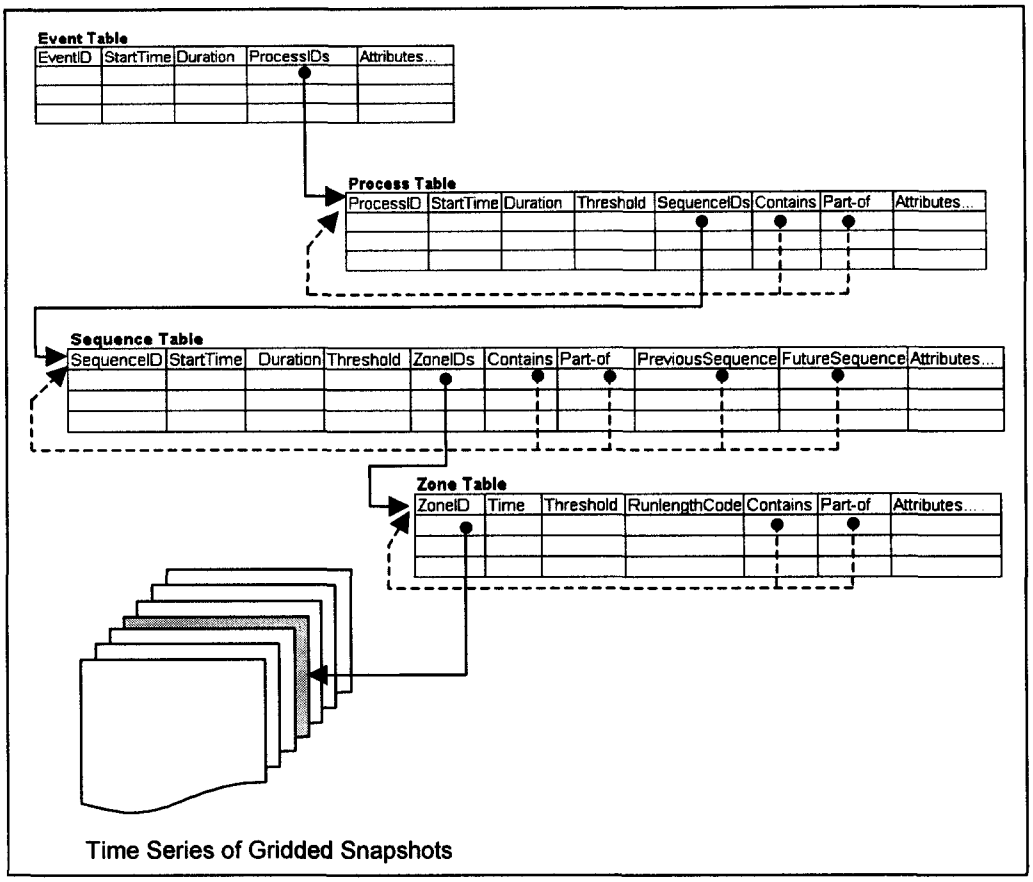


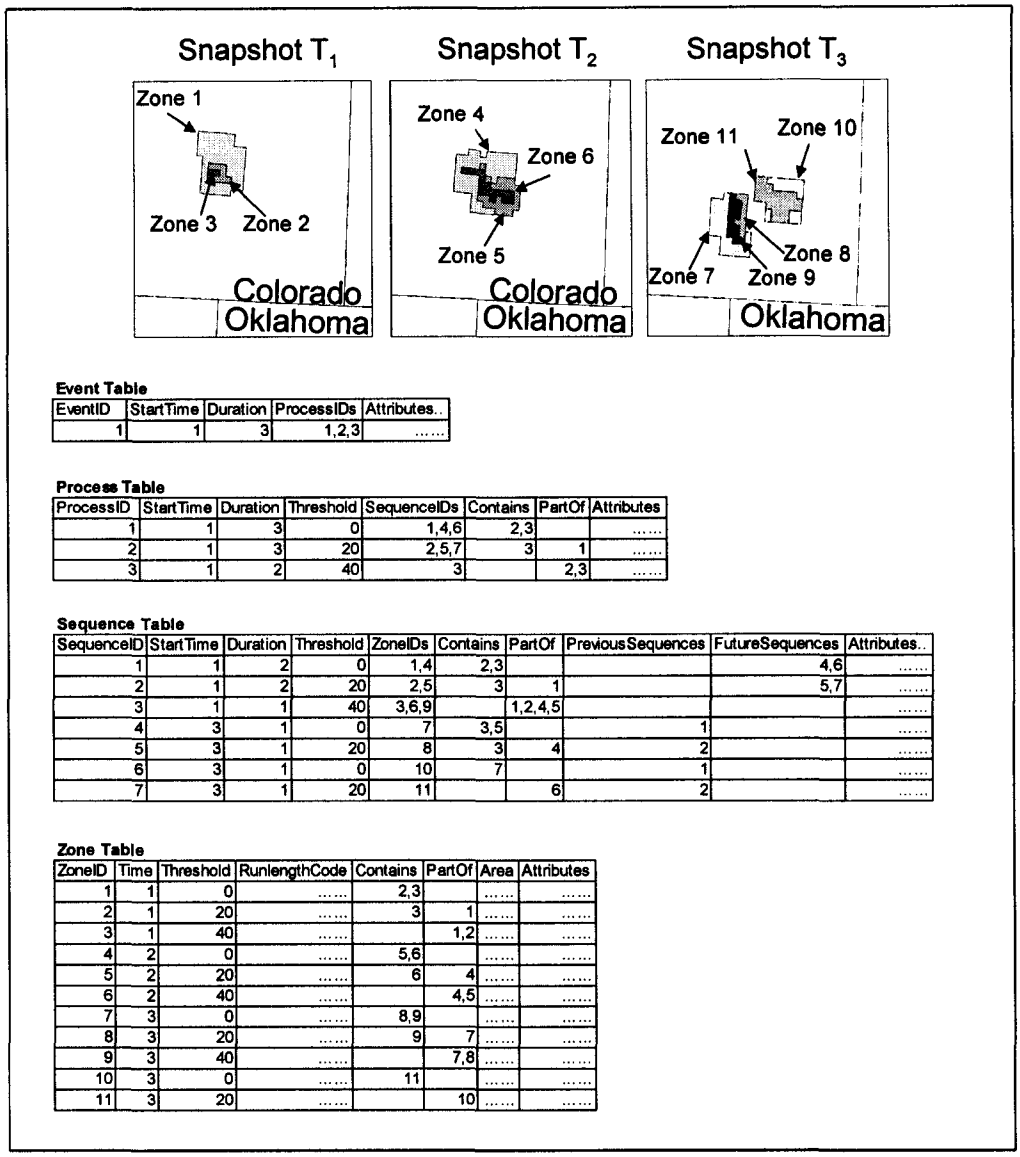
Figure 3. This diagram illustrates data structures used to implement the framework. The solid arrows indicate fields that can be used to associate objects between tables. The dashed arrows indicate fields that can be used to associate objects within the tables.

The proposed framework uses process objects to capture temporally contiguous sequences that may be related through splits or mergers. A process begins when zones of a selected threshold value appear at a snapshot and ends when none of the zones in the subsequent snapshot overlaps with zones in the current snapshot. Like sequences, the framework identifies distinct processes for each of the selected thresholds and stores *contains* and *part-of* relations among them. Finally, events are formed to capture all processes identified by the same boundary threshold, under the assumption that the

very presence of zones constitutes an event. These processes may be separate in space but they collectively persist over consecutive periods of time.

Taking a dual approach, the proposed framework includes both field and object representation schemes. The field scheme consists of a time series of raster layers (snapshots). The object scheme is implemented through a series of tables that store the zone, sequence, process, and event characteristics (Figure 3). Object geometry is stored in the zone table as run length codes corresponding to cells in raster layers with a time stamp to link the object and field schemes. In addition to the geometry, the zone table includes an identifier, a time stamp, the threshold used to delineate the geometry of object-like features, as well as *contains* and *part-of* relations with zones of different boundary thresholds. Attributes include zone area, maximum value, and other informaion to support domain specific analysis.

The geometry of the higher level objects (i.e. sequences, processes, and events) are not stored explicitly but can be easily inferred from their associated zones. The sequence table stores the sequence identifier, the start time and duration, the identification numbers of the zones that form the sequence. Previous sequence and future sequence identifiers are also included when the sequence begins or ends as a result of merging or splitting. The process table stores a process identifier, start time, duration, threshold, the sequence identifiers that form the process, as well as *contains* and *part-of* relations with other spatially overlapping processes identified from different thresholds. Similarly, the event table stores an event identifier, the start time, duration,



Event Table

EventID	StartTime	Duration	ProcessIDs	Attributes..
1	1	3	1,2,3	.....

Process Table

ProcessID	StartTime	Duration	Threshold	SequenceIDs	Contains	PartOf	Attributes
1	1	3	0	1,4,6	2,3		.....
2	1	3	20	2,5,7	3	1	.....
3	1	2	40	3		2,3	.....

Sequence Table

SequenceID	StartTime	Duration	Threshold	ZoneIDs	Contains	PartOf	PreviousSequences	FutureSequences	Attributes..
1	1	2	0	1,4	2,3			4,6	.....
2	1	2	20	2,5	3	1		5,7	.....
3	1	1	40	3,6,9		1,2,4,5			.....
4	3	1	0	7	3,5		1		.....
5	3	1	20	8	3	4	2		.....
6	3	1	0	10	7		1		.....
7	3	1	20	11		6	2		.....

Zone Table

ZoneID	Time	Threshold	RunlengthCode	Contains	PartOf	Area	Attributes
1	1	0	.....	2,3		.....	.....
2	1	20	.....	3	1	.....	.....
3	1	40	.....		1,2	.....	.....
4	2	0	.....	5,6		.....	.....
5	2	20	.....	6	4	.....	.....
6	2	40	.....		4,5	.....	.....
7	3	0	.....	8,9		.....	.....
8	3	20	.....	9	7	.....	.....
9	3	40	.....		7,8	.....	.....
10	3	0	.....	11		.....	.....
11	3	20	.....		10	.....	.....

Figure 4. This figure illustrates how the framework would represent the three snapshots at the top. The zones delineated by three different thresholds are indicated in the snapshots and the sequences, processes, and events that they form are indicated in the respective tables (see text for further discussion)

and the processes associated with the event. In the actual implementation, additional fields representing other attributes of interest can be associated with any of these tables.

Figure 4 presents a simple example of a series of three temporal snapshots of hourly accumulated rainfall to illustrate the representation using the framework. Three thresholds are used. The lightest shade represents the area meeting the lowest threshold

(0mm/hour) with the middle and darkest shades corresponding to the middle (>20mm/hour) and highest threshold (>40mm/hour). This example shows a single event consisting of three processes, seven sequences, and eleven zones. The first process (process 1, including sequences 1, 4, and 6) models the boundaries of the area meeting the lowest threshold, the second (process 2, including sequences 2, 5, and 7), the middle threshold, and the third (process 3, including sequences 3), the upper threshold. The area of rainfall splits after T_2 but all three processes continue into the next period because both new rainfall areas overlap their respective parent process in T_2. Sequences 1 and 2 end at T_2 because they split into two new sequences. The resulting areas in T_3 are assigned new sequence identifiers. Zone 3 continues without dividing or merging, and therefore its identity continues in T_3. The zone table maintains the relationship between the new zones in T_3 with the parent sequences in T_2 in the *PreviousZoneID* field. Likewise, the *FutureZoneID* field contains the zone identifiers from the splitting of the original sequences 1 and 2. These relationships allow a user to track changes and the entire lifespan of object-like features.

Detailed spatial characteristics of the object-like features are associated with zones. Object geometry is stored as run length codes in the zone table. Since the run length codes store only locations, we use a simple binary scheme that alternates between in the zone and not in the zone. For example, in a 5 x 5 raster layer, a 2 x 2 square offset by one from the x and y axes would be represented as {6,2,3,2,12}. The run length codes and fields for *contains* and *part-of* relationship, previous identifiers, and future identifiers in all tables are stored as comma delimited lists for ease of use and

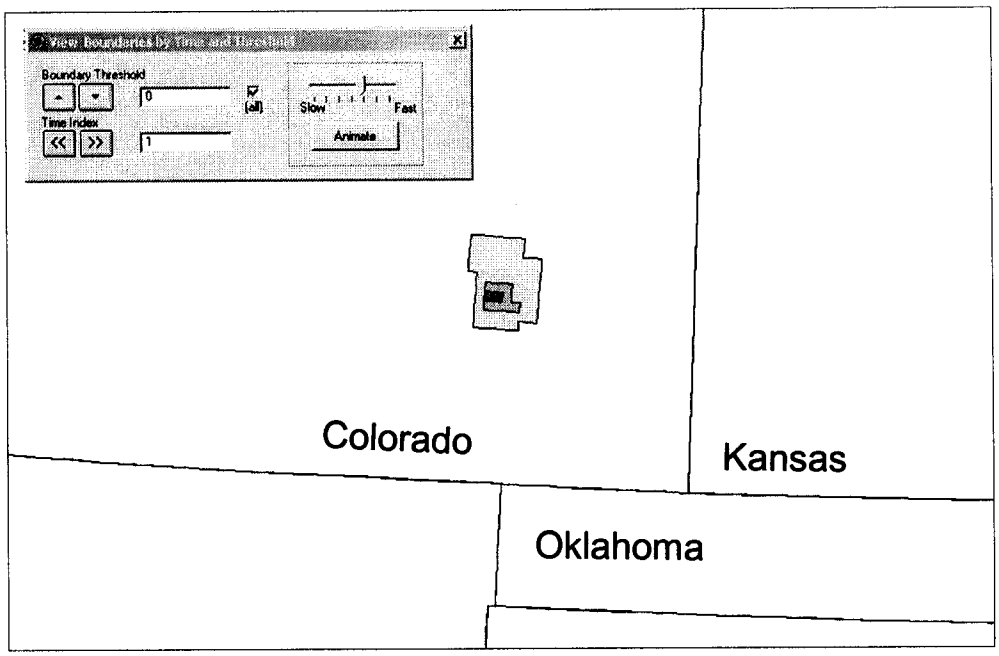


Figure 6. Viewer Dialog. This figure shows the viewer dialog with data included in Figure 4. The viewer allows the user to step through the data changing the boundary thresholds and time.

interpretation. In implementation, several functions have been written in this research to allow information queries based on these lists.

4. Implementation

A prototype was developed to demonstrate the use of the proposed framework in ArcView® GIS 3.2 software (Environmental Systems Research Institute Inc., Redlands, California). Although ArcView does not provide direct support for the proposed representational framework, its scripting language (Avenue™), the relational database support and display capabilities of ArcView provide the necessary tools to implement the framework.

Data preprocessing involved in importing rainfall data into GIS formats and developing algorithms to build zones, sequences, processes, and events from the rainfall

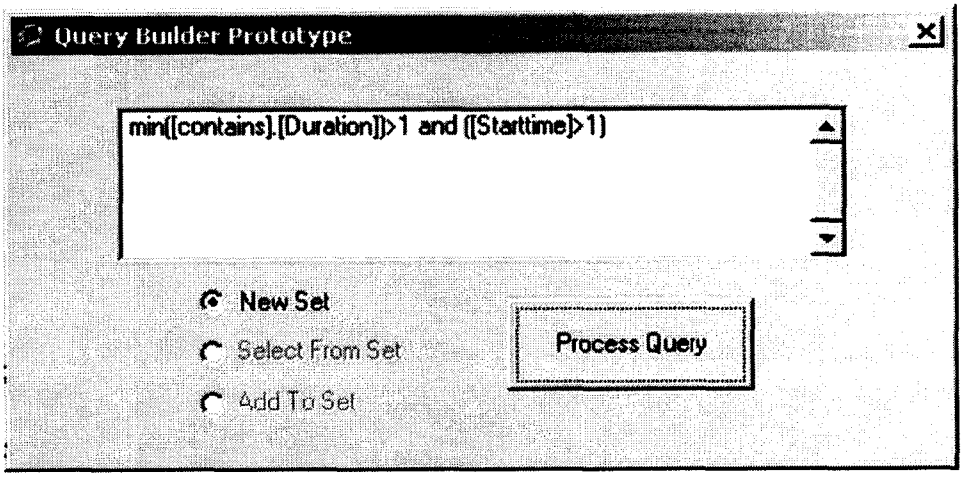


Figure 5. Query Builder Prototype. Applied to the sequence table, this query would return sequences that start after time 1 and contain sequences of a higher threshold with a minimum duration of more than 2 time intervals. Applying this query to the sequences table in Figure 4 would return sequences 1 and 2.

data. Our sample rainfall accumulation data from the Arkansas Red River Forecast Center is in the Hydrologic Rainfall Analysis Project (HRAP) coordinate system, in which the cell size increases away from its projection center. To avoid distortion resulting from large domains, the implementation uses a polygon theme representing the corrected position and shape of HRAP raster cells. Polygons corresponding to the cells represented in the run length codes are selected from this master theme and new themes of zones can be loaded individually or by sequence, process or event identifiers on the fly for review.

Several extensions to the standard query language (SQL) have been programmed to work with comma delimited lists in tables for zones, sequences, processes, and events. These extensions consist of commands for summary statistics that are important for exploratory analysis of these objects, including minimum, maximum, mean, variance, standard deviation, range, count, and sum (Figure 5). The prototype allows selection of

objects based on attribute fields such as *contains, part-of, future ID, Previous ID*. For example, the query in Figure 5 on the process table would return all processes beginning after Time 1 and contain sequences that have an average duration of more than one time interval. These summary characteristics can be calculated for any numeric attribute values that might be added to the basic representation such as movement or shape indices. This provides a means to explore attributes and relationships among zones, sequences, processes, and events. The summary functions work with SQL and can be used together to identify events, processes or sequences of interest.

Because insights are often gained by graphically reviewing spatial properties and relationships, a viewer has been created to allow the user to vary time and thresholds of the object representation (Figure 6). Furthermore, it allows the user to explore relationship between the various zones, sequences, processes and the thresholds used to derive these objects.

5. Case Study

The goals of the case study are to:

- Verify if the prototype provides enhanced querying and analysis capabilities to handle both fields and object-like features embedded in rain fields;
- Determine if the prototype is capable of handling multiple boundaries for object-like features and maintaining their topological and temporal relationships;
- Investigate scaling issues of the proposed framework regarding storage space and processing requirements.

We chose rainstorms to test the working of the proposed framework and prototype. Intense springtime storms are common in the Southern Plains, USA. These storms can result in flooding and associated features such as wind or hail, which can damage crops and structures. Weather radars produce massive rainfall field estimates making it difficult to manually or interactively search for specific spatiotemporal patterns.

Patterns of the most intense rainfall often indicate the storm's structure. In order to store and reason about the structure of storms, or the association of specific parts of the storm system with other severe weather events, it is necessary to incorporate object-like rain features into the analysis. For example, a linear alignment of relatively high rainfall moving perpendicular to the line might suggest a squall line of convective cells. Meteorologists have associated the morphological and structural characteristics to storm dynamics. Scheisser et al (1995) studied the structure of heavy rainfall events in Switzerland and classified storms based on the relative intensity of rain field derived from radar. The storms were categorized based on the object-like features including the shape and position of stratiform rainfall, characteristics, and the leading edge. Hagen et al (1999) studied thunderstorms in southern Germany and identified three classes of storms based on these object-like characteristics - isolated cells, events which follow along a line, and linear aligned thunderstorms that move roughly perpendicular to the major axis. In the U.S., Houze et al (1990) evaluated severe springtime rainstorms in Oklahoma. Storm organization was graded according to the degree to which it matched an idealized model of a leading line/trailing stratiform structure. Factors considered

were shape, orientation, movement of the storm area, characteristics of the leading edge and the presence of stratiform rainfall.

For research such as Houze (1990), Scheisser (1995), or Hagen et al (1999) object-like features of most intense precipitation associated with the cells and the stratiform rainfall areas are of interest. Hence, several rainfall thresholds would be needed to delineate the object-like features of different rainfall intensity. Other uses of rainfall data may still have other requirements. For example if the rainfall data is being studied to improve the efficiency of fertilizers or pesticides, the appropriate boundary might be the presence of rainfall (i.e. >0mm).

Data used in the case study are digital precipitation arrays (DPA) from the National Weather Service's Arkansas-Red River Forecast Center (ABRFC), covering the entire state of Oklahoma and portions of surrounding states (Figure 7). The DPAs are in a raster format and consist of approximately 4km x 4km grids in the Hydrologic Rainfall Analysis Project (HRAP) coordinate system and are archived in the NetCDF format (Arkansas-Red Basin River Forecast Center, 2002). Each grid contains the distribution of hourly accumulated rainfall estimates based on a composite from next generation radars (NEXRAD) and observations at ground weather stations (Schmidt et al. 2000).

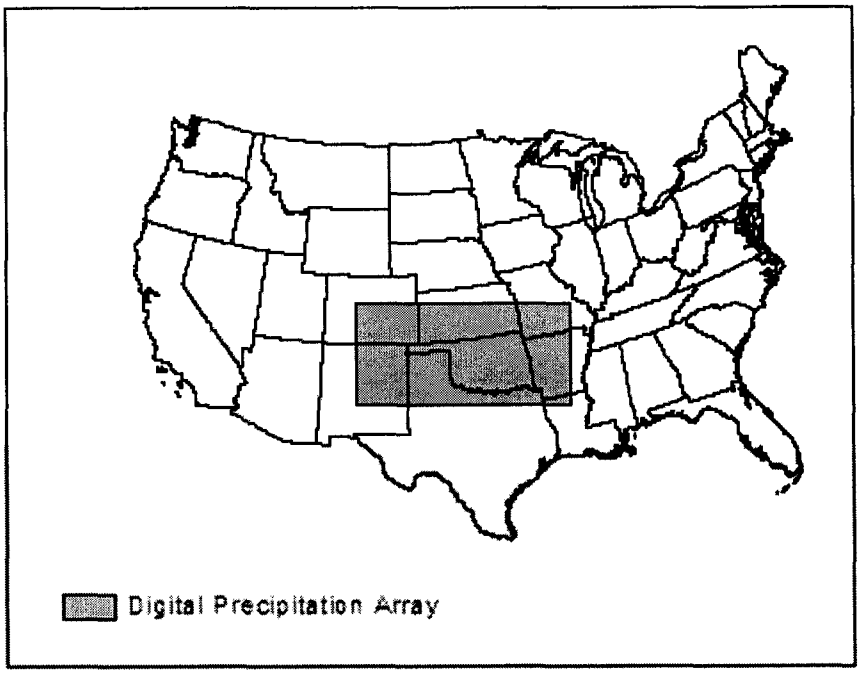


Figure 7. Arkansas-Red Basin River Forecast Center Digital Precipitation Array (DPA)

The DPAs are generated in real time and are used by the ABRFC for flood forecasting. They can also be valuable for other purposes such as climate analysis, risk assessment, facilities planning or agronomy. DPAs are in essence field representations of accumulated rainfall, ideal for hydrologic modeling and flood forecasting where the estimates of hourly rainfall accumulation are needed for discrete locations within the modeled domain. For other types of analysis, object-like characteristics within DPAs will be needed, such as spatiotemoporal patterns and structures of rainstorms.

The framework was implemented using rainfall data from March 15, 2000 to June 15, 2000. We developed several Java scripts to download DPAs from the ABRFC and process the data for input into the prototype. The downloaded DPAs were converted to grids with a storage requirement of about 47 MB. Three thresholds were used to

delineate rainfall "zones" (>0mm/hour, >20 mm/hour, and >40 mm/hour). Houze et al (1990) identify several important structural aspects of springtime rainstorms, including areas of light stratiform precipitation, areas of intense rainfall corresponding to convective cells, and areas of heavy precipitation indicating features such as a squall line. The thresholds selected for the case study are intended to capture these features in DPAs.

Rainfall zones are linked to form sequences based on overlaps in consecutive snapshots (DPAs). Rainfall processes link sequences related by merges or divisions and their predecessors and descendents. Rainfall events represent consecutive periods with rainfall somewhere in the modeled domain, with the understanding that rainfall processes within the same rainfall event may or may not related. The rainfall events, processes, sequences and zones are linked to the DPAs based on a time-date index. In addition to the basic elements of the framework described in section 3, we include attributes relevant to the rainfall dataset. These include area, centroid movement (speed and direction), elongation and orientation of the major axis.

The proposed framework was tested on its ability for enhanced querying for the meteorological case study. By defining zones, sequences and processes based on multiple thresholds and relating these objects through part of and contains relations, the framework provided a means to investigate more complex spatio-temporal patterns than those supported by a dual representation with a single boundary threshold. In the rainfall case study, the framework showed its strength in support of queries on low threshold rainfall processes that contain higher intensity processes. For example,

applying the following query to the zone table would identify rainfall sequences with a threshold of greater than 20 mm of rainfall per hour that contain at least one sequence of higher intensity rainfall.

(min([contains].[Threshold]) > 20) and ([Starttime]<46007)

In our implementation, 1/1/95 at 0 GMT is the reference date for the time index and corresponds to 00001 so the start time of 46007 corresponds to zones or sequences occurring earlier than 4/1/00. Figure 8 shows one of the zones returned by the query (this zone is part of rainfall event over the border of Texas and Oklahoma on 3/27/00). Inclusion of attribute information such as elongation and orientation, the framework supports queries based on characteristic patterns similar to those used in the springtime rainstorm typologies proposed by Houze et al (1990) and Scheisser et al (1996). Including additional attributes can further refine the searches. For example, querying for low intensity processes that contain higher intensity processes that have a significant range in speeds may suggest rainstorms with rotation. Objects selected by the query can be related back to the original gridded data based on the time index number. Boundaries of these objects can be automatically loaded from the run length codes in the zone table for display and analysis in the GIS.

One of the tradeoffs between storing multiple representations versus calculating object-like characteristics on demand is the need for additional storage space. The incremental storage requirements are dependent on the number of boundaries maintained, the complexity of the zones meeting the threshold, and the number of distinct areas modeled in each snapshot. A worst case scenario in terms of incremental

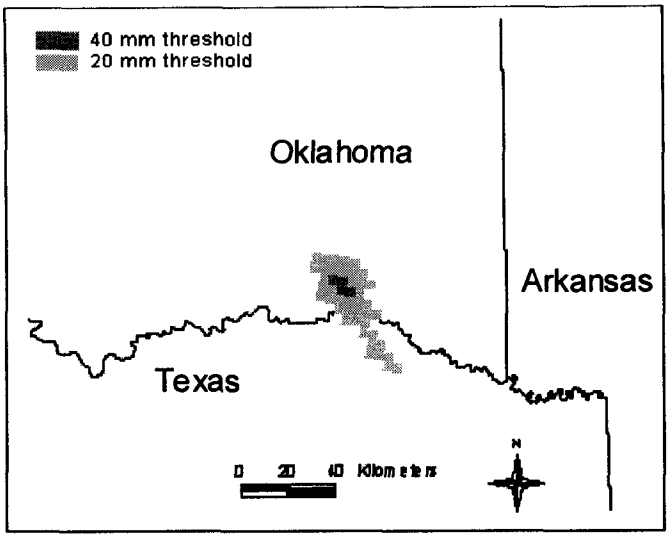


Figure 8. One of the seven sequences returned by the query described in the text. This sequence is part of a rainfall event that occurred over the border of Oklahoma and Texas on 3/26/00 and 3/27/00. The 40 mm/hour threshold sequence shown in the figure had a duration of one hour and occurred on 3/27/00 at 03 GMT.

data costs occurs with maximum spatial and temporal variability in the conceptual objects being modeled. Assuming that zones are defined based on adjacency in one of the four cardinal directions, the worst case scenario would have a checkerboard pattern of zones. This checkerboard pattern would shift each time period so there would be no temporal linkages between the zones. In other words, each zone would also represent the extent of a sequence object and a process object. Under these conditions, the incremental storage cost above an uncompressed raster representation would be proportional to the number of thresholds squared plus the number of thresholds.

The incremental storage costs include the costs of storing the zone, sequence, and process objects. The storage costs for all but the *contains* and *part_of* relations would be constant for each zone, sequence, or process object. The combined cost of the *contains*

and *part_of* relations is related to the number of thresholds minus one (n_{thr} –1) so the combined storage costs (*S*) for the zone, sequence and process objects would be

$$S = O(3n_z*(n_{thr} - 1)+c) \quad (1)$$

where n_z is the number of zones in all modeled time intervals, n_{thr} is the number of thresholds modeled, and c is a constant representing the storage costs of other attributes. Assuming a checkerboard pattern, the maximum number of zone objects at a given representation would be half of the number of cells so the total number of zone objects would equal

$$Z_{max}= 1/2 * n_c * n_{ts} * n_{thr} \quad (2)$$

where n_c is the number of cells in the raster layer and n_{ts} is the number of time steps. The storage costs for a snapshot model using raster representation assuming no compression is $O(n_c * n_{ts})$ so the proportional increase implementing the framework under this scenario would be $O(3/2*[n_{thr}^2 + c*n_{thr}])$.

Although it is theoretically possible to have extremely high costs associated with implementing the framework, the spatial and temporal of the scenario described above is unlikely to occur. Most natural phenomena exhibit some degree of spatial and temporal autocorrelation and far fewer objects can actually be meaningfully identified. We find that modeling rainfall data with three levels of boundaries based on the thresholds described above, the incremental cost is about 3.3 percent, including storing the relationships, attributes, and geometry of the objects.

Since this implementation of the framework stores the geometry of the boundaries as run length codes, it requires processing to convert the stored geometry

into GIS data objects that can be displayed and analyzed using the GIS. In the implementation, shapefiles representing zones can be generated on demand from selected event, process, or sequence objects. The average duration of a process object in our data set is about 4 hours with an average area of about 672 square kilometers (42 grid cells). The average storage space required to store the geometry of process objects of this duration and size using run length codes is about 0.88 KB compared to an average of 4.88 KB of required to store the geometry of the process objects as shapefiles. It only takes a few seconds to create shapefiles from the run length codes for processes of this duration and size making interactive display and analysis of the objects feasible.

6. Conclusions

A framework has been proposed that builds on a dual representation approach to represent both fields and object-like features embedded in distributed geographic phenomena. The framework explicitly stores multiple boundaries to represent object-like features which inherently have fuzzy boundaries. By representing object-like features explicitly, the proposed framework provides a means to summarize patterns and structures in distributed phenomena and to maintain object identity for analysis of spatial and temporal relationships (such as *contain, part-of, previousID,* and *futureID*). The proposed framework categorizes object-like features embedded in distributed geographic phenomena as zones, sequences, processes, and events according to boundary thresholds and spatial and temporal continuity. A zone represents a spatial cluster of grid cells that meet a certain threshold. Temporally and spatially continuous

zones constitute a sequence, and continuous sequences form a process. All processes that span over a period of time comprise an event. While these processes may be disjoint spatially or temporally, collectively they persist over a certain period of time. Each of these features is associated with an attribute table.

Although features defined by different boundary thresholds can be derived automatically from original data collected for distributed phenomena, the use of multiple representations provides some advantages. It stores the complex set of relationships between the objects over time which would be impractical to derive on demand. By implementing this framework, centralized data holders could provide a means for distributed users to query and do some types of analysis with relatively little overhead and avoid duplication of effort. The case study has demonstrated the working of the proposed framework using a prototype developed in the study with digital precipitation array data from the Arkansas-Red River Forecast Center. The prototype implemented the proposed framework in the relational data structure contained in ArcView, and it is shown to support complex queries that involve containment and temporal constraints. While the framework could also be implemented in an object oriented data structure, the widespread familiarity with relational databases and the ease of construction and update make the relational database a reasonable choice to implement the framework.

The prototype framework has some limitations. It is organized on data at fixed spatial and temporal scales based on the spatiotemporal granularity of the observed snapshots. A basic assumption of the framework is that the behavior or change in the

conceptual entities is continuous at the spatial and temporal granularity of the snapshots (Galton 1997, Wilcox et al 2000). This assumption allows us to reasonably use the behavior of the object-like parts of the as a basis for queries and analysis. If this assumption is not met, then there is no logical basis for the combination of states into zones, and processes.

This paper did not investigate how scale and the semantics associated with boundary definitions interact. These relationships should be explored in future work. Future investigations might also include extending the proposed framework to work with data with a variable temporal resolution.

Acknowledgements

This research was funded by the National Imagery and Mapping Agency (NIMA) through the University Research Initiative Grant NMA202-97-1-1024. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the NIMA.

7. References

Arkansas-Red Basin River Forecast Center, 2002, ABRFC Precipitation Products, http://www.srh.noaa.gov/abrfc/pcpnpage.html

Blaschke, T., Lang S., Lorup E., Strobl J. and Zeil P., 2000, Object oriented images processing in an integrated GIS/Remote Sensing environment and perspectives for environmental applications, in *Environmental Information for Planning, Politics and the Public*, edited by A. Cremers and K. Greve (Marlburg: Metropolis-Verlag), pp. 555-570.

Burrough, P. A., 1996, Natural Objects with Indeterminate Boundaries, in *Geographic Objects with Indeterminate Boundaries, GISDATA 2*, edited by P.A. Burrough and A. Frank (London: Taylor & Francis), pp. 3-28.

Burrough, P.A. and McDonnell, R., 1998, *Principles of Geographical Information Systems*, (Oxford: Oxford University Press).

Buttenfield, B.P., Object-Oriented Map Generalization: Modeling and Cartographic Considerations, in *GIS and Generalization: Methodology and Practice.*, edited by J.C. Muller, J.P. Lagrange and R. Weibel (London: Taylor and Francis), pp. 91-105.

Couclelis, H., 1992, People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. In *Theories and methods of spatio-temporal reasoning in geographic space*, edited by A. U. Frank, I. Campari, and U. Formentini (Berlin: Springer Verlag), pp. 65-77.

Davis, C. and Laender, A., 1999, Multiple Representations in GIS: Materialization Through Map Generalization, Geometric, and Spatial Analysis Operations, *ACM-GIS 1999*, pp. 60-65.

Egenhofer, M., and Mark, D., 1995, Naive Geography, In *Spatial Information Theory: A Theoretical Basis for GIS*, edited by A.U. Frank and W. Kuhn (Berlin: Springer-Verlag), pp. 1-15.

Erwig,M. and Schneider,M., 1997, Vague Regions, *5th International Symposium on Advances in Spatial Databases (SSD'97)*, edited by M. Scholl and A. Voisard (Berlin: Springer), pp. 298-320.

Galton, A., 1997, Continuous Change in Spatial Regions, In Spatial Information Theory: A Theoretical Basis for GIS (Proceedings of International Conference COSIT'97), edited by S. C. Hirtle and A. U. Frank, (Berlin: Springer -Verlag), pp. 1-13.

Hagen, M., Bartenschlager, B., and Finke,U., 1999, Motion characteristics of thunderstorms in southern Germany, *Meteorological Applications*, 6, pp. 227-239.

Houze, R., Smull, B., and Dodge, P., 1990, Mesoscale Organization of Springtime Rainstorms in Oklahoma, *Monthly Weather Review*, 118, pp. 613-654.

Jones, C., Kidner, D., Luo, L., Bundy, L., and Ware, J., 1996, Database design for a multi-scale spatial information system, *International Journal of Geographical Information Systems*, 10, pp. 901-920.

Mountrakis, G., Aguouris, P., and Stefinidis, A., 2000, Navigating through hierarchical change propagation in Spatiotemporal Queries, In *Seventh International Workshop on Temporal Representation and Reasoning* (Nova Scotia: IEEE Press), pp. 123-131.

Parent, C., Spaccapietra, S., 2000, Database Integration: the Key to Data Interoperability, in *Advances in Object-Oriented Data Modeling*, edited by M. P. Papazoglou, S. Spaccapietra, Z. Tari, (Cambridge, Massachusetts: The MIT Press), pp. 221-253.

Parent C., Spaccapietra, S., Zimanyi E., 2000, MurMur: Database Management of Multiple Representations, *AAAI-2000 Workshop on Spatial and Temporal Granularity, Austin, Texas, July 30, 2000,* http://lbdwww.epfl.ch/e/publications/articles.pdf/AAAI-STgranularity.pdf

Peuquet, D., 1988, Representations of Geographic Space: Toward a Conceptual Synthesis, *Annals of the Association of American Geographers*, 78, pp. 375-394.

Rigaux, P. and Scholl, M., 1994, Multiple Representation Modelling and Querying, in *Proceedings of the International Workshop on Advanced Research in Geographic Information Systems, Monte Verità, Ascona, Switzerland,* edited by J. Nievergelt, T. Roos, H. Schek, and P. Widmayer, (Berlin: Springer-Verlag), pp. 59-69.

Schiesser, H., Houze, R., and Huntrieser, H., 1995, The Mesoscale Structure of Severe Precipitation Systems in Switzerland, *Monthly Weather Review*, 123, pp. 2070-2097.

Schmidt, J., Lawrence, B., Olsen, B., 2000, A Comparison of Operational Precipitation Processing Methodologies, NOAA Technical Memorandum *NWS SR-205,* http://www.srh.noaa.gov/abrfc/p1vol.html

Timpf, S., 1997, Cartographic objects in a multi-scale data structure, *Geographic Information Research: Bridging the Atlantic,* edited by M. Craglia and H. Couclelis (London: Taylor & Francis), pp. 224-234.

Vangenot,C., Parent, C., and Spaccapietra, S., 1999, Multiple representations and multiple resolutions in geographic databases, in *Proceedings of the Advanced Database Symposium (ADBS'99), Tokyo, December 6-7, 1999,* http://lbdwww.epfl.ch/e/publications/ADBS99.pdf.

Wilcox, D., Harwell, M., and Orth, R., 2000, Modeling Dynamic Polygon Objects in Space and Time: A New Graph-based Technique, *Cartography and Geographic Information Science,* 27, pp. 153-164.

Winter, S., 1998, Bridging Vector and Raster Representation in GIS, in *Advances in Geographic Information Systems,* edited by R. Laurini, K. Makki and N. Pissinou (Washington D.C.: The Association for Computing Machinery Press), pp. 57-62.

Yuan, M., 2001, Representing Complex Geographic Phenomena with both Object and Field-like Properties, *Cartography and Geographic Information Science,* 28, pp. 83-96.

Chapter 3

Object-Based Characterization of Surface Changes to Enhance Querying and Analysis in Geographic Information Systems

Object-Based Characterization of Surface Changes to Support Enhanced Querying and Analysis Capabilities in Geographic Information Systems

John McIntosh and May Yuan
Department of Geography
University of Oklahoma

Abstract

Regularly scheduled measurements from satellites and other remote sensors, such as weather radar, are acquiring ever increasing amounts of spatiotemporal data. These expanding datasets can potentially provide new insights about the processes and events that produce these data. Much of the newly available data measures phenomena with properties that vary across space. Examples include soil moisture and temperature, which are often referred to as distributed phenomena. Distributed phenomena often are characterized as surfaces. Nevertheless emerging features of extreme values in a raster field may be important indicators of the distributed phenomena. Current geographic information systems (GISs) characterize surfaces or predefined portions of the surfaces with global statistics such as mean, variance and range. Global statistics treat each grid cell independent of the others, and consequently coherent patterns of features within the field are overlooked. However, these patterns are the key to draw insights to how the surface phenomenon evolves. Explicit representation of these patterns and surface changes can enhance spatiotemporal querying and analysis in a GIS. Here we propose a general framework to characterize change in surfaces based on emerging patterns of features in the surface. We use and object-oriented approach to describe these features

and consequently develop object-based characterization to enhance query and analysis support for surface change. Indices describing the characteristics of the patterns and surface change are stored explicitly and these indices provide a basis for queries. Concatenation of these indices over time can facilitate queries about surface evolution that has not been fully addressed in GIS literature.

Key Words: spatiotemporal characterization, spatiotemporal query, field-object representation

1. Introduction

Characterization of objects based on shape, spatial relations, and location have long been used to help understand processes at work in the landscape (Wentz, 2000). Similarly, surface patterns and textural features such as variance, contrast, and entropy have been used to characterize distributed phenomena with properties that vary continuously across space. Spatial variation embedded in such a phenomenon forms a surface, and changes in the surface indicate the working of certain processes.

For both object and surface characterization, most of the tools and functions of GISs are geared towards static analysis (i.e. the shape of an object at a given time or descriptive summaries of a surface at a given time). To understand how a surface changes we must go beyond static analysis to examine the surface spatially and temporally.

Analytical needs for surface change have become pronounced as the production of geospatial data has increased exponentially over the last two decades. Most significantly is the increase in data through remote sensing for phenomena with surface characteristics. Examples include terrain, temperature, precipitation, and soil moisture. Much of the data is collected regularly in time creating temporal datasets consisting of snapshots that document change. Accompanying the increased availability of data has been an increased interest in the representation of time and dynamics in GIS models. A number of studies have proposed frameworks to characterize the movement, evolution, and relations of points and objects (Claramunt and Theriault 1996, Egenhofer and Al

Taha 1992, Galton 1995, Galton 2000). However, less research has been done on generic methods to characterize and represent change in phenomena. Central to such research is the question: How should change and behavior of distributed phenomena such as population density or air pollution be characterized and represented in a GIS?

Conventionally surface phenomena are represented as raster fields. Features identifiable within surfaces , such as zones of relatively high or low values, can be important indicators of surface change, but are not represented in the raster approach. This research investigates the use of these features to characterize spatiotemporal patterns in surface phenomena. We attempt to augment the traditional representation by characterizing changes in the features identifiable in temporally consecutive snapshot pairs. Such changes serve as building blocks to characterize and analyze processes that drive surface changes.

Here we propose a general framework to characterize spatiotemporal patterns in distributed phenomena based on indices representing the state and changes of identifiable features imbedded in surfaces. The framework is intended to be applicable to a wide range of distributed phenomena, but is tested using springtime rainstorm events over the state of Oklahoma from March 15, 2000 to June 15, 2000.

The next section reviews approaches to characterize continuous surfaces. In section 3, we propose a framework to characterize change between snapshot pairs of surfaces based on the form, distribution and behavior of emerging features (conceptual objects) imbedded in the surfaces. The fourth section describes the implementation of

the proposed framework using rainfall data. The final section identifies strengths and weaknesses of the proposed framework and discusses areas for future work.

2. Background

We begin with a review of methods used to characterize static surfaces (i.e. techniques that evaluate a snapshot independent of preceding and subsequent snapshots) because temporal sequences of these characteristics are useful to characterize processes or events in the surfaces. This is followed by a review of techniques used for characterizing change and events in continuous surfaces. Recognizing that distributed phenomena have both field-like properties (ie. surfaces) and object-like properties (ie. emerging features), our review considers characterization techniques for both conceptual views.

2.1 Characterizing Surfaces

Surface characterizations are important in many applications ranging from astronomy (Zhao 2000) to quality control for paper production (Abidi et al, 1999). Sadahiro (2001) identified three general approaches to analyzing surfaces: statistically based, mathematically based and object-based methods. Statistical characterizations typically describe surface texture and can include characteristics such as mean, median, variance, contrast, entropy, or spatial autocorrelation. Textural indices have been applied to a diverse range of geographic phenomena.

Mathematical approaches utilize an equation of known properties that best fits a surface. The equation type, coefficients and constants of the equation indicate

characteristics of the surface. For example if the best fit model of a surface is a first order polynomial equation, then the surface will be a plane and the coefficient and constant of the equation characterize the slope and base height of the plane. Higher order polynomials map more complex undulations. Mathematical approaches are often used to investigate characteristics of surfaces such as trends or delineate structures with noisy or incomplete data. For example, Likkason (1993) found that fourth degree polynomial equation was best able to represent long wavelength gravitational anomalies associated with uplifting of the mantle.

Object-based methods include qualitative approaches that characterize objects or portions of surfaces in terms of membership in a category emphasizing the distributed nature of the surface. One approach is to distinguish surfaces in terms of the presence or absence of surface features such as pits, peaks, slopes, cols, concave or convex hillsides (Abidi et al 1999, Okabe and Masuda 1984, Sadahiro 2001). Such features are conceptual objects in that they emerge from the surface and can also be of interest in their own right. The framework proposed here assumes that conceptual objects are the most important aspect of surfaces and that the form, distribution, and behavior of the conceptual objects can provide a basis for interpretation and analysis of surface change and evolution of the respective distributed phenomena.

2.2 Characterizing Conceptual Objects in Surfaces

The shape, orientation and distribution of conceptual objects imbedded in surfaces have long been utilized in geographic analysis. In biogeography, shape characterization and distribution of features in density or probability surfaces have been

important in understanding species distributions. For example Ruggiero et al (1998) used equiprobalistic surfaces to define geographic ranges of species in South America. The ranges were then evaluated in terms of the shape, area, and location relative to mountains and Oceans. Dieleman and Mortensen (1999) incorporated shape measures in their investigation of the evolution of weed patches which were defined based on a seeding density surface. They evaluated the weed patches based on area, orientation, distribution of patches, and weed density within the patches. Furthermore, Yuan and Perault (1998) applied fractals to evaluate landscape change in the Olympia National Forest.

Shape and distribution of conceptual objects in continuous surfaces have also played an important role in climatological investigations. Hagen et al (1999) identified three classes of thunderstorms in southern Germany based on object-like characteristics of lightning density surfaces. Storms were classified on how isolated the cells (zones in a density surface with relatively high rates of lightning strikes) were, whether cells were linearly aligned, the direction of movement relative to the major axis of aligned cells, and characteristics of the leading edge. In the U.S., Houze et al (1990) evaluated severe springtime rainstorms in Oklahoma based on radar reflectivity surfaces. Storm organization was graded according to the degree to which conceptual objects perceived in the reflectivity surface matched an idealized model of a leading line/trailing stratiform structure.

In artificial vision and pattern recognition, characterization of raster images (e.g. video images) is an important component. This often involves analysis of shape of

imbedded conceptual objects. However, the goal is generally to categorize objects, which are defined a priori, rather than exploring characteristics of surfaces. Consequently, a significant emphasis in pattern recognition work placed on characterizing the boundaries or edges of objects, which are assumed to be distinct (Forte and Greenhill 1997, Gunsel and Tekalp 1998, Huang et al 2000, Kuappinen et al 1995, Kumar and Georgou 1990, Sako and Fujimura 2000). In contrast, many of the conceptual objects in distributed geographic phenomenon do not have well defined boundaries. As a result, methods of object identification and comparison are often of less utility for geographic analysis.

2.3 Characterizing change in Distributed Phenomena

All three approaches to characterize surfaces (section 2.1) are also of interest for analysis of processes or events in the fields. Mathematical and statistical approaches each have equivalents for characterizing spatiotemporal patterns. Physically based models have long been used to describe the behavior of dynamic phenomena in physics and meteorology for example. Statistical methods also can be applied to characterize spatiotemporal behavior of surfaces over time. For example, Dielman and Mortensen (1999), in addition to using object-based descriptors of the weed patches, also used lag correlation to investigate surface change over time.

Object-based approaches can also be used to describe changes in surfaces over time. For example Sadahiro (2001) proposed a qualitative framework to model change in surfaces representing population density based on the appearance, disappearance and change in topology between *peaks*, *slopes* and *bottoms* of the surface. Considerable

research in GI Science has been directed at developing generic qualitative descriptors to describe change (Egenhofer and Al Taha 1992, Claramunt and Theriault 1996, Hornsby and Egenhofer 1997 and 2000, Wilcox et al, 2000). While much of this work focuses on change in objects, some of the approaches may be applied to conceptual objects within surfaces. For example Claramunt and Theriault (1996) proposed a semantic model of spatial change based on describing events involving objects using sets of "process primitives". These primitives are arranged in temporally organized sequences and represent events in terms of basic changes such as appearance, disappearance, expansion, contraction, deformation, displacement and rotation. Similarly, Hornsby and Egenhofer's Change Detection Language (1997 and 2000) provides explicit qualitative description of change relating to objects. This language focuses on the appearance/disappearance of entities and on identity transitions from one object to another over time. The Spatiotemporal Graph Model (Wilcox et al 2000) facilitates analysis of processes recorded as series of regular temporal snapshots by explicitly storing topological changes in polygon objects including *create, continue, lost, split,* and *merge.*

Characterizing change and events has also been an active research topic in computer science. Recent research in image analysis has focuses on dynamic objects. As with static images, the goal typically is to create temporal image retrieval systems that will find images related to a particular event. For example, Srinvasa and Abuja (1999) developed an artificial neural network (ANN) based on correlated features for retrieval of series of images. Chang et al (1998) used a hierarchy of indices including visual and

textural features describing the principal objects and actions to retrieve scenes from video.

The review in section 2.1 and this section suggest a wide range of possible approaches to characterize and represent change. Approaches favored in image processing and artificial vision tend to rely on metrics that are good at categorizing but provide little semantic meaning. For example coefficients from a Fourier transform of a boundary representation (Kumar and Georgou 1990, Kuappinen 1995) may provide a basis to successfully assign a given shape or temporal patterns to categories, but provide little support for interpreting new patterns, which is important in geographic investigations.

Alternatively, the GIScience research tends to focus on qualitative aspects of shape and behavior. These qualitative approaches provide semantic information that is valuable for understanding patterns. However, qualitative descriptions are categorical (e.g. "growth" or "no growth"). There is often a continuum within categories that may be missed by strictly qualitative measures. For example, a qualitative description such as growth applies equally to an object that increases size by 0.01% as to an object that increases size by 110%. Clearly there is a continuum. The metric of percentage increase or the absolute change in area complements the semantic information in the qualitative description.

3. Proposed Indices for Object-Based Characterization of Surface Change

In this paper we take the conceptual view that surfaces can be described by the presence and behavior of emerging features (ie. conceptual objects) in the surface. In contrast to Sadahiro (2001), objects are not merely a means of describing the surface but are assumed to possess intrinsic value in their own right. People perceive these objects and analyze their properties to improve understanding of processes that generate these objects. For example, plumes, areas of population density (urban areas), or jet streams can be perceived and reasoned as conceptual objects.

The state and behavior of conceptual objects imbedded in surfaces serve as the basis of our proposed indices. Our work complements the work of Yuan (2001) and Mennis et al (2000). While the frameworks used by Yuan and Mennis et al represent object-like characteristics within distributed phenomena, characterization of the behavior of the objects is limited. Our proposed representational framework is intended to provide building blocks to characterize spatiotemporal patterns and behaviors of conceptual objects imbedded in the surface of a distributed phenomenon.

The review in the section 2 suggests that there is a perceptual dichotomy between object and surface based indices. Suites of indices generally characterize either object-like properties or surface characteristics, but not both. Geometric and topological indices typically describe shape and other boundary properties of objects. Indices used for fields typically describe global patterns or characteristics such as variance and spatial autocorrelation. In many applications, a combination of object- and field-like properties provides a more complete representation than an object or field based index alone. We

may perceive and model a mountainous region in an elevation field as an object. A characterization based on the extent of the mountainous area is incomplete without a characterization of objects like ridges, slopes and aspect within the mountainous region.

Measures that are invariant to factors such as location, size, orientation, or time are desirable in designing indices for object characterization (Gardoll et al 2000, Wentz 2000). This invariance allows the indices to be applicable to a wide range of phenomena regardless of location, scale, or time. While the modeled area may be fixed, location invariance is needed because conceptual objects may occur throughout the modeled domain, and change in distributed phenomena is often not fixed to a specific location. For example in meteorology, areas of high or low pressure may occur throughout the modeled domain. Conceptual objects such, as jetstreams and anticyclones, shift location from time to time. Perceived similarity of conceptual objects is based on other criteria such as the magnitude of the high, the shape of the high pressure area (e.g. a pressure ridge), movement and spatiotemporal relations to other objects. Orientation invariance, like location invariance, is desirable because absolute orientation of perceptual features is often less important to the orientation relative to some other consideration such as movement direction. For example the absolute orientation of a pressure ridge relative to the cardinal directions is assumed to be less important than the orientation of the major axis relative to the direction of movement.

Temporal invariance allows spatiotemporal similarity to be independent of the absolute time in which the event occurred. We assume that change relative to the timeframe of the event is more important than the absolute time when processes

occurred. Size invariance is included in order to maintain generic indices that can be used for phenomena at a variety of scales. We assume that important spatiotemporal properties are of use in characterizing phenomena at a variety of spatial scales. Therefore example characteristics about the form of the conceptual objects, such as the elongation, are valid for small scale objects as well as large scale objects.

3.1 Design Basis

Our indices are designed to characterize surface change using conceptual objects. We assume that both the characteristics of objects and the relationships between objects in the surface are important. Hence, we include descriptions of both objects and object relationships. Both static and dynamic descriptors are used. Static descriptors describe the current state of the object or object relationships and provide a baseline for the dynamic descriptions. While we do not explicitly model change in the static indices, these indices can be part of an overall characterization of the objects history when they are linked together in temporal sequences. Finally although the proposed indices describe characteristics of the conceptual objects, field-like properties that reflect spatial variation within a conceptual object are also considered when appropriate for the design of the proposed indices. Table 1 summarizes the design basis for object-based characterization of surface change. The following sections list the proposed indices and discuss in detail their computation and utility.

Table 1 - Conceptual Framework

	Objects	Object Relationships
Static	Characteristics such as shape or orientation of objects	Characteristics such as distribution of objects
Dynamic	Characteristics such as growth or the granularity of change of in objects	Characteristics such movement of the objects relative to other nearby objects

3.1.1 Object Characterization

We use two measures, *elongation* and *orientation,* to characterize the state of individual conceptual objects. Elongation is a measure of the overall length of the object in relation to the width. Orientation is a measure of the direction that the longest axis of the object. Elongation only provides a very general description of shape, and there are many other shape descriptors available such as squareness, edge characteristics or the fractal dimension that can provide a more refined characterization of objects. While it may make sense to use these indices with discrete objects, conceptual objects imbedded in fields often have indeterminate boundaries, so indices based on the details of the boundaries are less desirable. Elongation applies to objects with both distinct and indistinct boundaries and captures the general form of the objects.

There are several methods to calculate elongation and orientation. One of the most common approaches is to use the orientation and dimensions of the minimum bounding rectangle. For shapes with more complex boundaries the ratio of the major and minor axis of the best ellipsoidal approximation is a good alternative. We define

elongation based on the principal moments of inertia calculated from all values in the object. Elongation (E) is defined as follows

$$E = \frac{(I_{Max} - I_{Min})}{(I_{Max} + I_{Min})} \quad (1)$$

where E is the elongation index of an object and I is the rotational inertia about the axis of rotation (Gardoll et al 2000). Inertia varies depending on the orientation of the axis of rotation. The inertia (I) at a given angle (θ) is based on the sum of the contribution of each cell in the object using the following equation

$$I_\theta = \sum_{i=1}^{n} r_i^2 z_i \quad (2)$$

where r is the distance to the axis of rotation and z is the value of the grid (Giancoli 1984). Inertial is evaluated at one degree intervals to find the minimum and maximum values.

In Figure 1, the angle θ describes the orientation of the measured shape along the minimum magnitude of inertia axis. Theta (θ) is relative to the cardinal directions. Relative orientation of objects can be more useful than absolute orientation in some cases. For example, in meteorology, the relative orientation of a band of precipitation to the direction of movement may provide insights that cannot be obtained with orientation based on a fixed reference frame. A band or rain parallel to the direction of movement is associated with quite different events than a band of rainfall with its

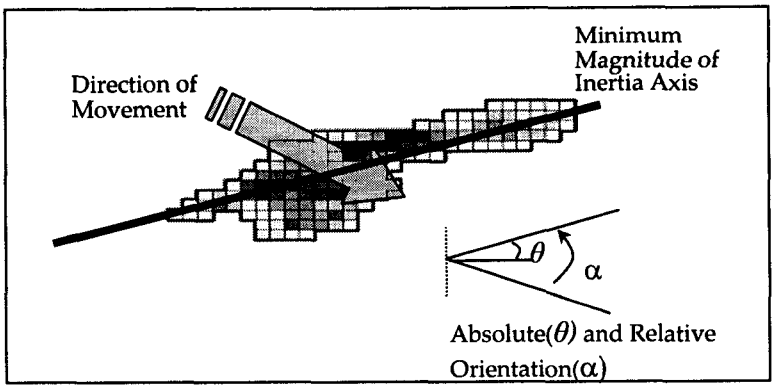


Figure 1. Orientation of a Conceptual Object

orientation perpendicular to the direction of movement, even if the absolute orientation of both examples is the same.

The orientation index (*O*) is calculated as the difference in degrees (α) between the direction of movement and the orientation of the major axis moving in a counter clockwise direction. The relative direction is normalized by 180 so the index value varies between 0 and 1. A value of 0.5 occurs when the major axis is perpendicular to the direction of movement and a value of 0 occurs when the orientation is parallel with the direction of movement. In cases where the object exists in a single time period and there is no movement, the orientation index is set to null.

We define two indices to characterize change in a conceptual object. The first index is percent growth and the second is granularity of change. The growth index (G) is calculated as the change in area from time t_i to t_{i+1} relative to the area at time t_i.

$$G = \frac{Area_{t2} - Area_{t1}}{Area_{t1}} \tag{4}$$

The index is positive for growth and negative for a decrease in area. Disappearance of the object corresponds to an index value of -1.

Change in conceptual objects between time steps may involve non-uniform change to many small areas or more uniform change involving larger areas. The goal of the granularity of change index is to characterize the type of change based on the dominant spatial scale at which it occurs. Many geographic phenomena have been observed to operate at distinct spatiotemporal scales. Numerous methods have been proposed to identify the operational scales of the phenomena to assist in diverse activities such as the best scale of observation, appropriate scales for modeling, or appropriate resolution for remote sensing. One of the earlier methods proposed to quantify the dominant scale was developed by Tobler and Moellering (1972). Their method proposes that variance will peak at spatial scales that best represents the dominant processes. The geographic variance is calculated using the population variance of the layer

$$Var_{pop} = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n^2} \qquad (4)$$

where n is the number of spatial units and x is the value of a unit.

The geographic variance method was originally proposed for hierarchical data. Woodcock and Strahler (1987) proposed a local variance method for remote sensing. Their method is similar to the geographic variance method but does not require a strict hierarchy (Cao and Lam, 1997). For this approach, local variance is defined as the

average variance calculated by passing an overlapping 3x3 window over the entire image. The average local variance is calculated as follows

$$\overline{Var}_{local} = \frac{\sum_{i=1}^{n} \text{var}_i\ 3x3}{n} \qquad (5)$$

where n = the number of 3x3 windows and var_{3x3} is the variance within the analysis window. In order to find the resolution of the operational scale, the cells on the surface are aggregated and the average local variance is calculated. The cell resolution with the maximum average local variance is assumed to be the operational scale of the scene. At coarser resolutions, variation within the aggregated spatial units is lost and the variance goes down. At finer resolutions, the average variance is lower because the 3x3 windows do not cover large enough areas to capture much variation. By aggregating the cells and testing the average local variance, an optimum spatial resolution can be found that is a balance between the variance represented and the number of cells used to represent the variance. This is the resolution that will resolve most of the variation.

Our granularity of change index (R) is based on the local variance method applied to the change in the surface values of the conceptual objects. A difference grid is calculated to show the change in states of a conceptual object at the current and next time step. The average local variance is calculated using overlapping 3x3 windows at increasingly coarser grid resolutions. Coarser grid resolutions are achieved by aggregating cells with the average value of the component cells. As the grid resolution decreases, a cell covers a greater area and its value approaches the average value of the conceptual object. Our algorithm stops at the resolution when the average variance of

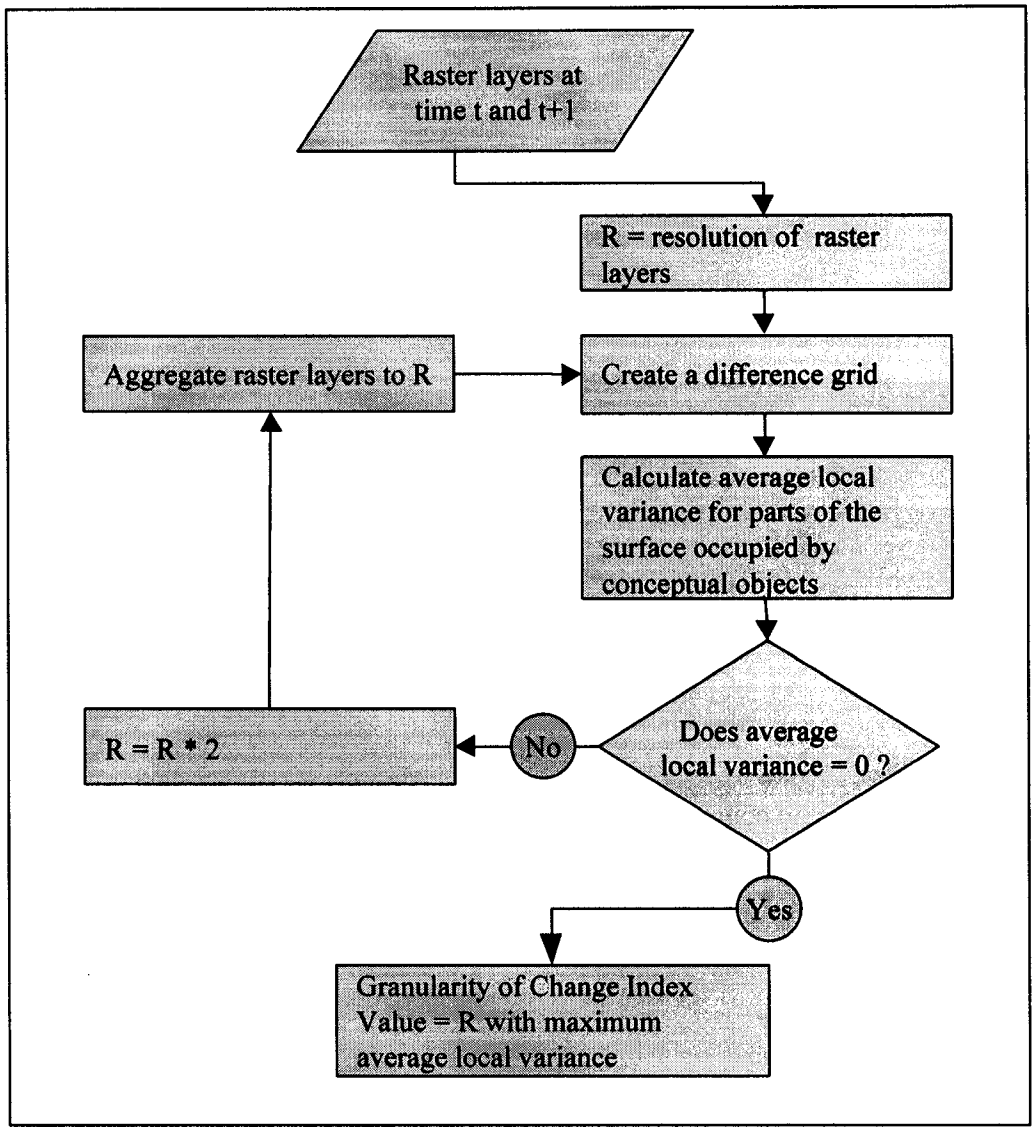


Figure 2. Granularity of Change Index Calculation

change within the conceptual objects reaches zero (Figure 2). This method is designed to identify the optimal resolution to capture change of the surface of the conceptual objects. In some cases the operational scale may be at resolutions finer than the observation in both the spatial and temporal domains. In such cases the maximum variation will occur at the resolution of the input layer. Local maximums may also occur at more than one resolution. In such cases, the peak with the greatest variance is used as the index value.

While the other peaks are not used for this index, they could be used as additional attributes for domains with processes or events that typically operate at multiple resolutions.

3.1.2 Object Relationships

We use two indices to characterize object relationships. For our static characterization, we use a *distribution index* that characterizes the general distribution of conceptual objects, and for the dynamic characterization we use a *movement variation index* that characterizes the variation in velocity these objects. The distribution index is intended to characterize the overall distribution of conceptual objects within the surface. We use a nearest neighbor approach based on the relative location of each conceptual object's centroid. As the degree of clustering increases, the average distance between the objects and their nearest neighbors decreases. Likewise, as the objects become more dispersed, the average distance between their centroids and their nearest neighbor increases. This index is based on the average nearest neighbor distance between each possible pair of centroids.

While the average nearest neighbor distance provides a means to characterize spatial distribution, it is specific to the number of points and the area in which the points are distributed. We standardize the average nearest neighbor distance by relating it to the expected average nearest neighbor distance in a random distribution of the same number of points in the given area. The standardized distribution index (D) is the ratio of the average nearest neighbor distance ($\overline{NND}$) divided by the estimated average random nearest neighbor distance ($\overline{NND}_R$).

$$D = \frac{\overline{NND}}{\overline{NND_R}} \qquad 6)$$

The average nearest neighbor distance is calculated using the basic formula

$$\overline{NND} = \frac{\sum_{i=1}^{n} NND_i}{n} \qquad (7)$$

where n represents the number of points. We calculate the average random nearest neighbor distance using the method described in McGrew and Monroe (2000).

$$\overline{NND}_R = \frac{1}{2\sqrt{d}} \qquad (8)$$

where density (d) equals the number of points divided by the area. A random spatial pattern will result in an index value of 1. Index values of greater than one are more dispersed while index values of less than 1 are more clustered.

The distribution index describes the general distribution of the objects at specific times, but does not describe how the objects move relative to each other. For example, whether the difference in velocity (speed and direction) varies significantly over short distances or whether nearby objects tend to have the same patterns of movement. We

define the movement variation index to characterize the relative movement of the conceptual objects. This index is analogous to the concept of wind shear used in meteorology where shear (S) is defined as the difference in velocity between 2 points divided by the distance between those points

$$S = \frac{\left|\vec{V}_{diff}\right|}{d} \tag{9}$$

where were $\vec{V}_{diff}$ is the difference in velocity and d is the distance between the two points (centroids of the conceptual objects). In this index we compare the velocity of each object with the velocity of other objects at the same time to find the maximum shear value. The movement variation index (MV) is defined as the average maximum shear value for each conceptual object

$$MV = \frac{\sum_{i=1}^{n} S\max_i}{n} \tag{10}$$

where *Smax* is the maximum shear for an object and n is the number of objects at the given time. The minimum MV index is zero which corresponds to spatial objects with identical velocity (ie. moving parallel at the same speed).

3.1.3 Calculating Index Values for a Hierarchy of Conceptual Objects

When surface changes, conceptual objects may be aggregated spatially or temporally to additional conceptual objects at a higher level of abstraction. As a consequence, conceptual objects at multiple levels of abstraction form a hierarchy in which spatial and temporal characteristics manifest in the surface. The indices we

defined can be applied to measure conceptual objects at all levels of the hierarchy and furthermore relate their spatiotemporal characteristics across different levels of abstraction.

4. Case Study

We use hourly rainfall in the Southern Plains, USA to test the proposed indices. Rainfall exhibits distinct spatiotemporal patterns that possess both object- and field-like properties making it a good test phenomenon for the proposed indices. Data used in the case study are digital precipitation arrays (DPA) from the National Weather Service's Arkansas-Red River Forecast Center (ABRFC), covering the entire state of Oklahoma and portions of surrounding states (Figure 3). The DPAs are in a raster format and consist of approximately 4km x 4km grids in the Hydrologic Rainfall Analysis Project (HRAP) coordinate system and are archived in the NetCDF format (Arkansas-Red Basin River Forecast Center, 2002). Each DPA contains the distribution of hourly accumulated rainfall estimates based on a composite from next generation radars (NEXRAD) and observations at ground weather stations (Schmidt et al. 2000).

The DPAs are generated in real time and are used by the ABRFC for flood forecasting. They can also be valuable for other purposes such as climate analysis, risk assessment, facilities planning and agronomy. The framework was tested using rainfall data from March 15, 2000 to June 15, 2000. We developed several Java scripts to download and convert the DPAs to an ArcInfo® grid format. Indices were calculated

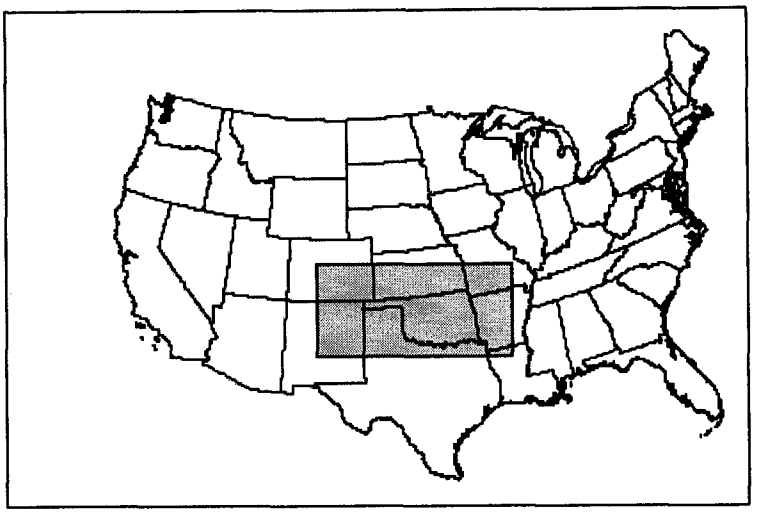


Figure 3. Arkansas-Red Basin River Forecast Center
Digital Precipitation Array (DPA)

using the ArcView® GIS 3.2 (Environmental Systems Research Institute Inc., Redlands, California) scripting language, Avenue™.

Two thresholds were used to delineate rainfall "zones" (>0mm/hour and >20mm/hour). Houze et al (1990) identify several important structural aspects of springtime rainstorms, including areas of light stratiform precipitation, areas of intense rainfall corresponding to convective cells, and areas of heavy precipitation indicating features such as a squall line. The thresholds selected for the case study are intended to capture these features in DPAs.

Starting with zones, we construct a hierarchy of conceptual objects of sequences processes, and events through spatiotemporal aggregation. Rainfall zones are linked to form sequences based on overlaps in consecutive snapshots (DPAs). Rainfall processes link sequences related by merges or divisions and their predecessors and descendents.

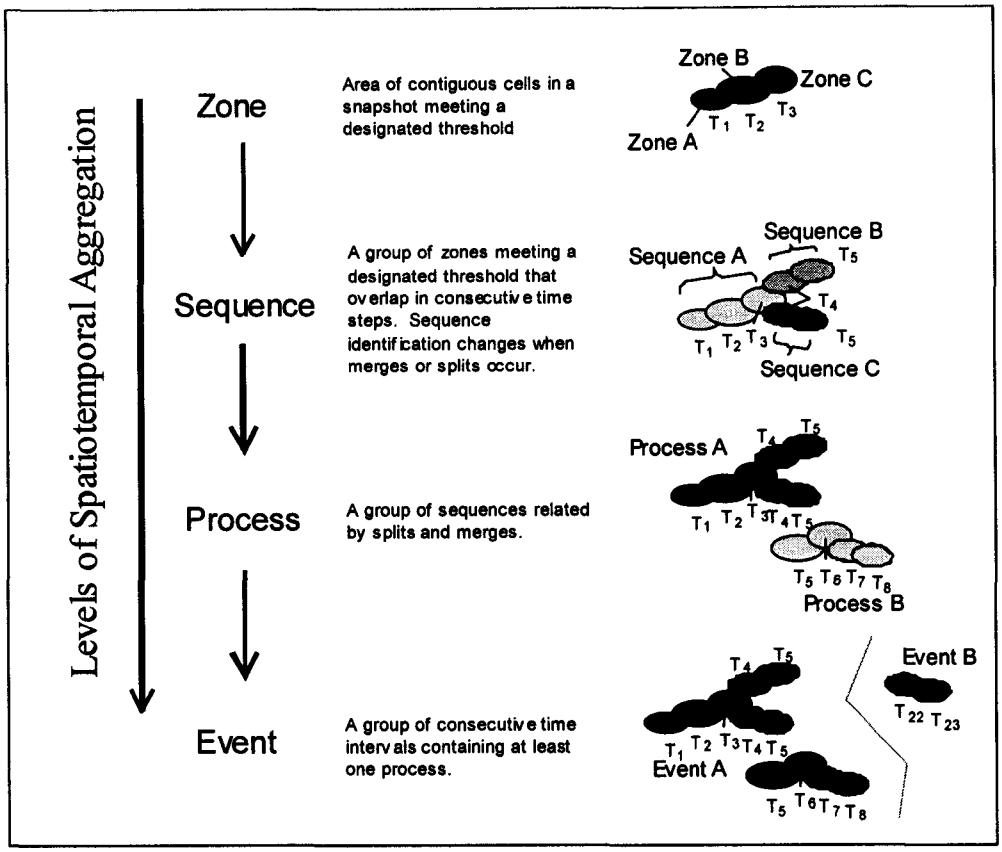


Figure 4. A hierarchy of conceptual objects

Rainfall events represent consecutive periods with rainfall somewhere in the study area (Figure 4).

The indices were calculated to characterize surface change based on these conceptual objects and their change from the given time to the next. The dataset consists of 1567 temporal pairs of snapshots (DPAs at subsequent times) containing 40870 conceptual objects in the study area between March 15, 2000 and June 15, 2000.

We investigate the indices using process and event objects. The indices associated with processes characterize change in the rainfall surface the indices associated with events characterize change in the surface of a storm system. Object-like

indices (elongation, orientation, growth, and granularity of change) are measured at the sequence level in our implementation. For processes and events, the object-like indices are the average index values of component sequences. Since sequence objects only include a single zone at any given time period, the distribution and relative movement indices are only calculated for process and event objects, which have the potential for more than a single zone in a given time period. The distribution index calculation requires area. For the event objects, we use the entire study area. For process objects, the area of the convex hull around the zones associated with the process at the given time is used.

4.1 Evaluation

We apply an approach similar to Wentz (2000) to evaluate the proposed indices. Specifically we would like to answer the following questions:

- Does each index describe a unique aspect of the spatiotemporal patterns?
- Are the indices descriptive enough to cluster snapshot pairs that are similar within the cluster and dissimilar with other clusters?

Ideally, each member of the suite of indices should characterize a unique aspect of surface change. Wentz (2000) suggested that this occurs when there is no correlation between any pairs of indices. A correlation analysis was performed on the indices to determine the efficiency of the characterizations. We tested the correlation of the index values for both process and event objects. Tables 2 and 3, which show correlation for all possible index pairs, demonstrate that there is little overlap among the proposed indices. Of the 15 possible index pairs for event objects, the highest correlation is between object

Table 2. Index Cross Correlation Matrix for Processes

	Elongation	*Orientation*	*Growth*	*Granularity of Change*	*Distribution*	*Relative Movement*
Elongation	1.000					
Orientation	-0.096	1.000				
Growth	-0.088	0.001	1.000			
Granularity of Change	-0.158	-0.023	0.241	1.000		
Distribution	0.003	0.014	-0.106	-0.001	1.000	
Relative Movement	0.167	-0.033	-0.022	-0.251	-0.009	1.000

Table 3. Index Cross Correlation Matrix for Events

	Elongation	*Orientation*	*Growth*	*Granularity of Change*	*Distribution*	*Relative Movement*
Elongation	1.000					
Orientation	-0.184	1.000				
Growth	-0.085	0.031	1.000			
Granularity of Change	-0.123	0.033	0.009	1.000		
Distribution	0.217	-0.156	0.151	-0.321	1.000	
Relative Movement	-0.126	0.052	0.103	-0.128	0.233	1.000

movement variation and the granularity of change index (r= - 0.25). While much higher than the other values, the correlation coefficients suggest that the common information between these indices is less than 7%. The index values at the process level are slightly more correlated. The strongest relationship (r= -0.321) is between the distribution index and the granularity of spatial change index.

Given the large number of index pairs, even combinations with relatively low correlations are statistically significant. Critical values were calculated to differentiate between the statistically significant correlations using an alpha of 0.05. Because the index values are time series and exhibit some autocorrelation, the significance can be overestimated. To avoid this, an effective sample size based on the degree of autocorrelation was used in the determination of the critical values. The effective sample size was calculated for both sets of index values using the following formula

$$n' = n\left(\frac{1-\rho}{1+\rho}\right) \tag{11}$$

where n' is the effective sample size, n is the actual sample size, and ρ is the lag-1 correlation of the time ordered index values (Wilks 1995). To be conservative, the minimum of the effective sample size was used as the degree of freedom to determine the critical values. Using this approach for the process indices, elongation had a statistically significant correlation with the granularity of change and relative movement indices. Granularity of change also had a statistically significant correlation with distribution and relative movement.

For the event indices, all correlations were significant except for (1) elongation and growth; (2) orientation and growth granularity of change and relative movement; and (3) growth and granularity of change. While many of the relations between the various indices are statistically significant, the variance explained is still low suggesting that the indices characterize unique aspects of the spatiotemporal patterns in surfaces.

To determine if the proposed indices sufficiently discriminate among spatiotemporal conceptual objects, a k-means cluster analysis was performed to generate groups of similar temporal pairs of snapshots based on event objects. The indices were converted into standard anomalies prior to clustering. This was to avoid indices with higher variance dominating the clustering process (Wilks 1995). K-means is a nonhierarchical clustering technique that requires the final number of clusters to be selected a priori. We tested the data iteratively using an approach that compares the

sum of the squares between successive numbers of clusters to find the optimum number of clusters. This method suggests that it is justifiable to add another cluster when

$$\frac{\sum wss_k}{\sum (wss_{k+1} - 1)} * (n_{row} - k - 1) > 10 \qquad (12)$$

where wss_k is the sum of squares of the members of clusters from the mean with k clusters, wss_{k+1} is the sum of squares of the members of clusters from the mean with k +1 clusters and n_{row} is the number of rows in the data set (Hartigan 1975, SPLUS Language Reference, 6.1.2 Release 1). Hierarchical (complete linkage) clustering was used to generate the initial cluster centers.

K-means analysis suggests twenty-five clusters of events in the DPA data set. Figure 5 shows the normalized cluster centers and Table 4 provides summaries of the cluster groups. Figures 6, 7 and 8 provide some examples from several large cluster groups with conceptual objects defined as rainfall exceeding 2mm/hour. A preliminary inspection reveals a similarity in the shape and orientation among the clusters, which is not unexpected given the dominant weather patterns during the spring time in the study area. For example clusters 1 and 3 are characterized by elongated zones oriented from southwest to northeast of roughly the same size. While these clusters look similar, legitimate distinctions are captured. The zones in cluster 1 move roughly perpendicular to the axis of orientation, while the movement in cluster 3 is more or less in parallel to the axis or orientation. Some of the examples from cluster 5 also contain elongated zones. However, the less clustered distribution dominates.

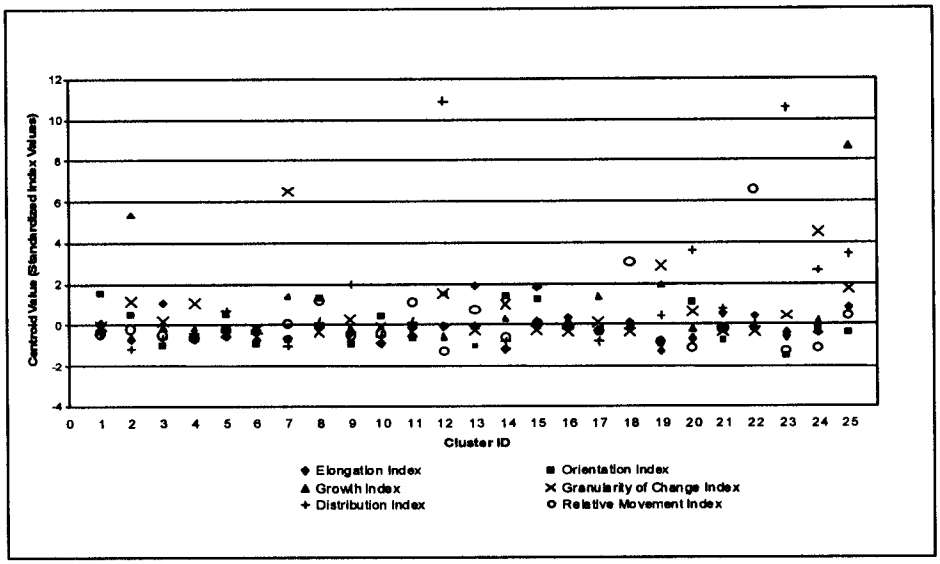


Figure 5. Twenty-five Clusters from K-means Analysis

Table 4. Cluster Characteristics

Cluster ID	Number of Snapshots per Cluster	Within Cluster Sum of Squares	Within Cluster Standard Deviation	Average Number of Zones per Snapshot	Average Zone Area	Standard Deviation of Zone Area
1	110	140.782	1.136	24.736	120.818	786.338
2	6	25.810	2.272	19.500	12.752	26.586
3	105	143.762	1.176	20.600	94.591	520.683
4	92	133.299	1.210	31.250	98.992	585.337
5	160	189.588	1.092	38.025	144.981	758.246
6	145	152.122	1.028	34.159	122.677	841.379
7	2	0.509	0.713	15.000	9.033	18.320
8	76	108.259	1.201	47.237	161.804	995.984
9	29	59.957	1.463	17.448	252.538	1194.179
10	127	114.333	0.953	33.315	95.068	621.659
11	121	126.711	1.028	51.694	138.115	785.056
12	3	10.881	2.332	3.000	508.222	1042.107
13	79	124.008	1.261	36.494	159.951	979.361
14	47	106.510	1.522	29.489	113.585	663.864
15	75	120.725	1.277	31.347	162.880	873.332
16	163	116.090	0.847	42.736	88.422	589.276
17	45	73.207	1.290	23.222	35.100	311.512
18	39	90.328	1.542	49.974	167.241	1056.358
19	10	41.486	2.147	31.100	75.621	327.041
20	10	28.519	1.780	10.300	240.806	888.911
21	107	115.388	1.043	36.243	154.141	781.364
22	8	30.330	2.082	33.875	320.723	2128.675
23	2	2.063	1.436	3.500	1110.714	1822.276
24	5	24.245	2.462	14.400	62.083	140.267
25	1	0.000	--	109.000	12.220	39.704

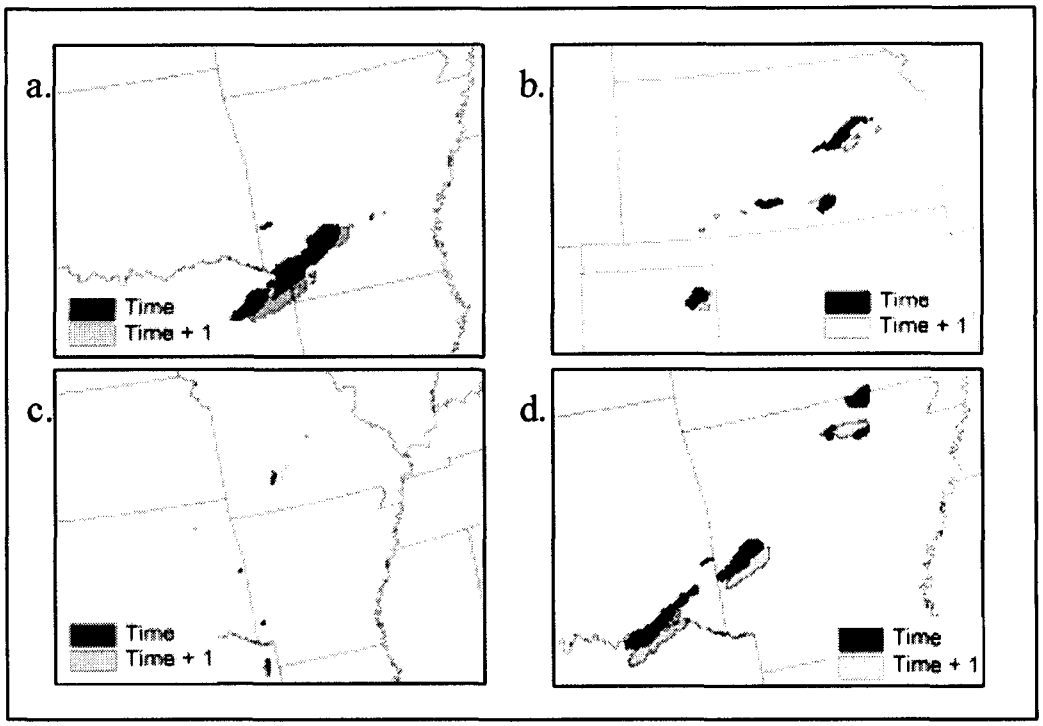


Figure 6. Examples from Cluster 1

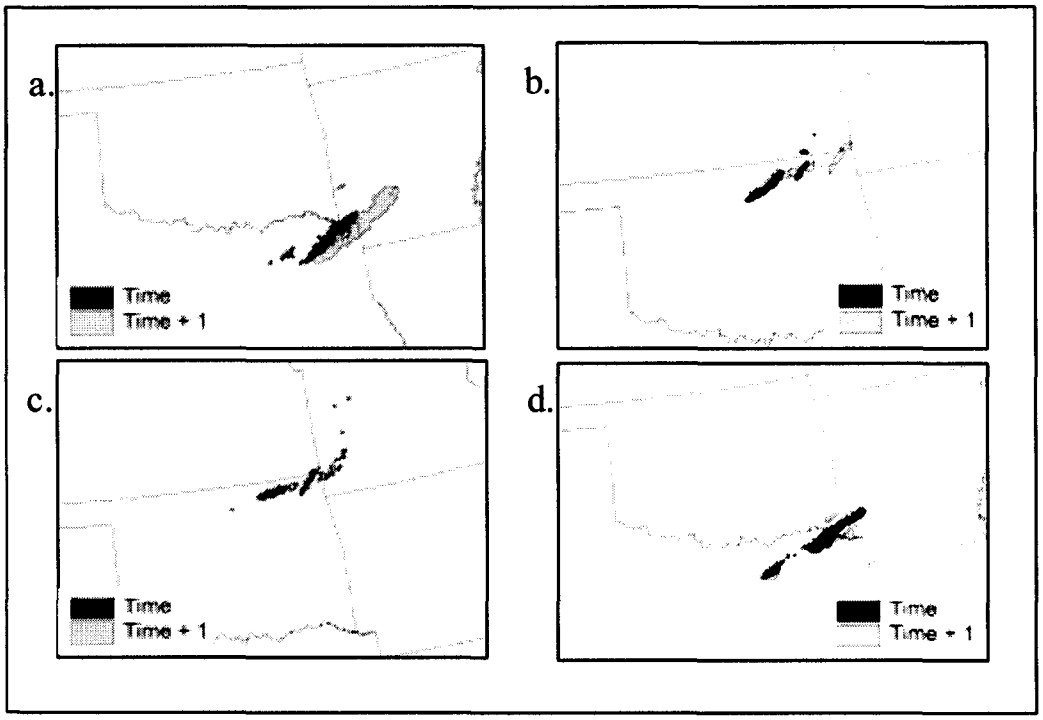


Figure 7. Examples From Cluster 3

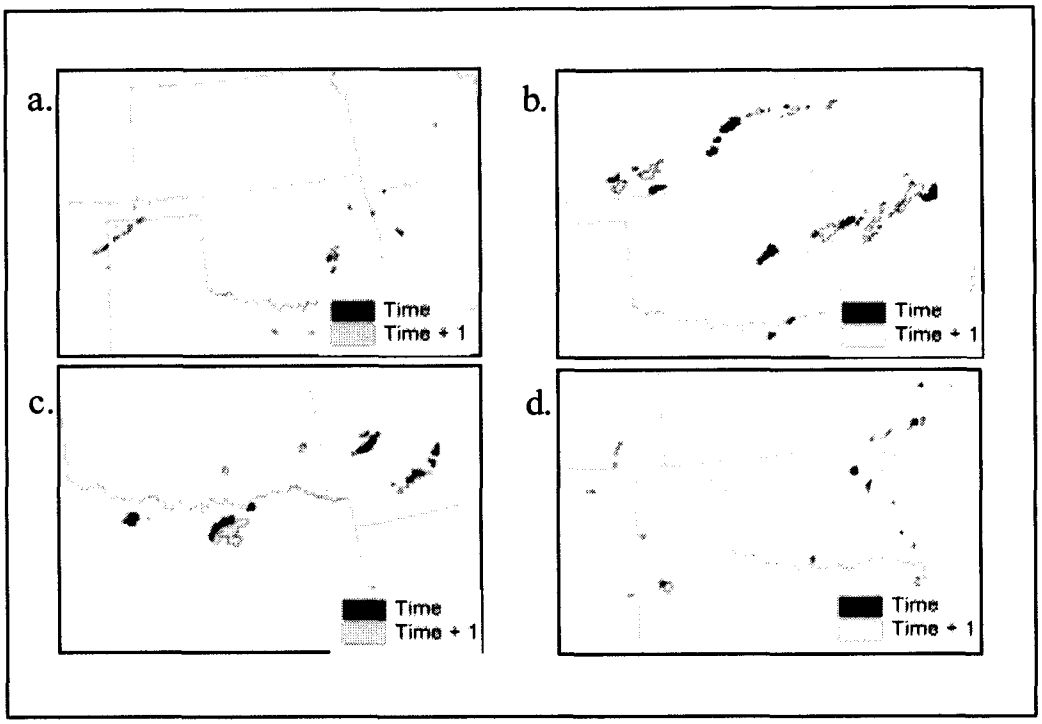


Figure 8. Examples From Cluster 5

5. Conclusions

A suite of indices has been proposed to characterize spatiotemporal patterns in distributed phenomena at discrete intervals in time. The basis of the indices is the idea that surfaces of distributed phenomena may be characterized based on spatiotemporal patterns of conceptual objects imbedded in the surface. Our indices expand upon those established in many disciplines, such as geography, ecology and meteorology.

A case study was conducted to implement and evaluate these indices. The study used hourly rainfall data from the Arkansas-Red River Forecast Center. Zones, of contiguous cells meeting a 0 mm/hr or 20 mm/hr threshold, were explicitly represented. Sequence, process, and event database objects based on temporal collections of the zones

were also explicitly represented in the database. Six indices based on object-like characterizations and object relationships were proposed.

Two criteria were proposed to judge the usefulness of these indices: (1) The framework should be efficient with little overlap between the information characterized in the indices, and (2) that the framework should cluster examples of the spatiotemporal pattern primitives that are intuitive. The results of the case study suggest that the proposed framework meets the first condition. The correlation analysis shows that there was very little overlap between the indices at both the event and process level. The cluster analysis shows that these indices effectively group cases that exhibit similar spatiotemporal characteristics. The majority of the dataset are included in the most dominant clusters (about 66% in the larges 8 clusters). This is probably due in part to the fact that the cluster represents dominant patterns of springtime rainstorms in the study area. The combination of indices is not robust to outliers and some of the smaller clusters are dominated relatively extreme index values. Figure 9 shows examples from three of the smallest clusters - 12, 23 and 25. Clusters 12 and 23 have an extremely high distribution index that indicates that the zones are highly dispersed. This is an artifact of the band of precipitation along the western extent of the domain. Cluster 25 is also a result of error in the dataset.

This paper did not address how the spatiotemporal pattern primitives can be linked together to characterize whole events or processes, which requires multiple snapshots that cover the entire lifecycle of a rainstorm event. Future work should investigate how the overall similarity of entire processes or events represented by these

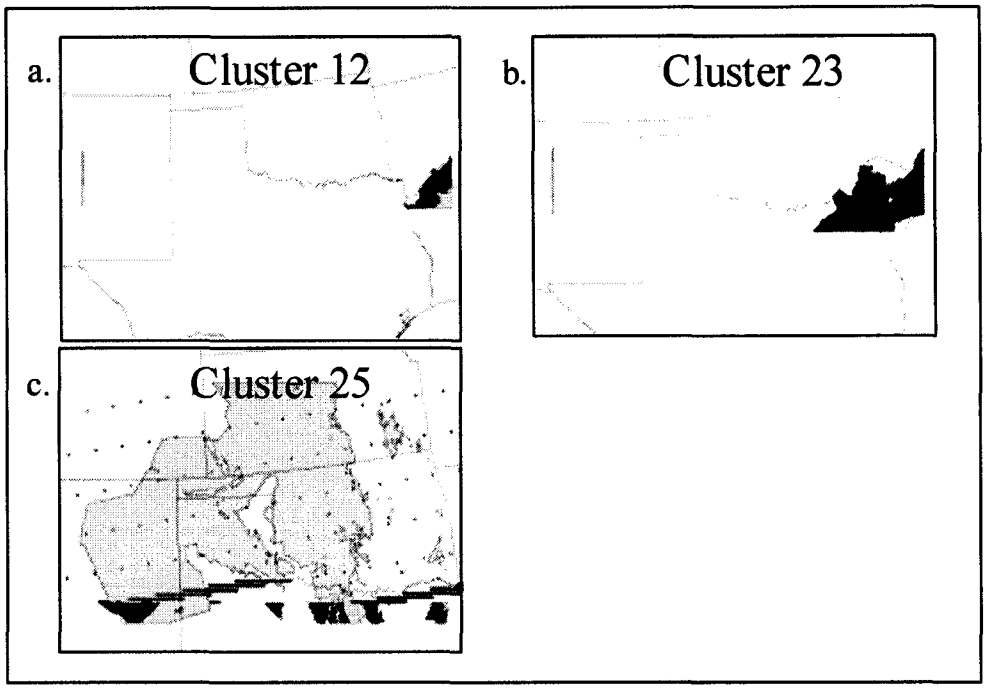


Figure 9. Examples from Clusters with Low Membership. The circles in clusters 12 and 13 show noise along the western edge of the DPA. Cluster 25, a single member cluster, exhibits noise through out the modeled area

indices can be used for query and analysis.

Acknowledgements

This research was funded by the National Imagery and Mapping Agency (NIMA) through the University Research Initiative Grant NMA202-97-1-1024. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the NIMA.

6. References

Abidi, B.R., Sari-Sarraf, H., Goddard, J.S. and Hunt, M.A., 1999, Facet Model and Mathematical Morphology for Surface Characterization, Photonics East, Proceeding No. 3837.

Adams, E.S., 2001, Approaches to the Study of Territory Size and Shape, *Annual Review of Ecological Systems*, 32, pp. 277-303.

Arkansas-Red Basin River Forecast Center, 2002, ABRFC Precipitation Products, http://www.srh.noaa.gov/abrfc/pcpnpage.html.

Cao, C., and Lam, N/S., 1997, Understanding the Scale and Resolution Effects in Remote Sensing and GIS, Scale in *Remote Sensing and GIS*. Edited by D.A. Quattrochi and M.F. Goodchild, (Boca Raton, FL: CRC Lewis), pp. 57-72.

Chang S. et al, 1998, A fully automated content-based video search engine supporting spatiotemporal queries, *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5), pp. 602-616.

Claramunt C and M Theriault, 1996, Toward Semantics for Modelling Spatio-Temporal Processes within GIS, *Advances in GIS Research II, Proceedings of the 7th International Symposium on Spatial Data Handling*, edited by M.J. Kraak and M. Molenaar, (London: Taylor and Francis), pp. 47-63.

Dieleman, J.A. and Mortensen, D.A., 1999. Characterizing the spatial pattern of Abutilon theophrasti seedling patches, *Weed Research*, 39, pp. 455-468.

Egenhofer M and K. Al-Taha, 1992, Reasoning about Gradual Changes of Topological Relationships, in *Theories and methods of Spatio-Temporal Reasoning in Geographic Space, International Conference GIS- From Space to Territory: Theories and methods of Spatio-Temporal Reasoning, Pisa Italy September 1992*, edited by A. Frank, I. Campari, U. Formentini, (Berlin: Springer Verlag), pp. 196-219.

Forte, P. and Greenhill, D., 1997, A Scale-Space Approach to Shape Similarity, *Lecture Notes in Computer Science*, 1252, pp. 333-336.

Galton, A., 1995, Towards a Qualitative Theory of Movement, in Spatial Information Theory: A Theoretical Basis for GIS, *Lecture Notes in Computer Science*, 988, pp. 377-396.

Galton, A., 2000, *Qualitative Spatial Change*, (Oxford: Oxford University Press).

Gardoll, S.J., Groves, D.I., Knox-Robinson, C.M., Yun, G.Y. and Elliott, N., 2000, Developing the tools for geological shape analysis, with regional to local-scale examples from the Kalgoorlie Terrane of Western Australia, *Australian Journal of Earth Sciences*, 47, pp. 943-953.

Giancoli, D.C., 1984, *General Physics*, (Englewood Cliffs, New Jersey: Prentice-Hall).

Gunsel, B. and Tekalp, A. M.,1998, Shape similarity matching for query-by-example, *Pattern Recognition*, 31(7), pp. 931-944.

Gupta, A. and Santini, S., 2000, Toward Feature Algebras in visual Databases: The Case for a Histogram Algebra, *VDB-5, International Conference on Visual Databases, Tokyo,* http://www.sdsc.edu/~gupta/publications/halgebra.pdf.

Hagen, M., Bartenschlager, B., and Finke,U., 1999, Motion characteristics of thunderstorms in southern Germany, *Meteorological Applications*, 6, pp. 227-239.

Hamazaki, T., 1996, Effects of patch shape on the number of organisms, *Landscape Ecology*, 11, pp. 299-306.

Hartigan, J., 1975, *Clustering Algorithms*, (New York: Wiley).

Hornsby, K. and Egenhofer, M., 2000, Identity-based change: a foundation for spatio-temporal knowledge representation, *International Journal of Geographical Information Science*, 14(3), pp. 207-224.

Hornsby K. and Egenhofer, M., 1997, Qualitative Representation of Change, *Lecture Notes in Computer Science*, 1329, pp. 15-33.

Houze, R., Smull, B., and Dodge, P., 1990, Mesoscale Organization of Springtime Rainstorms in Oklahoma, *Monthly Weather Review*, 118, pp. 613-654.

Kauppinen, H., Seppanen, T. and Pietikainen, M., 1995, An Experimental Comparison of Autoregressive and Fourier Based Descriptors in 2-D Shape Classification, *IEEE Transactions on pattern Analysis and Machine Intelligence*, 17(2), pp 201-207.

Kriegel, H.P. and Seidl, T.,1998, Approximation-Based Similarity Search for 3-D Surface Segments , *GeoInformatica* ,2(2), pp.113-147.

Li, X., 1996, A New Method to Improve Classification Accuracy with Shape Information, *International Journal of Remote Sensing*, 17(8), pp. 1473-1482.

Longley, P., Batty, M., Shepherd, J., 1991, The size, shape and dimension of urban settlements, *Transactions, Institute of British Geographers*, 16(1), pp. 75-94.

Longley, P., Batty, M., Shepherd, J., 1992, Do green belts change the shape of urban areas? An analysis of the settlement pattern of South East England, *Regional Studies*, 26(4), pp. 437-452.

McGrew, J. C. and Monroe, C.B., 2000. *An Introduction to Statistical Problem Solving in Geography, 2nd edition.* (Boston: McGraw-Hill).

Medda, Nijkamp and Riveld 1997, Recognition and Classification of Urban Shapes, Tinbergen Institute, Rotterdam, Paper 97-061.

Mennis, J., Peuquet, D. and Qian, L., 2000, A Conceptual framework for incorporating cognitive principles into geographical database representation, *International Journal of Geographical Information Science*, 14(6), pp. 501-520.

Mesev, T. V., Batty, M. Longley, P. A. and Xie, Y., 1995, Morphology from imagery: detecting and measuring the density of urban land use, *Environment and Planning A*, 27, pp. 759–780.

Okabe, A. and Masuda, S., 1984, Qualitative Analysis of Two-Dimensional Urban Population Distribution in Japan, *Geographical Analysis*, 16, pp. 301-312.

Osada,R., Funkhouser, T., Chazells, B. and Dobkin, D., 2002, Shape distributions, *ACM Transactions on Graphics*, 21, pp. 807-832.

Ruggiero, A., Lawton, J., and Blackburn, T., 1998, The geographic ranges of mammalian species in South America: Spatial patterns in environmental resistance and anistrophy, Journal of Biogeography, 25, pp. 1093-1103.

Sadahiro, Y., 2001, Analysis of Surface Changes Using Primitive Events, *International Journal of Geographical Information Science*, 15, pp. 523-538.

Sako, Y. and Fujimura, K., 2000, Shape similarity using homotopic deformation, *Visual Computer*, 16, pp. 47-61.

Schmidt, J., Lawrence, B., Olsen, B., 2000, A Comparison of Operational Precipitation Processing Methodologies, *NOAA Technical Memorandum NWS SR-205*, http://www.srh.noaa.gov/abrfc/p1vol.html.

SPLUS Language Reference, 6.1.2 Release 1, Insightful Corporation.

Srinivasa, N. and Abuja, N., 1999, A topological and Temporal Correlator Network for spatiotemporal pattern leaning recognition and recall, *IEEE Transactions on Neural Networks*, 10(2), pp. 356-371.

Tobler W., and Moellering, H., 1972, Geographical Variances, *Geographical Analysis*, 4, pp. 34-50.

Wentz, E., 2000, A Shape Definition for Geographical Applications Based on Edge, Elongation, and Perforation, *Geographical Analysis*, 32(1), pp. 96-112.

Wilcox, D., Harwell, M., and Orth, R., 2000, Modeling Dynamic Polygon Objects in Space and Time: A New Graph-based Technique, *Cartography and Geographic Information Science*, 27, pp. 153-164.

Williams S.E. and Pearson R.G.,1997, Historical rainforest contractions, localised extinctions and patterns of vertebrate endemism in the rainforests of Australia's Wet Tropics, *Proceedings of the Royal Society B*, 264, pp. 709-716.

Wilks, D.S., 1995, *Statistical Methods in the Atmospheric Sciences*, (San Diego:Academic Press).

Woodcock, C.E., and Strahler, A.H., 1987, The factor of scale in remote sensing, *Remote Sensing of Environment*, 21, pp. 311-332.

Yuan, M., 2001, Representing Complex Geographic Phenomena with both Object and Field-like Properties, *Cartography and Geographic Information Science*, 28, pp. 83-96.

Yuan, M. and Perault, D. R., 1998. Measuring the Fractal Dimensions of a Temporal Forest Landscape. *Applied Geographic Studies*, 2(2): 131-144.

Chapter 4

Assessing Similarity of Geographic Processes and Events

Assessing Similarity of Geographic Processes and Events in Distributed Phenomena

John McIntosh and May Yuan
Department of Geography
University of Oklahoma

Abstract

The increased availability of spatiotemporal data collected from satellite imagery and other remote sensors provides opportunities for enhanced analysis of geographic phenomena. Much of the new data includes regular snapshots of the environment. Comparison of these snapshots can provide information about changes to the modeled phenomena. However, challenges to geographers' ability to effectively use such spatiotemporal data are limitations in conventional data GIS data models and analytical tools to handle massive multidimensional data. One of the fundamental tools necessary to meet such challenges is query support to retrieve and summarize data that correspond to events and processes with certain spatiotemporal characteristics. To this end, this paper proposes a method to assess similarity of geographic events and processes that generate phenomena with varying characteristics in space and time. Such spatiotemporal characteristics can form significant features identifiable in the phenomenon. The proposed method, Dynamic Time Warping, is based on temporal sequences of six indices that describe characteristics and behaviors of these feature in space and time to assess similarity of events and processes that drive the change of the phenomenon of interest.

Key Words: Spatiotemporal Query, Field-Object Representation, Dynamic Time Warping

1. Introduction

Most geographic information systems are based on a map paradigm where each layer represents the state of a geographic theme at a given time or temporal interval. While many geographic phenomena change continuously, data is often collected at discrete times and stored as snapshots. A comparison of pairs of snapshots representing subsequent time intervals can reveal spatial and attribute changes of the modeled phenomena and support queries about "what changed," "where the change occurred," and "when the change occurred." However, patterns of change over time that reflect the evolution of events or processes are often of interest. Even if snapshots are taken at an adequate temporal resolution to capture important changes in a phenomenon as an event unfolds, comparison of snapshot pairs only addresses change between two time instants and does not reflect the lifecycle of the event.

Many geographic phenomena are distributed in that their properties vary across an extended area. The properties of such distributed phenomena are typically modeled as continuous surfaces. In this paper we present a conceptual framework to assess similarity of events that drive changes in such phenomenon captured in regular snapshots. The proposed framework is intended to enhance support of spatiotemporal analysis in geographic information systems by supporting queries based on patterns of change that reflect "how" events unfold. Temporal sequences of indices characterizing change in zones of relatively high or low values (conceptual objects) within the surface of the distributed phenomena of interest are the basis for the similarity assessment.

In the next section, we provide a brief background on recent advances in temporal GIS in the treatment of events and processes. Section 3 introduces the conceptual basis for characterizing geographic events in surfaces based on the behavior of conceptual objects, a description of the proposed indices used for the characterization of the surface and similarity measure selected to implement the query framework. Section 4 describes the implementation of the framework using hourly rainfall as a case study. The final section identifies strengths and weaknesses of the proposed framework and discusses areas for future work.

2. Background

A basic requirement for queries about events is to track change that is caused by the events. Tracking change in space and time requires abilities to handle time in a GIS. This section briefly describes some approaches to incorporating time in GIS and relates these approaches to the proposed framework.

The main impetus driving the evolution of spatiotemporal models described in the is: (1) efficient storage of spatiotemporal data; (2) better support for queries based on change; and (3) improved support of semantic concepts. One of the most commonly used spatiotemporal models is the snapshot model (Peuquet 2001). The snapshot model is conceptually simple. The states of a theme are recorded at regular time intervals and each time step is represented by a new layer. Although the snapshot model allows information about the state of a geographic theme to be retrieved for various time periods, there are several disadvantages. The model only stores information about the

theme at specific times. While information on change can be derived by comparing consecutive snapshots, an exhaustive search is required to determine how much change occurred over a specified period of time. In addition, snapshot models often store redundant data when no change occurs. To reduce data redundancy, the amendment vector approach (Langran 1992) stores a base map and changes in the geometry of objects. However, search for change is restricted to location. Alternatively, the TEMPEST model (Peuquet and Wentz 1994) is organized around a timeline rather than location and thus increases the efficiency of data searches that that seek when change occurs. More recent approaches such as the TRIAD framework (Peuquet 1994, Peuquet and Qian 1996) and the Three Domain Representation (Yuan 1996) include semantic information and relationships in addition to spatial and temporal information. These approaches provide capability to store and retrieve information based on time, position, and semantic concepts.

While much of the focus has been on modeling objects, new approaches such as the Event-Based Spatiotemporal Data Model (Peuquet and Duan 1995) have been proposed to improve querying capabilities and data storage efficiency for raster representation. This model stores change on a cell-by-cell basis. Temporal raster models track changes in attribute values at individual cells or global changes over the entire modeled domain. However, changes driven by geographic events modeled as raster layers often involve more than a single cell but occupy less than the entire modeled domain. While changes in cells may be tracked, current GIS technology lacks the capability to link and summarize cells that collectively correspond to change occurring

in an event. For example, an air pollution plume may only involve a portion of the modeled area and characterizations based on the entire domain, or at the individual cell level, may not be useful in isolation.

One solution is to identify cells of interest as features and to explicitly model these features as data objects. Mennis et al (2000) used gridded satellite derived cloud to temperatures to identify mid latitude cyclonic storms in the Midwest United States. Yuan (2001) used radar derived gridded hourly rainfall accumulations to track springtime storm events in Oklahoma. In both cases, the events (i.e. storms) were modeled and tracked as objects, thus providing a basis for temporal queries and analysis. While these and other spatiotemporal models provide information about the state of the phenomena at given times and change from one snapshot to another, they do not explicitly support queries based on patterns of change that span multiple snapshots. This is a significant deficiency because these patterns of change can reflect the way that events or processes evolve and provide insights into the factors affecting the change.

3. Assessing Spatiotemporal Similarity

Our method is designed according to a dual representational approach and models phenomena as both objects and surfaces similar to Yuan (2001). Features identifiable in a phenomenon are modeled as database objects. While the phenomenon is represented as a field to model the spatial variation of the phenomenon's properties, multidimensional temporal sequences of indices representing the state and behavior of these objects at regular time intervals are stored as attributes of data objects representing

individual processes and events. Querying about patterns of change is accomplished by comparing the sequence of index values of a target to the stored sequence of index values for an event object in the database. A nearest neighbor approach is used to order the events in terms of similarity to the target and the most similar event, or most similar set of events, from this list is returned.

The proposed method is based on three foundations:

1. Important spatiotemporal features in distributed phenomena can be modeled as conceptual objects.
2. The behavior of conceptual objects characterizes the behavior of the modeled phenomena.
3. Temporal sequences of indices describing the state and behavior of these objects at a sufficient temporal resolution can characterize the evolution of the surface.

The remainder of this section discusses indices used to describe properties of conceptual objects that form the basis of the surface characterization and proposes a distance measure to assess the similarity of events that drive changes in the surface.

3.1 Characterizing Processes and Events in Surfaces Using Conceptual Objects

Indices describing the state and behavior of conceptual objects are used to characterize patterns of change associated with events acting on the surfaces, including indices for object and object relationships in both static and dynamic descriptions. The static descriptors describe the current state of the object, and object relationships provide a baseline for the dynamic descriptions. Assembly of static indices in temporal sequences will provide historical information that reflects how changes occurred in space and time. Table 1 summarizes the proposed indices.

Table 1 – The proposed indices to characterize objects and object relationships of features imbedded in a distributed phenomena.

Index Type		Index Name	Description
Static	Object	Elongation	A measure of the width relative to the length of the best ellipsoidal approximation of the object derived using the maximum and minimum moments of inertia.
		Orientation	Orientation of the major axis of the conceptual object relative to the direction of movement.
	Object Relationship	Distribution	A measure of clustering based on the average nearest neighbor distance method.
Dynamic	Object	Growth	Percent growth.
		Granularity of Change	The operational scale of change in the conceptual object's surface based on the local variance method.
	Object Relationship	Relative Movement	A measure of the variation in movement velocity among the conceptual objects over distance calculated as the average maximum shear between objects.

The index values are arranged in temporal sequences of varying lengths corresponding to discrete patterns of change observed in temporal series of raster snapshots.

3.2 Assessing Similarity of Events Using Temporal Sequences

The previous section describes indices used to characterize surfaces based on the state and behavior of imbedded conceptual objects measured at regular time intervals. A critical element of the proposed framework is to effectively use these sequences to support querying of events or processes that generated change spanning multiple time frames. Traditional databases support querying by retrieving records that match an

explicit set of conditions. The complexity of possible sequence values makes it difficult to formulate temporal queries based on Boolean logic. Instead, the framework utilizes a nearest neighbor, or similarity based approach. Similarity queries seek records that closely match a model provided in the query.

The similarity of objects can be based on a measure of the difference between the objects, based on correspondence between the objects, or a combination of the two. Assessing similarity using either approach is not straightforward, and numerous methods have been proposed. The remainder of this section discusses conceptual models used to assess similarity and proposes Dynamic Time Warping, a method developed in computer science for speech and pattern recognition, to assess similarity of events and processes based on resulting change in space and time.

3.2.1 Conceptual Models of Similarity

Much of the work relating to spatiotemporal similarity has focused on correctly identifying objects or activities, and consequently, tends to be application specific. In contrast, the proposed method is intended to work with spatiotemporal patterns possessing a range of characteristics where the ultimate goal is an intuitive measure of similarity rather than correctly categorizing spatiotemporal patterns that are defined a priori. The fundamental question of similarity is to find a common reference frame for measurement. Winter (2000) pointed out that because there are so many possible aspects of physical, linguistic, and semantic correspondence, the statement "A is similar to B" is meaningless unless the framework for the similarity assessment is stated.

Cognitive science identifies four major models to assess similarity: geometric, featural, alignment based and transformational. Geometric approaches base similarity on the measured differences between the objects. As the differences decrease, the assumed similarity between the objects increases. Geometric models are commonly used to analyze similarity (Goldstone 1999). They are based on representing entities as points in n dimensional space. Typically, the similarity of two entities is judged to be inversely related to the distance in the attribute space of interest.

Featural models of similarity are based on both elements or features of objects that are common and different. Tversky (1977) argued that similarity should be based on both the number of common features and the features that are not in common. Instead, he proposed a model, which defined similarity between objects as a weighted difference of the common and distinct features. Geometric and the featural models of similarity are good for comparing collections of indices, commonly used for shape and image comparisons. However, neither model is good for highly structured data. Goldstone (1999) suggested that for structured data, it is often better to evaluate features hierarchically. This provides a means of comparing corresponding elements.

In the alignment method, matching features are aligned based on the role that they play within their objects. For example, a house with a blue door and a truck with a blue hood both share the feature blue, but the common color does not increase the similarity much because the house's door does not correspond to the truck's hood. Similarity under the transformational approach is based on how many changes would be required to make the two objects equivalent (transformational distance). The

similarity of two objects is assumed to be inversely proportional to the number of operations required to transform one object so it is equivalent to the other. For example, XXXXO requires only one transformation to become XXXOO.

3.2.2 Similarity Assessment of Temporal Sequences

Geometric models are commonly used to assess similarity in geographic research. Euclidean Distance, one of the most widely used geometric models, is equivalent to the straight line distance of sequences mapped to points in n space where n is the length of the sequence. Euclidean Distance is defined as

$$d\ (x,y) = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2} \qquad (1)$$

where n is the number of dimensions and x and y are the sequences. Euclidean Distance treats each element or dimension as independent so each element of a sequence is compared only to the corresponding element in the other sequence (x_i is compared with y_i, x_{i+1} with y_{i+1}, and so on). This assumption of element independence is reasonable for many applications. For example if temperature, humidity, and pressure are stored in sequences and the Euclidean Distance between two sequences is calculated, the temperature would be compared with temperature, humidity with humidity, and pressure with pressure.

However, with temporal sequences, the assumption of independence may be too restrictive. Temporal elements are typically not independent of each other and Euclidean Distance may not be consistent with perceived similarity if the sequences are not perfectly aligned. A person may recognize that spatiotemporal sequences have

similar features based on relative position of the patterns of high and low values. However, if the patterns are not perfectly aligned in sequences, as is often the case with temporal sequences, high values may be compared with low values, and therefore similar patterns based on relative location will not be reflected in the distance value.

For example, the two sequences in Figure 1a appear more similar than the two sequences in Figure 1b. In Figure 1a, both sequences exhibit similar peaks and pits, although these features are misaligned along the time axis. In Figure 1b, the patterns are quite different, one sequence with a constant value and the other with peaks and pits. Using the Euclidean Distance measure, there is less distance between the sequences in Figure 1b than in Figure 1a. The perceived similarity in Figure 1a is based on the relative positions of features, but the perceived similarity is not reflected in the Euclidean Distance because the patterns are out of phase.

Another characteristic of Euclidean Distance and other geometric methods is that the sequences must be of the same dimension or length. Events and processes that act in the landscape may vary slightly in duration, yet possess common features that are perceived by humans as similar. Sequences representing processes or events of different lengths cannot be compared without adjustments to one or both of the sequences. The adjustments can include increasing the length of the shorter sequence or removing values from the longer sequence. However, these adjustments either add or remove information from the sequences, which can lead to erroneous results.

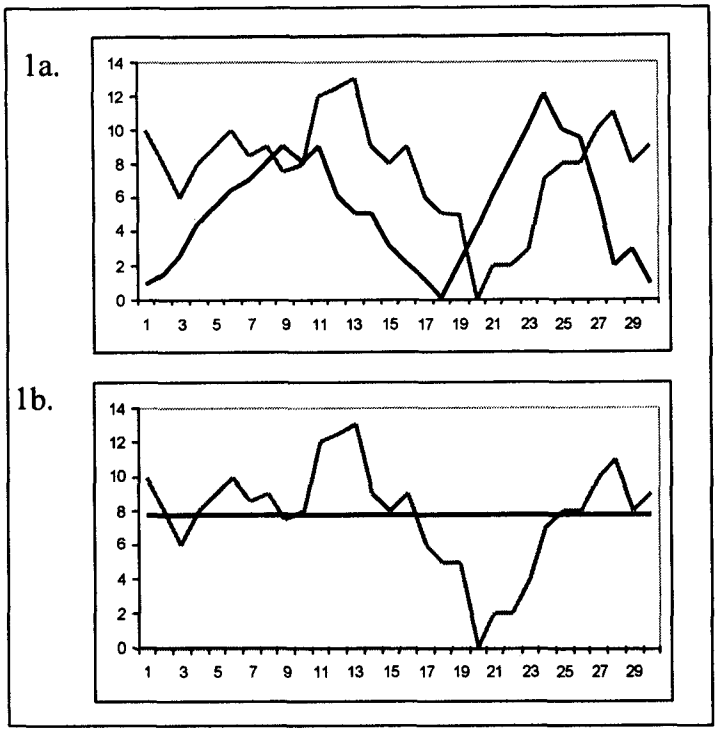


Figure 1. Significant features are not aligned on the temporal axis in 1a. Using Euclidean Distance, the sequences in 1b are more similar than the sequences in 1a.

Alternatives to Euclidean Distance have been developed, mostly in computer science and molecular biology to provide more flexibility for activities such as DNA analysis and voice recognition. For example, in genetics, sequence analysis of DNA attempts to match patterns of DNA in sequences that are not aligned. Similarly, people speak at different rates so speech recognition systems must be able to adjust phonemes for correct word matching. For these types of applications, geometric distance measures such as Euclidean Distance are not suitable. Techniques such as Quadratic Form Distance, String Edit methods, and Time Warping were developed to meet the challenge. Since then these methods have been applied to a diverse array of other

applications such as analysis of stocks, and computer vision, and activity analysis in the social sciences (Wilson 1999).

Quadratic Form Distance, in contrast to Euclidean Distance, does not assume that elements in a sequence are independent (Faloutsos et al 1994, Hafner et al 1995). The distance calculation includes the difference between all sequence element pairs with the greatest weight applied to the sequence element pairs that are most closely aligned in the sequences. Quadratic Form Distance is calculated as follows:

$$d_A(x,y) = \left[(x-y)\cdot A\cdot(x-y)^T\right]^{1/2} = \left[\sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j}(x_i - y_i)(x_j - y_j)\right]^{1/2} \qquad (2)$$

where x and y are the sequences, A is a weighting matrix that specifies the weighting of the distance between the sequence element pairs in the calculation and N is the length of the sequence. Because values that are most closely aligned in the sequence are weighted highest, slightly misaligned features are reflected in the distance value. Quadratic Form Distance, like Euclidean Distance also requires sequences of the same dimension.

For both Euclidean and Quadratic Form Distance, the magnitude of differences in the objects is used to quantify similarity. For sequences, similarity is assumed to be inversely related to the distance of the corresponding elements. Another approach is to focus on the parts of the sequences that are similar. String Edit methods take this approach. For example, the local alignment method proposed by Smith and Waterman (1981) determines similarity based on the length of the longest common subsequence in two sequences.

The similarity score is typically based on the length of the longest common subsequence normalized by the average length of the sequences being compared. The longest common subsequence need not be exactly matching but points are deducted from the similarity score for each element that does not match. The optimum solution is usually found using dynamic programming techniques. String Edit methods were originally developed for DNA analysis in molecular biology and consequently work with "alphabets" of nominal scale data. The concept of what constitutes a match must be adjusted to work with interval or ratio scale data. In one approach, elements are considered a match if the value of the second sequence falls within a specified range of the element value in the first sequence (Vlachos et al, 2002).

A significant disadvantage of the local alignment method with non-nominal scale data is that that the degree of difference of the nonaligned portions of the sequences is overlooked in the similarity measure. A portion of the sequence that does not match may have significant differences, but the score does not reflect how different the unmatched parts are (Figure 2).

3.2.3 Dynamic Time Warping

Dynamic Time Warping avoids problems associated with the methods described above. Dynamic Time Warping was introduced in the late 1970s (Sakoe and Chiba 1978, Rabiner et al 1978, Rabiner and Schmidt 1980). It is based on stretching portions of the sequences being compared along the temporal axis by repeating elements, so the dominant features, patterns of high and low values, are optimally aligned. The distance is the difference of the transformed sequences plus a cost associated with stretching the

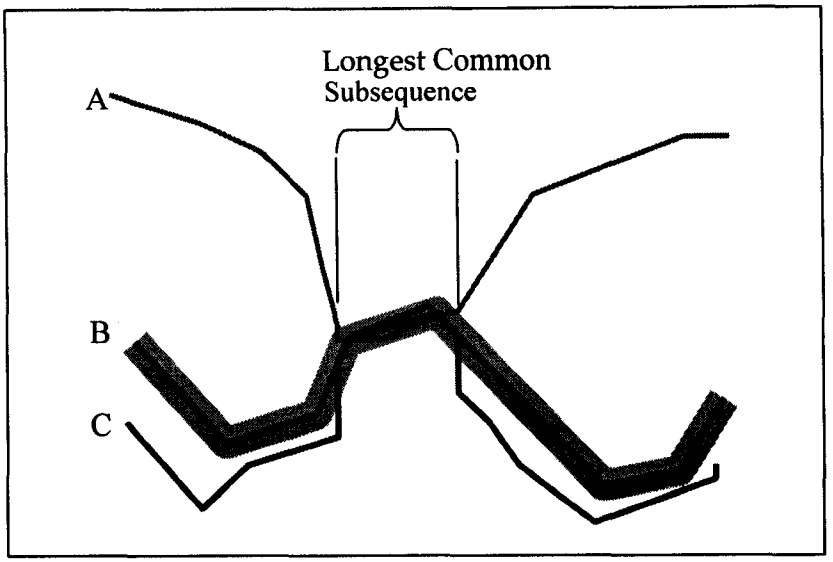


Figure 2. Sequences A and C share the same common subsequence with Sequence B. Based on the longest common subsequence method are equally similar even though the non matching portions of Sequence C are closer to Sequence B.

sequences to align features. Figure 3 shows the two sequences in Figure 1 and those sequences transformed using Dynamic Time Warping.

Dynamic time warping works with sequences of different lengths, and features need not be perfectly aligned to be reflected in the distance score. Like the local alignment method, the method is based on aligning sequences to maximize the similarity of features. Geometric distance is used to compare the transformed sequences utilizing all elements of the sequence. Since Dynamic Time Warping works with interval or ratio scale data, the magnitude of the differences is incorporated into the distance score.

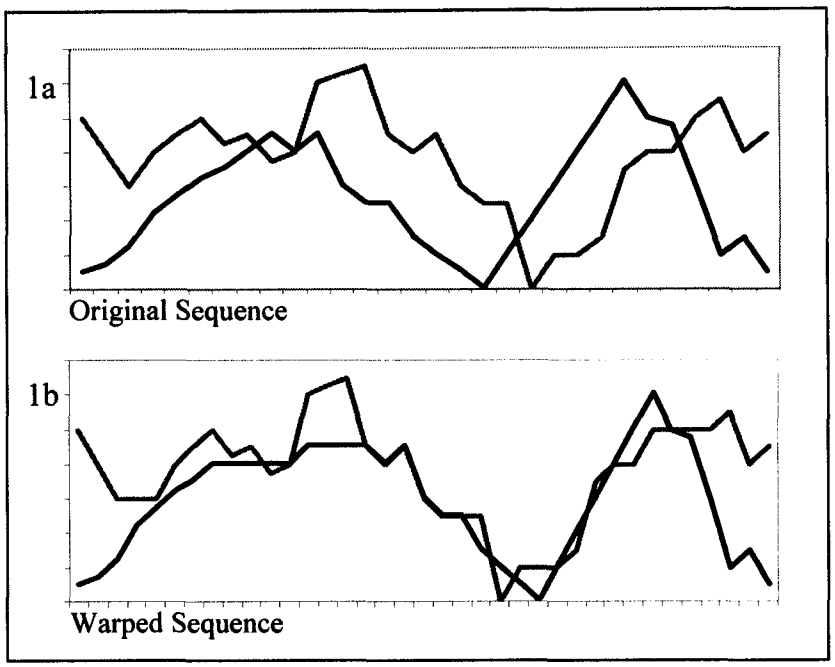


Figure 3. Warped Sequence

Time Warping is computationally expensive because all possible alignments to be evaluated to determine the optimal transformation of the original sequences. To minimize computation, Time Warping is typically performed using a dynamic programming algorithm. The dynamic programming algorithm is based on solving sub problems once and then reusing those answers to solve larger problems. Because Dynamic Time Warping involves testing all possible combinations of sequence elements, the distance between specific element pairs may be used for multiple alignments. The dynamic programming algorithm only requires the distance between possible element pairs to be calculated once thus avoiding re-computation of the answer for every possible alignment.

The basic algorithm for Dynamic Time Warping involves four procedures. Beginning with two sequences, A and B, with lengths *n* and *m* respectively, an empty *n* x *m* matrix (M) is created. Beginning at the lower left corner, the matrix is filled as follows:

1. Calculate the local cost for the lower left corner of the matrix ($M_{0,0}$). The local cost is the Euclidean Distance of A_0 and B_0. Move up first column sequentially adding the local cost plus the local cost of the cell below and a compression/dilation penalty.

2. Calculate the cumulative costs for the next column starting at the first row working up. The cost includes the local cost (Euclidean Distance of the corresponding sequence elements) plus the minimum of the following.

    a. The value of cell $M_{i-1,j-1}$
    b. The value of cell $M_{i-1,j}$ plus the compression/dilation penalty
    c. The value of cell $M_{i,j-1}$ plus the compression/dilation penalty

3. Repeat step 2 until the entire matrix is filled in.

4. The final distance is the value in the upper right corner of the matrix ($M_{n-1,m-1}$)

Figure 4 shows the completed matrix (M) for the sequences in Figure 3. The line through the matrix shows the corresponding elements of the sequences that are used to calculate the distance between the sequences. The position of the path indicates the corresponding elements of the transformed sequences. Note that every value in the original sequences is used; the warping merely stretches the sequences in order to find an optimal solution with the least distance. The time warp distance is the value in the upper right cell of the matrix, which is 88 in Figure 4. We standardize the time warp distance by dividing it by the average number of time steps in the sequences, which is 30 in Figure 4.

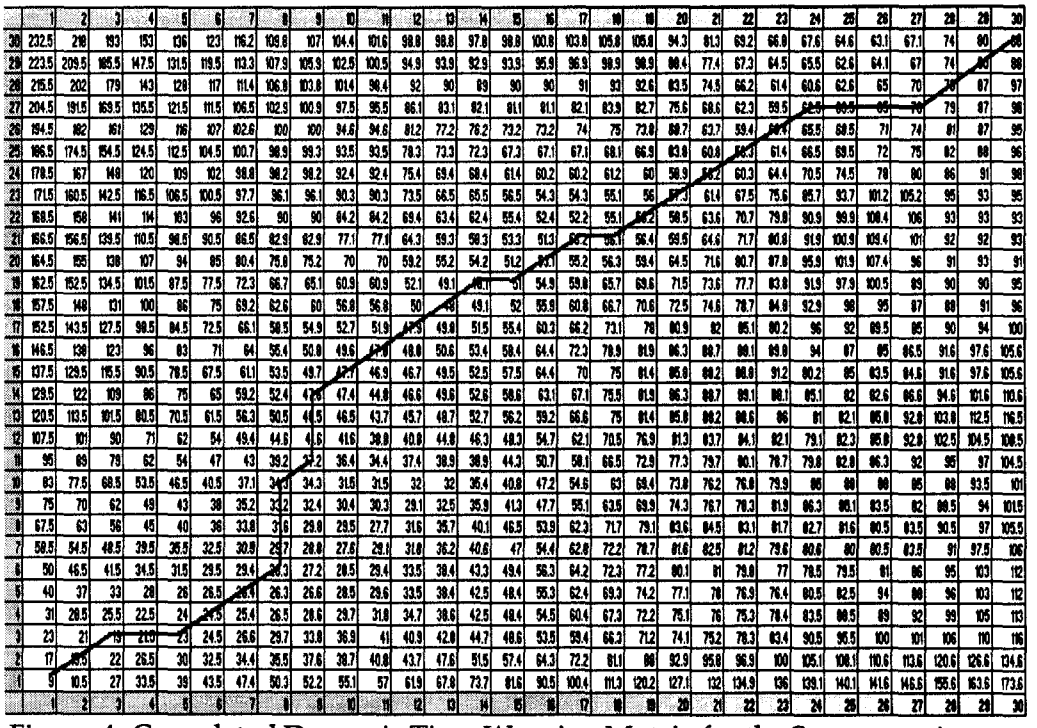


Figure 4. Completed Dynamic Time Warping Matrix for the Sequences in Figure 3.

4. Case Study

The proposed method was implemented using hourly rainfall in the Southern Plains. The Arkansas-Red River Basin Forecast Center (ABRFC) archives hourly 4 km digital precipitation arrays (DPAs) for Oklahoma and portions of surrounding states (Figure 5). The data is available from 1995 to the present. It is arranged by date and time with individual files containing raster layers that can be downloaded on demand. The progression of rainstorms (processes and events that drive change in precipitation) can be seen in the snapshots from hour to hour. Information regarding how storms evolve in the region is potentially of value for diverse uses such as climatology,

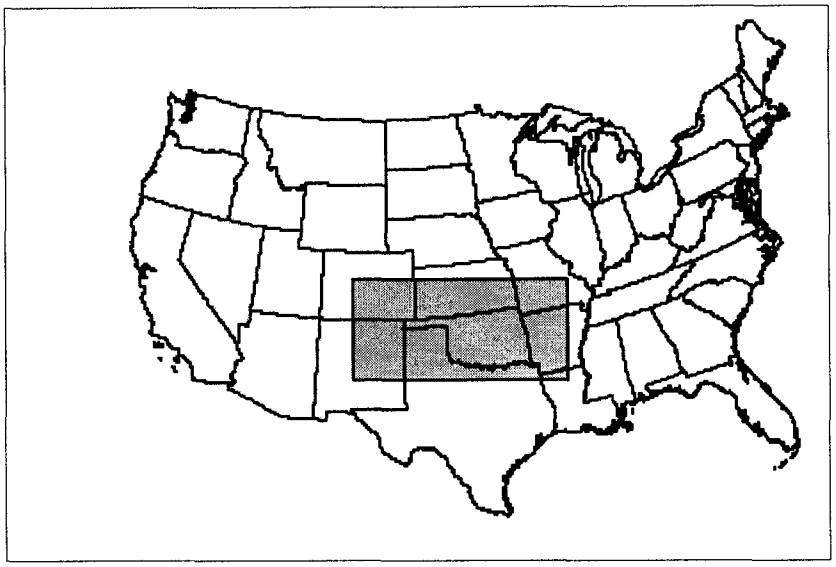


Figure 5. Arkansas-Red Basin River Forecast Center Digital Precipitation Array (DPA)

agriculture, and resource management. Over eighty-seven hundred raster layers are added to the archive each year. The large size of the data set makes it difficult to manually search for specific spatiotemporal patterns. Consequently, information regarding how processes and events evolve is contained in this data set is largely inaccessible. The proposed framework could be useful in cases like this by providing a means to query the data based on spatiotemporal patterns in the modeled surface.

The case study aims to test the performance of the proposed framework and applicability of the approach for querying based on events or processes that can be deduced from changes imbedded in the rainfall data. Specifically we want to address the following questions:

- Is the ordering of events or processes under the framework consistent with human perception?
- Is the distance measure robust with regards to perceptually similar events or processes of differing lengths?

- Do perceptually similar events and processes cluster?

The remainder of this section describes the methodology and results of the case study.

4.1 Methods

The framework was tested using rainfall data from March 15, 2000 to June 15, 2000. We developed several Java scripts to download and convert the DPAs to a grid format. Indices were calculated using Avenue™, the ArcView® GIS 3.2 software (Environmental Systems Research Institute Inc., Redlands, California) scripting language. Areas of rainfall form "zones" in the DPAs. Three thresholds were used to delineate rainfall zones (> 0 mm/hour, > 20 mm/hour and > 40 mm/hour). Houze et al (1990) classified springtime rainstorms based on arrangement and shape characteristics of rainfall zones. The thresholds selected for the case study are intended to capture these features in DPAs.

We organize the data into zones, sequences, processes and events (Figure 6). Rainfall zones are linked to form sequences based on overlaps in consecutive rainfall snapshots (DPAs). Rainfall processes link sequences related by merges or divisions and their predecessors and descendents. Rainfall events represent consecutive periods with rainfall somewhere in the modeled domain. Some events may consist of multiple processes. The processes correspond to discrete storms, which may be of as much interest as the overall patterns in the modeled domain. For example, Houze et al (1990), evaluated storms, which correspond to process objects in our representation. Index values described in section 3.1 were calculated for both events and processes.

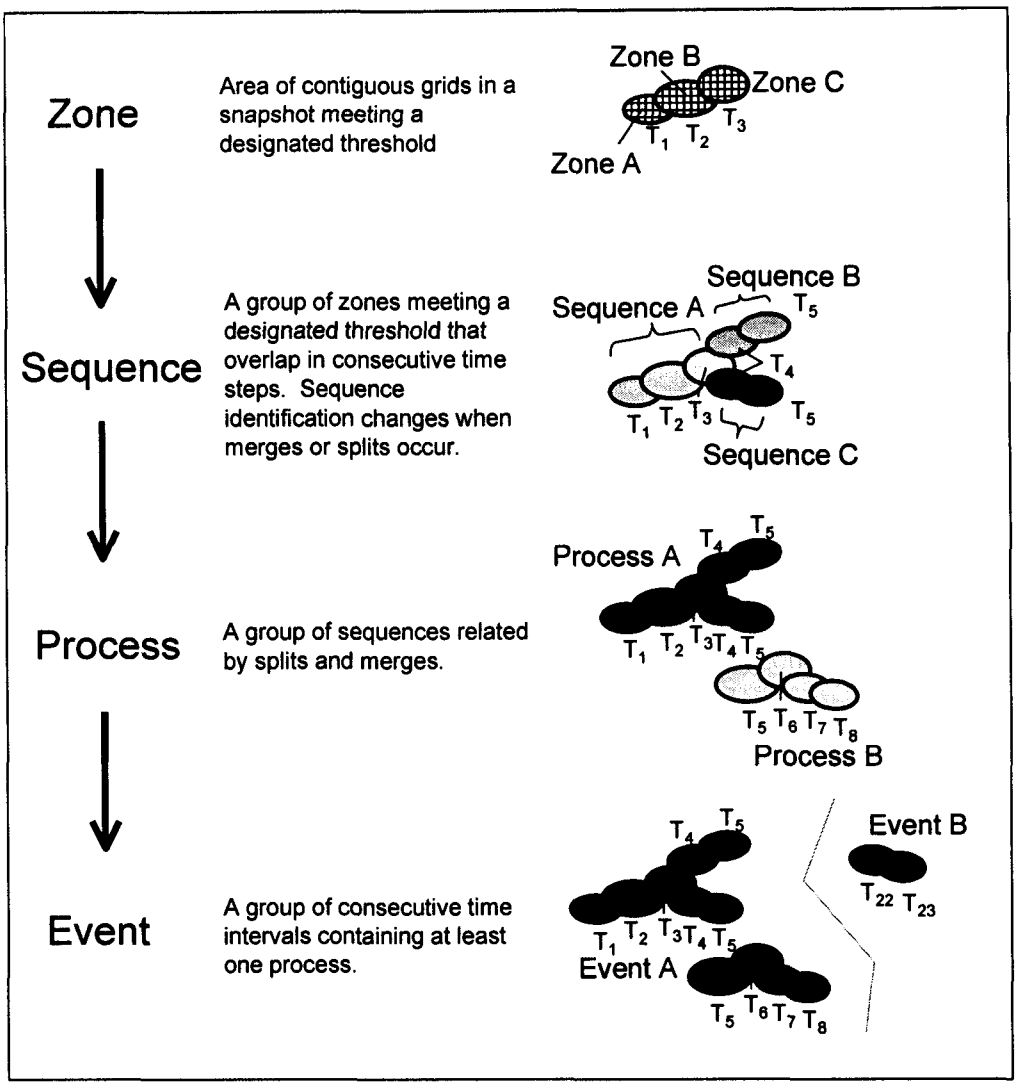


Figure 6. Representational Framework.

Similarity queries require a target, or model, in the same form as the database objects. The temporal sequences include six indices per time element and events and processes of interest may span many hours. In order to facilitate querying, we developed a "query by example" interface to create target sequences for use in the queries. The user sketches a series of time stamped rainfall areas on the screen at the desired rainfall threshold (Figure 7). Scripts calculate the index values and store them in a table. The query interface allows the user to store a query and select a stored query or

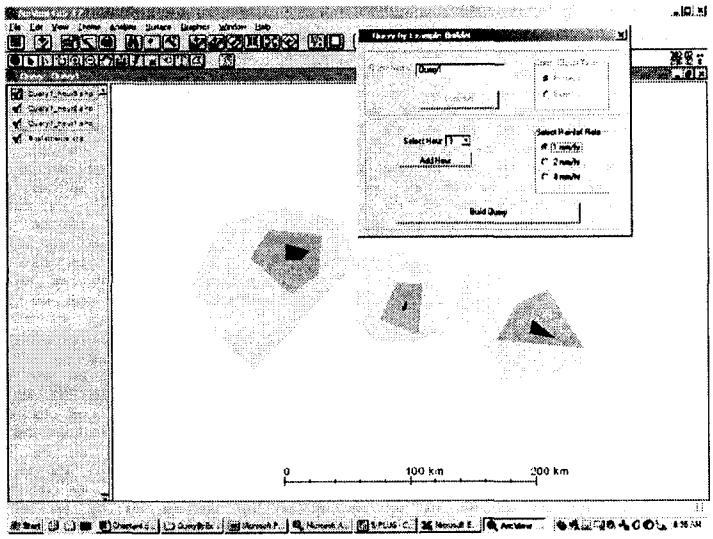


Figure 7. Query by Example Interface.

database object. Response to the query is a list of storms ordered from most similar to least similar. Since the relative importance of various characteristics change depending on use, the query interface allows the user to select the indices used in the similarity assessment and select the relative weight of these indices.

4.2 Evaluation

Our evaluation seeks to (1) find if the query results are consistent with human perception, (2) find if perceptually similar processes and events are clustered accordingly, and (3) find if the framework performs satisfactorily with sequences of different lengths.

It is relatively straightforward to assess performance if the goal is to correctly identify or classify objects into predefined categories because errors of omission and commission can be easily measured. In contrast, the proposed method focuses on comparing database objects to a target without fixed categories defined a priori. There is

no predefined "correct" answer. Due to the complexity of the event objects, we use a small sample of 20 processes with durations of 3 to 5 hours to illustrate how the method orders typical processes. The processes in the test set are based on zones defined by rainfall greater than 0 mm/hour. Figure 8 shows the object representation for each of the processes throughout its lifecycle.

Each of the processes in this test set was compared to the other 19 processes using Dynamic Time Warping to create a distance matrix (Figure 9). Lower scores on the distance matrix correspond to less distance, or a greater similarity. The Dynamic Time Warping scores are normalized by the length of the sequences which allows the distances between processes to be compared, regardless of length if the same set of indices are used for the analysis. Many of the processes in the test set appear very similar. For example, a number of these processes are elongated with movement roughly parallel to the orientation of the major axis. Process 1 exhibits this general pattern. The closest matches to process 1, processes 2, 7, and 14 are also distinguished by these characteristics. Consistently, the time warped distances etween these processes and Process 1 are very similar In contrast, the distance between Process l and the least similar matches, processes 4, 8, 11, and 19 is are at least twice as high as those in the previous group.

The relative distances to Process 1 suggest that the combination of indices and distance measures are able to order examples by degrees of similarity. It is also important for the framework to provide intuitive scores and ordering for processes and events that are not perceptually very similar. The distances between Process 8 and all of

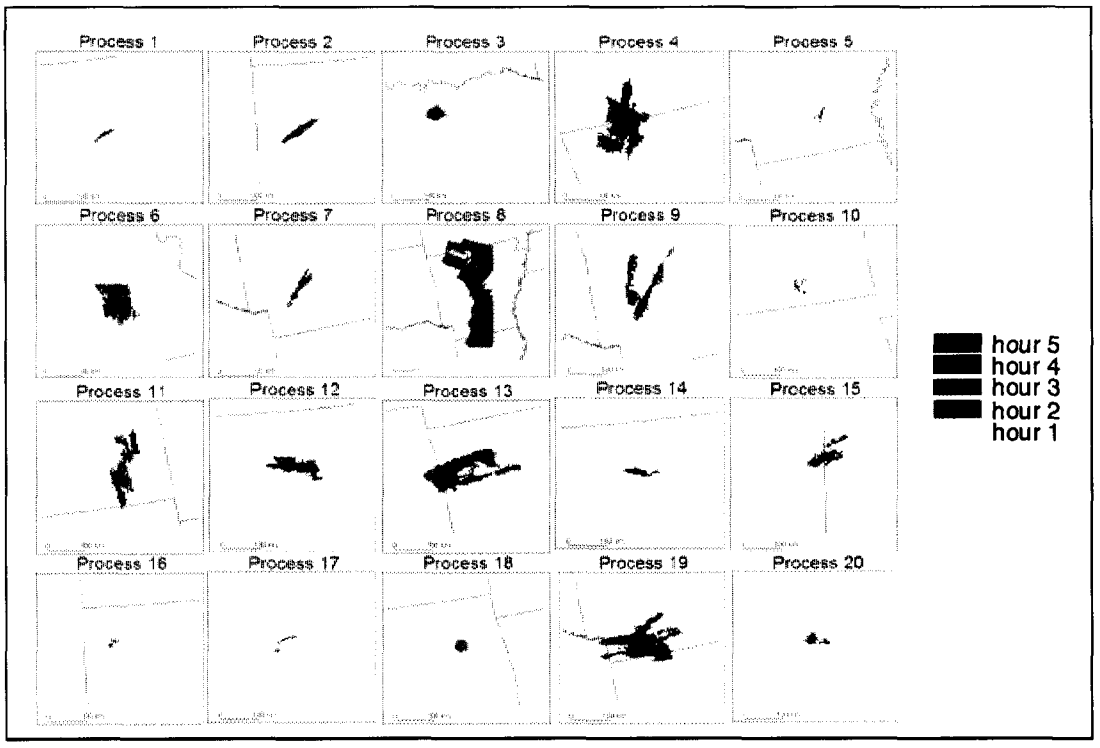


Figure 8. Test set of processes with durations between 3 and 5 hours.

the other processes appear high suggesting that there is no "good" match. In contrast to Process 1, which had a best match with a distance of 0.63, the best match for Process 8 is Process 11 with a distance of 2.35. While the distances are high relative to the best matches for process 1, examination of the charts in figure 8 suggests that the distance measure is able to order examples by degrees of freedom, even if there are no close matches.

As mentioned in the previous section, processes that share a similar pattern of features may have slightly different durations. The Dynamic Time Warping algorithm balances costs of stretching the sequences with the benefits of aligning the features. In our implementation we use a unit cost (=1) for each time that the sequences are stretched along the time axis. The indices are converted to standard anomalies to avoid problems

Process Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0																			
2	0.77	0																		
3	1.59	1.62	0																	
4	2.21	2.50	2.47	0																
5	1.91	1.57	1.59	2.06	0															
6	1.45	1.67	1.58	1.71	1.90	0														
7	0.63	0.69	1.41	2.62	1.54	1.71	0													
8	3.38	3.61	3.58	2.68	3.71	2.54	3.78	0												
9	0.84	1.16	2.05	2.06	1.69	1.54	1.33	3.18	0											
10	1.24	1.03	1.51	2.40	1.56	1.46	1.23	3.66	1.55	0										
11	2.31	2.63	2.81	2.29	2.40	1.99	2.68	2.35	1.94	2.54	0									
12	1.14	1.20	1.55	2.10	1.61	1.90	0.95	3.74	1.44	1.34	2.49	0								
13	1.75	1.62	2.09	1.66	1.93	1.56	1.88	2.71	1.55	2.26	1.80	1.81	0							
14	0.74	0.86	1.84	2.52	1.80	1.77	0.90	3.92	1.35	1.11	2.74	0.93	1.95	0						
15	0.82	0.96	1.74	2.23	1.63	2.02	0.89	3.90	1.31	1.31	2.68	0.83	2.05	1.15	0					
16	1.43	1.69	1.28	2.11	2.03	1.17	1.56	2.94	1.52	1.35	2.04	1.53	1.79	1.30	1.90	0				
17	1.01	0.57	1.59	2.41	1.53	1.73	0.98	3.28	1.23	1.19	2.54	1.25	1.78	1.16	1.16	1.85	0			
18	1.70	1.38	1.08	2.59	1.68	1.51	1.31	3.33	1.95	1.00	2.60	1.27	2.20	1.21	1.60	0.63	1.44	0		
19	2.01	2.20	2.39	1.41	1.87	1.64	2.39	3.00	1.22	2.21	2.09	2.39	1.32	2.40	2.45	1.90	2.22	2.46	0	
20	1.57	1.92	1.64	1.87	1.64	1.65	1.81	2.99	1.61	1.70	2.03	1.19	1.85	1.78	1.60	0.94	1.80	1.29	2.13	0

Figure 9. Dynamic Time Warping Distance Matrix of Test Set

associated with units of measurement. Consequently a unitary cost is fairly substantial unless there is a misalignment of significant features.

Another goal of the case study is to find if the method performs satisfactorily with sequences of different lengths. The results should reflect similarity of processes even if the lengths are not equal. Even with processes of short duration used in our test set, slight variation in length does not disproportionately increase the distance scores. For example Process 1 has a duration of four hours, yet the four closest matches are of different lengths.

Geographic analysis and interpretation often involves developing and assigning categories. We performed a hierarchical cluster analysis using the complete linkage method with the distance matrix of the test set (Figure 9). The dendogram in Figure 10 shows the results. If a height of two is selected, six clusters emerge.

As Table 4 indicates, three of the clusters have a single member. The remaining clusters show reasonable groupings. For example, Cluster 1 contains processes that are

Table 4. Test Set Clusters

Cluster ID	Process Number	Cluster ID	Process Number
1	1	2	3
	2		6
	5		16
	7		18
	10	3	9
	12		13
	14		19
	15	4	4
	17	5	8
	20	6	11

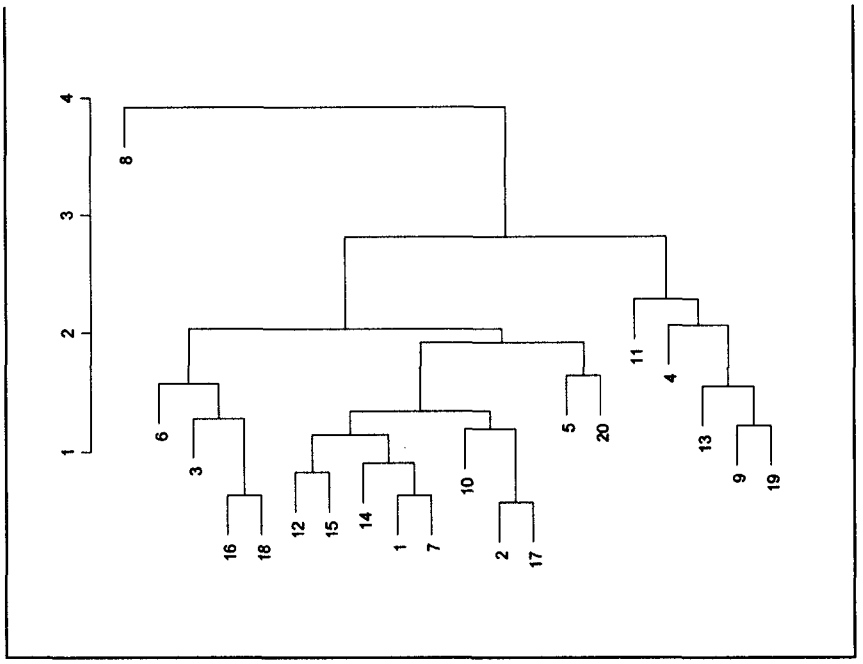


Figure 10. Dendogram of test set clusters.

elongated and move roughly parallel to the axis or orientation and Cluster 2 includes processes that appear round and show little movement. Because of the small test set, the clusters tend to be defined by just a few factors such as elongation and orientation, with the exception of process 8 which has a significantly higher spatial scale of change.

5. Conclusions

In this paper we proposed a method to assess similarity of processes and events that drive changes in phenomena with properties varying across space. The proposed method utilizes temporal sequences of index values characterizing the state and behavior of identifiable features in the phenomenon of interest as the basis for the similarity assessment. Dynamic Time Warping is proposed to assess similarity because of its ability to work with sequences of varying lengths and ability to account for features that are not perfectly aligned over time.

Because the proposed similarity assessment is intended to be used to support spatiotemporal queries the measure must provide satisfactory results so the users can retrieve desired database objects that appear similar to a target object. Three criteria were proposed to evaluate the performance of the framework: (1) to find if the similarity assessments are reasonable compared with visual inspections, (2) to find if similar processes judged by the method are consistent with results from clustering analysis, and (3) to find if the method is capable of assessing similarity among sequences of different lengths The results of the case study suggest that the proposed framework performs well on the specified criteria. The relative ordering of the test processes corresponded well with visual inspection. The method was not overly sensitive to sequences of different lengths and in a number of cases the best matches were of different lengths.

Three major clusters were distinguished based primarily on elongation and orientation. Other indices such as growth, granularity of change, distribution, and relative movement appear minor in the similarity and clustering results. This is probably due to the small number of samples in the test set. When the sample size becomes larger, more diverse processes can be discerned, and more indices may play determining roles in judging similarity using dynamic time warping and clustering methods.

There are several areas for future work. While the framework is intended to be general, it has not been tested with other phenomena. The combination of index values and similarity assessment techniques may be domain dependent. The method should be tested with other phenomena. Calculating distance using Time Warping is computationally intensive, even using the dynamic programming algorithm. The computation time is related to the length of the sequences, $O(n*m)$, so with large sequences, the computation time will be high. There are methods to optimize the calculation such as stopping the calculation if a specific distance threshold has been exceeded or by limiting the calculations to the most likely alignments of the sequences. Future work should incorporate these or other optimization techniques or develop heuristic algorithms to derive "good-enough" solutions.

Acknowledgements

This research was funded by the National Imagery and Mapping Agency (NIMA) through the University Research Initiative Grant NMA202-97-1-1024. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the NIMA.

6. References

Arkansas-Red Basin River Forecast Center, 2002, ABRFC Precipitation Products, http://www.srh.noaa.gov/abrfc/pcpnpage.html

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D. and Equitz, W., 1994, Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems*, 3, pp. 231-262.

Goldstone R., 1999, Similarity, in *MIT Encyclopedia of the Cognitive Sciences*, R. Wilson and F. Keil editors, (Cambridge, MA: MIT Press), pp. 757-759.

Hafner J., Sawhney H. S., Equitz W., Flickner M. and Niblack W. 1995, Efficient Color Histogram Indexing for Quadratic Form Distance Functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17 (7), pp. 729-736.

Houze, R., Smull, B., and Dodge, P., 1990, Mesoscale Organization of Springtime Rainstorms in Oklahoma, *Monthly Weather Review*, 118, pp. 613-654.

Langran, G., 1992, *Time in Geographic Information Systems*, (Washington DC.: Taylor & Francis).

Mennis, J., Peuquet, D., and Qian, L., 2000, A Conceptual framework for incorporating cognitive principles into geographical database representation, *International Journal of Geographical Information Science*, 14(6), pp. 501-520.

Peuquet D., 1994, Its about time: A conceptual framework for the representation of temporal dynamics in geographic information systems, *Annals of the Association of American Geographers*, 84(3), pp 441-461

Peuquet, D., 2001, Making Space for Time: Issues in Space-Time Data Representation, *Geoinformatica*, 5(1), pp. 11-32.

Peuquet, D. and Duan, N., 1995, An event-based spatiotemporal model (ESTDM) for temporal analysis of geographic data, *International Journal of Geographical Information Systems*, 9(1), pp. 7-24.

Peuquet, D., and Qian, L., 1996, An Integrated Database Design for Temporal GIS, in *Advances in GIS Research II, Proceedings of the 7th International Symposium on Spatial Data*

Handling , edited by M.J. Kraak and M. Molenaar, (London: Taylor and Francis, London), pp. 21-31.

Peuquet, D. and Wentz, E., 1994, An Approach for Time-Based Analysis of Spatiotemporal Data, *Advances in GIS research; proceedings of the Sixth International Symposium on Spatial Data Handling, Edinburgh Scotland,*(London: Taylor and Francis), pp. 489-504.

Rabiner, L. R., Rosenberg, A. E. and Levinson, S. E., 1978, Considerations In Dynamic Time Warping Algorithms For Discrete Word Recognition, *IEEE Transactions On Acoustics, Speech and Signal Processing*, 26, pp. 575-582.

Rabiner, L. R. and Schmidt, C. E., 1980, Application Of Dynamic Time Warping To Connected Digit Recognition, *IEEE Transactions on Acoustics Speech and Signal Processing*, 28, pp. 377-388.

Sakoe, H. and Chiba, S., 1978, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26, pp. 43-49.

Smith T. and Waterman M., 1981, "Comparative biosequence metrics", *Journal of Molecular Evolution*, 18, pp. 38-46.

Tversky, A., 1977, Features of similarity, *Psychological Review*, 84, pp. 327-352.

Vlachos, M ., Kollios, G., and Gunopulos, D., 2002, Discovering Similar Multidimensional Trajectories, *International Conference on Data Engineering, February 26 – March 1, 2002, San Jose CA,* pp. 673-684.

Winter, S., 2000, Location similarity of regions, *ISPRS Journal of Photogrammetry & Remote Sensing* 55, pp. 189-200.

Yuan, M., 1996, Modeling Semantic, temporal, and spatial information in geographic information systems, in *Geographic Information Research: Bringing the Atlantic* edited by M Craglia and H Couclelis, (London, Taylor and Francis), pp. 334-347.

Yuan, M., 2001, Representing Complex Geographic Phenomena with both Object and Field-like Properties, *Cartography and Geographic Information Science*, 28, pp. 83-96.

Chapter 5

Summary and Conclusions

1. Introduction

Geographical understanding of processes that shape the landscape is enhanced by information relating to how change occurs. In the past geographers often worked in data poor environments, but increased use of remote sensing and other innovations of geographic data acquisition are beginning to provide data rich environments for investigation of many phenomena. Sources such as earth observation satellites not only provide means of collecting new digital data, but also provide regularly scheduled measurements that document change in space and time.

Much of the recently available data is for geographic phenomena, with properties distributed across wide areas that are typically modeled with raster snapshots. The increasing availability of such spatiotemporal data holds great promise for improved understanding of geographic processes. However, a significant impediment to geographers' ability to use this data is the limited support to search through large spatiotemporal datasets based on patterns of change that reflect the effects of events and processes on the modeled phenomena. This dissertation presented a representational and query framework to overcome such limitations.

The dissertation research included three self contained, but related studies. The first study proposed an extension to the dual field/object data model that defines objects based on multiple thresholds and maintains temporal and topological relations between these objects for enhanced spatiotemporal analysis. The second study developed an approach to characterize change in the phenomena between subsequent snapshots of

using indices to describe the state and change in the modeled objects. The third study proposed a method to assess the similarity coherent patterns of change in the modeled phenomenon corresponding to processes and events utilizing temporal sequences of the indices developed in the second study.

The representational framework and proposed methods to characterize and assess similarity were evaluated with hourly digital precipitation arrays (DPAs) in the Arkansas-Red River Basin. However, the framework is intended to be generally applicable for dynamic phenomena that possess both field and object-like properties and in which change occurs in distinct processes and events. For example, other phenomena such as wildfires, disease outbreaks or air pollution could be modeled in a GIS using this approach. The remainder of this chapter summarizes the results of the three studies and proposes future work.

2. Summary of Results

The framework proposed in Chapter 2, "A framework to enhance semantic flexibility for analysis of distributed phenomena", defines object-like features embedded in distributed geographic phenomena. The framework categorizes objects including zones, sequences, processes and events based on boundary thresholds and spatial and temporal continuity. It stores a complex set of relationships between these objects over time to support enhanced query and analysis capabilities that would be impractical to derive on demand. The proposed framework was implemented to:

1. Verify if the prototype provides enhanced querying and analysis capabilities to handle both fields and object-like features embedded in rain fields;

2. Determine if the prototype is capable of handling multiple boundaries for object-like features and maintaining their topological and temporal relationships;

3. Investigate scaling issues of the proposed framework regarding storage space and processing requirements.

Implementation of the case study demonstrated that the prototype was able to handle multiple boundaries for the objects and maintain topological and temporal relationships. An analysis of storage requirements showed that the incremental costs to implement the framework is relatively low, about 3% of the total storage requirements of the raster part of the representation.

In Chapter three, " Object-Based Characterization of Surface Changes to Enhance Querying and Analysis in Geographic Information Systems," a suite of six indices was proposed to characterize spatiotemporal patterns in the dual field/object representational scheme proposed in the first study (Table 1). The method assumes that change in surfaces may be characterized by spatiotemporal patterns of embedded conceptual objects. Characteristics of both objects and object relationships are included in the indices. The proposed approach utilizes both static and dynamic indices. Static indices describe the current state of the object or object relationships and provide a baseline for the dynamic descriptions.

Table 1. Proposed Indices

	Objects	Object Relationships
Static	Elongation and orientation	Distribution
Dynamic	Growth and Granularity of Change	Relative Movement

Two criteria were proposed to evaluate the indices:

1. The indices should be efficient with little overlap of information characterized in each of the indices
2. Perceptually similar patterns of change should cluster.

A correlation analysis was performed to test the degree of information overlap between the proposed indices. The results of a correlation analysis showed that there is very little overlap between the indices at both the event and process level. A k-means cluster analysis was performed to see if perceptually similar spatiotemporal patterns of change cluster. The cluster analysis showed that the suite of indices effectively groups cases that exhibit similar spatiotemporal characteristics.

In Chapter 4, "Assessing Similarity of Geographic Processes and Events ," a method was proposed to assess the similarity of patterns of change spanning multiple snapshots in the raster surfaces associated with the lifecycles of processes and events using temporal sequences of the indices proposed in Chapter 3. Dynamic Time Warping was proposed to assess the similarity of the sequences. Three criteria were used to evaluate the performance of the proposed approach.

1. The similarity assessment should be consistent with human perception,
2. Perceptually similar processes and events should cluster accordingly,
3. The framework should perform satisfactorily with sequences of different lengths.

The proposed approach was tested using a subset of the rainfall data. The results of the case study suggest that the proposed framework performs well based on the specified criteria. The relative ordering of the test set corresponded well with a visual

inspection. A hierarchical cluster analysis of a small subset was performed. The cluster analysis showed that the similarity method clusters perceptually similar events together. Inspection of the closest matches showed that Dynamic Time Warping is not overly sensitive to sequences of different lengths.

3. Discussion and Future Research

The representational framework and approach to assess similarity proposed in this dissertation are intended to improve exploration and analysis of dynamic geographic phenomena with properties varying across space. Chapters two, three, and four show that the proposed approach accomplishes this goal with the rainfall data. The approach assumes that change within continuous fields occurs in discrete processes and events such as thunderstorms, fires, or floods. Because the representational framework models change in terms of distinct events and processes the proposed approach is not suitable for phenomena with change occurring continuously without logical divisions. It is also assumed that the distribution and behavior of zones of relatively high values within surfaces provide an adequate basis for queries based on how events and processes evolve. The query framework can only be implemented if the spatial and temporal granularity is adequate to capture relevant distribution and behavior of the process of interest. With too coarse of spatial or temporal granularity the behavior of processes may be inadequately described. Such inadequacy will invalidate the similarity assessment on which the queries are based.

While the evaluations in chapters three and four suggest that the suite of indices and dynamic time warping provide a basis for discriminating between perceptually similar spatiotemporal patterns and a reasonable basis for assessing similarity of processes and events, the accuracy was not tested in this research. The fundamental difficulty is that similarity is subjective and thus there are no universally recognized values for similarity to use as a basis to test for accuracy. However, there often is agreement among people about the relative ordering and general degree of similarity. While not in the scope of this research, the performance of the proposed framework could be evaluated based on how the magnitude of similarity and the relative ordering under the framework relates to the perceptions of human subjects.

The framework focuses on representation and querying of a single geographic theme. Most events involve more than a single data layer and event descriptions and comparisons would be more comprehensive to involve other correlated themes such as precipitation and temperature. Future work should investigate extending the framework to include multiple themes.

Most GIS data models represent fields in two dimensions with the third dimension representing the attribute values, but this is not adequate for some types of phenomena. For example, air pollution fields vary vertically as well has horizontally, and events modeled using two spatial dimensions will not capture this vertical variation. Future work should investigate extending the proposed framework to three dimensions. Finally, the framework and proposed approach to assess similarity have been presented as a means of exploring historical data. The approaches may have some

utility for prediction of future behavior by identifying analogs from the historical database.

Despite these limitations, the currently proposed framework should provide benefit to analysis of spatiotemporal data sets representing a broad range of geographic phenomena.