

Research Article

A framework to enhance semantic flexibility for analysis of distributed phenomena

J. MCINTOSH*[†] and M. YUAN[‡]

[†]Metcalf & Eddy, 5075 South Bradley, Suite 203, Santa Maria, CA 93455, USA

[‡]Department of Geography, University of Oklahoma, 100 E. Boyd Street,
Sarkeys Energy Center 684, Norman, OK 73019, USA

(Received 5th September 2002; in final form 10th August 2004)

While some geographic phenomena hold uniform properties, such as land-use zones, many geographic phenomena are distributed such 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 to meet such spatial analysis needs is that 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.

Keywords: Field-object representation; Spatiotemporal data modelling; Spatiotemporal query

1. Introduction

Over the past 20 years, the production of geospatial data has increased exponentially. Most significant 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. The shift poses representational challenges because different users may have distinct views to analyse and interpret these phenomena. Conventionally, geographic information

*Corresponding author. Email: john.mcintosh@m-e.com

systems (GIS) only represent and analyse distributed phenomena as raster fields, but individual features identifiable from distributed phenomena can be the primary indicators for how the phenomena evolve 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 combined strengths of the object and field representations enhance the ability to summarize and reason about spatiotemporal patterns within distributed phenomena. Cova and Goodchild (2002) acknowledged the need to extend geographic representation to include fields of spatial objects. They proposed an innovative means to map locations of discrete objects of significance in a continuous field. Hence, interactions between a field and an object space can be analysed dynamically as the location of interest in the field varies. Complementarily, our focus is placed in the identification of discrete objects in a field and examination of these objects as to how attributes vary in the space bound by individual objects and how spatial properties of these objects and spatial variations of their attributes change over time. With a focus on object identification, our approach emphasizes information needs on how objects move and evolve in space and time. Being embedded in continuous fields, objects identified in distributed geographic phenomena do not have clear boundaries. The boundary issue on geographic objects with indeterminate boundaries has been well addressed in a collection edited by Burrough and Frank (1996). Nevertheless, boundaries play a critical role in determining their spatiotemporal behaviours (e.g. determining how a low moves) and interactions (e.g. determining how 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 15 March 2000 to 15 June 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 analysing 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 modelled 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. Modelling 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 in terms of 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 behaviours, and spatial relationships. Figure 1 illustrates boundaries for an area of rainfall based on three thresholds: $>0 \text{ mm h}^{-1}$, $>20 \text{ mm h}^{-1}$, and $>40 \text{ mm h}^{-1}$. 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 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



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

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 different uses by allowing the data to be viewed in a variety of formats generated on demand from the original database (Parent et al., 2000). Although conceptually elegant, this approach is not always practical or even possible. For example, with temporal GIS, the geometry and attributes of modelled 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 modelled 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 boundaries of object-like features can be delineated automatically for various thresholds on an ad hoc basis, a multiple representational framework is taken in our proposed framework.

Maintaining multiple representations indeed increase 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 modelled 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. 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 explicitly stored in 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. As defined in equation (1), a zone (Z_n) is a spatially continuous aggregate of locations (S_x) with properties that meet a certain threshold (ℓ) at a given time (i.e. $T_x = \text{constant } C$) as indicated below, whereas $f(x)$ is the function to retrieve the value of the attribute of interest at location x .

$$Z_n = \{x(S_x, T_x) | \forall x \in S_x, f(x) \leq \ell, T_x = C\} \tag{1}$$

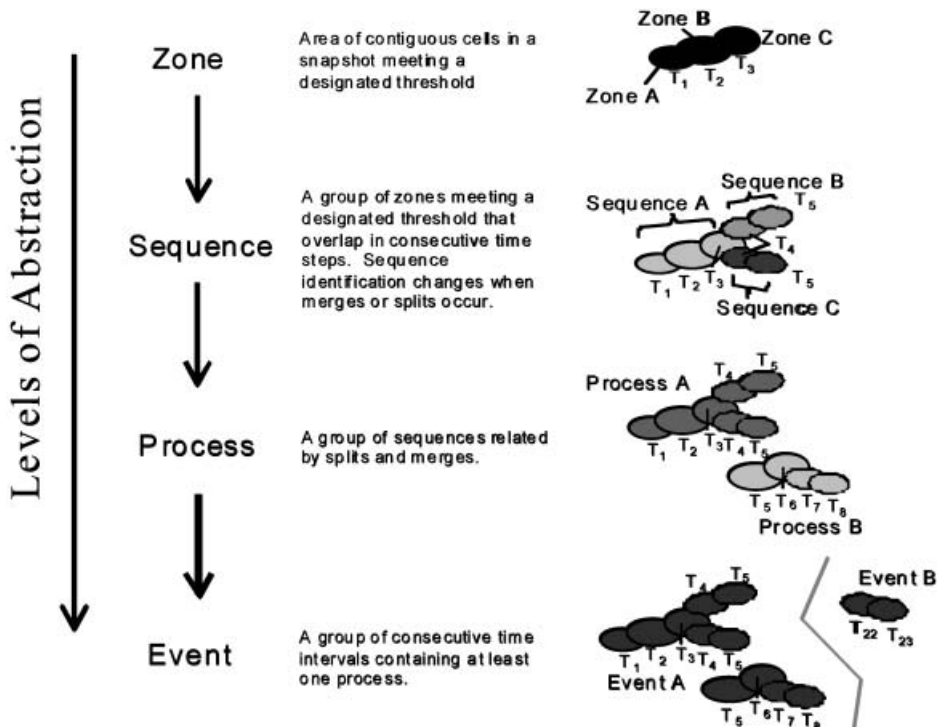


Figure 2. Organization of the object-based representational framework.

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. Hence, a sequence (Sq) is, defined in equation (2), an aggregate of overlapping zones (i.e. $S_{z_i} \wedge S_{z_{i+1}} \neq \phi$) over the space and time of these zones of interest (S_z , and T_z), and every zone in the sequence only has one and only one predecessor (i.e. $N(z_{i-1})=1$) and one and only one successor (i.e. $N(z_{i+1})=1$).

$$Sq = \{Z(S_z, T_z) | \forall z_i \in Z, S_{z_i} \wedge S_{z_{i+1}} \neq \phi, \forall z_i : N(z_{i+1}) = 1 \text{ and } N(z_{i-1}) = 1, T_{z_{i+1}} = T_{z_i} + 1\} \quad (2)$$

The strategy ensures that branching areas always have a distinct sequence identification number so that the junctures of branches and mergers can be easily identified. In addition, sequences at a higher threshold may be contained within a single lower threshold sequence. The framework explicitly stores *contains* and *part-of* relations for sequences and provide a means to link semantically related sequences to corresponding sequences at a higher or lower threshold. *Future* and *previous* relations are also stored to allow easy tracking of sequences that are involved in splits or merges as part of a process. Hence, a process (Pr) is defined here as a spatial (S_{sq}) and temporal (T_{sq}) aggregate of sequences (i.e. $Sq(S_{sq}, T_{sq})$) that share the same origin (i.e. sequences that are related through branching and merging). Therefore, for any sequence in a process, there is some time over the time of the process where the sequence overlaps with at least one other sequence in space and time (i.e. $T_{Sq_i} \wedge T_{Sq_j} \neq \phi, S_{Sq_i} \wedge S_{Sq_j} \neq \phi$).

$$Pr = \{Sq(S_{sq}, T_{sq}) | \forall sq \in Sq, \exists Tsq_j, \text{ where } T_{Sq_i} \wedge T_{Sq_j} \neq \phi, S_{Sq_i} \wedge S_{Sq_j} \neq \phi\} \quad (3)$$

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 of interest constitutes an event. Hence, an event is defined as a spatial and temporal aggregate of processes of interest (i.e. $Pr(S_{pr}, T_{pr})$). These processes may be separate in space,

but they collectively persist over consecutive periods of time (i.e. $T_{Pr_i} \wedge T_{Pr_j} \neq \phi$)

$$Ev = \left\{ Pr(S_{Pr}, T_{Pr}) \mid \forall Pr_i \in Pr, \exists T_{Pr_j} \text{ where } T_{Pr_i} \wedge T_{Pr_j} \neq \phi \right\} \quad (4)$$

With 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). 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, 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.

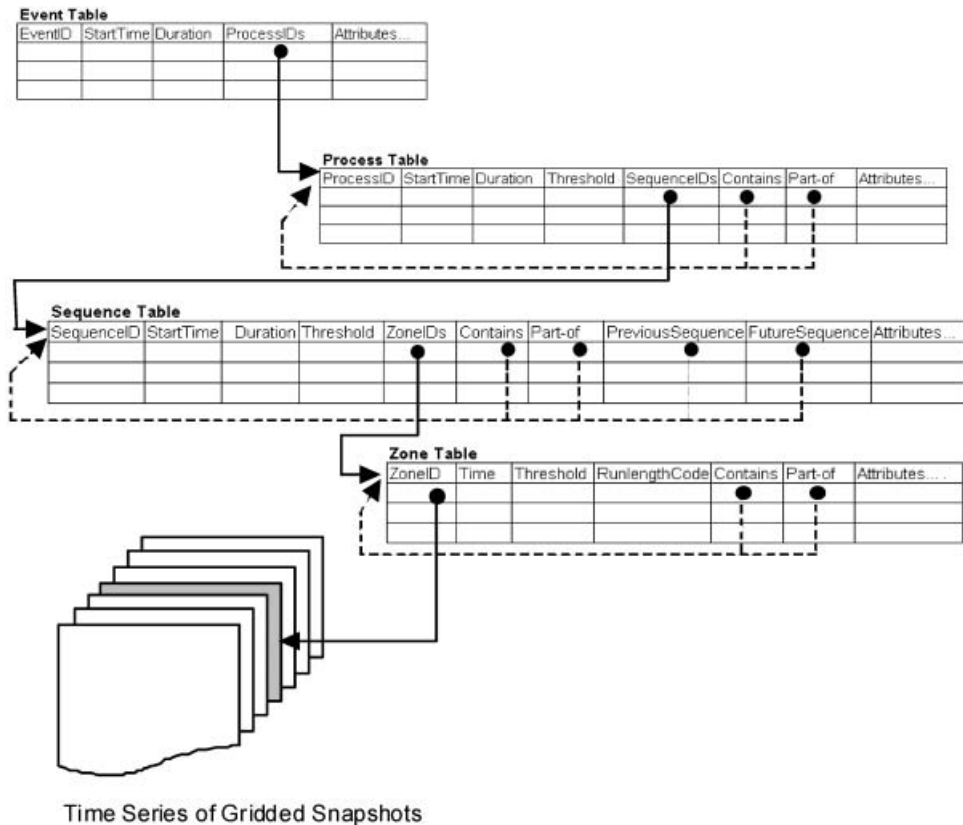


Figure 3. Data structures used to implement the framework. The solid arrows indicate fields that can be used to associate objects between tables. The dashed arrows indicate fields that can be used to associate objects within the tables.

The spatial extent of zones is stored as run length codes in the zone table. Since the zone table stores only the extent, we use a simple binary encoding scheme. Beginning at the lower left corner of the raster layer, cells are read from left to right advancing from the bottom to the top row. The code contains values representing the length of runs in the layer alternating between cells that are outside the zone and cells that are within the zone. In cases where the lower left cell of the raster layer is occupied by the zone, the first number in the code is set to zero preserving the ordering of the coding scheme. The square in the 5×5 raster layer in figure 4 would be represented as $\{6, 2, 3, 2, 12\}$. Beginning at the lower left cell, there is a run of six cells outside the zone followed by a run of two cells within the zone, followed by three outside the zone and so on. The run length encoding used here is independent of the data encoding of the original raster data set, and the origin for the runlength codes is set relative to the extent of the modelled area.

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 interpretation. In addition to the location information, 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.

The spatial extent 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 zone level stores cells of observations from raster layers. Sequences are aggregates of zones. The boundary of each sequence is not stored in the framework, so there is no duplicate boundary between zones and sequences. Rather, spatial properties and spatial relationships of a sequence are derived based on their corresponding zones. Similarly, processes are aggregates of sequences, and events are aggregates of processes. Their spatial properties are derived from corresponding objects at the lower levels.

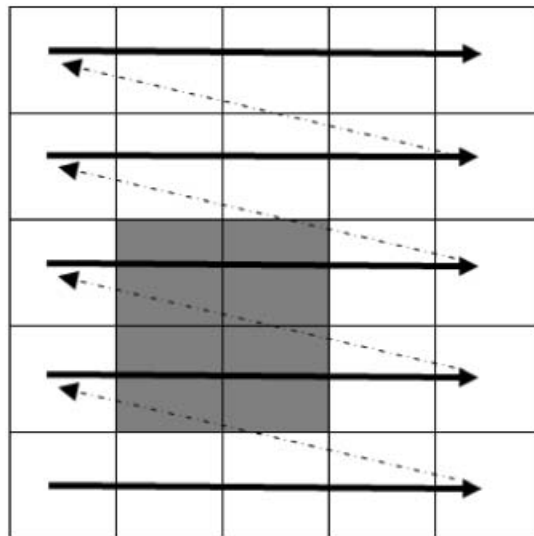


Figure 4. Run-length encoding example. The arrows indicate the order in which the cells are read. The square within the 5×5 raster layer is represented as $\{6,2,3,2,12\}$.

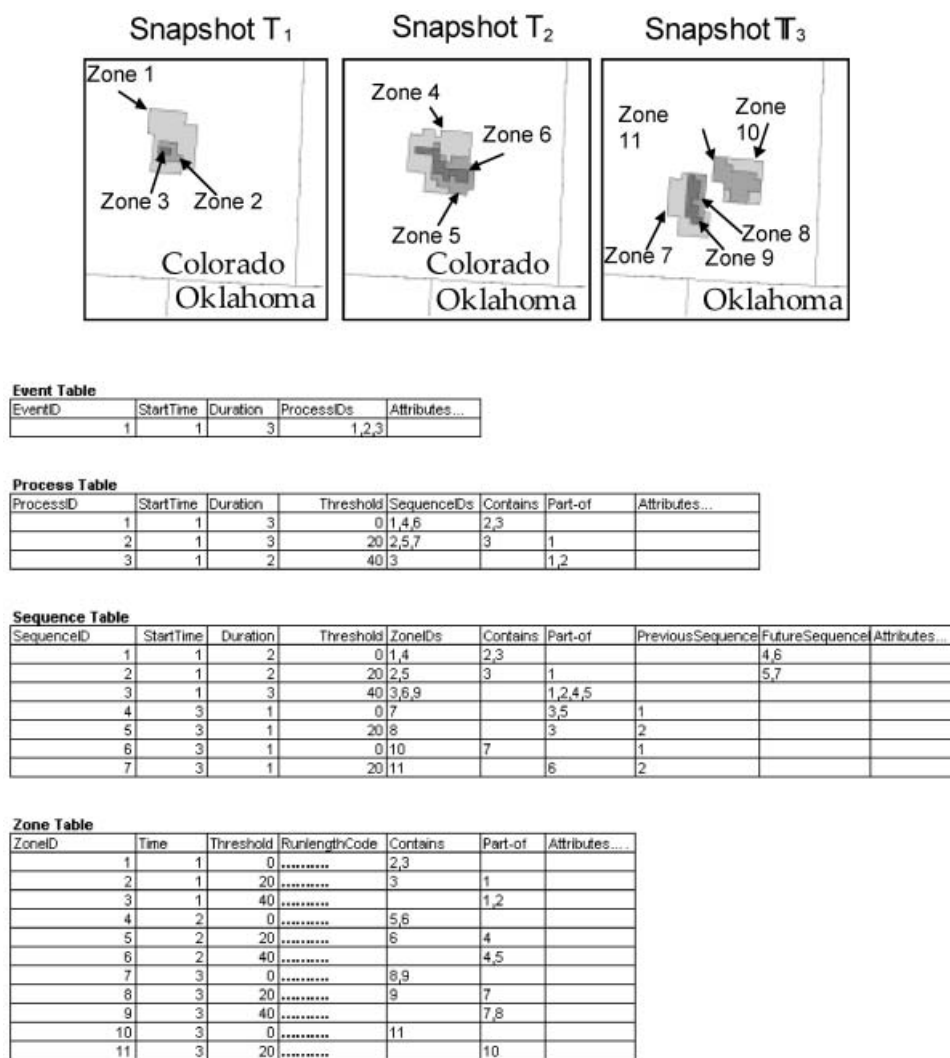


Figure 5. Illustration of how the framework would represent the three snapshots at the top. The zones delineated by three different thresholds are indicated in the snapshots, and the sequences, processes, and events they form are indicated in the respective tables (see text for further discussion).

Figure 5 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 (0 mm h^{-1}) with the middle and darkest shades corresponding to the middle ($>20 \text{ mm h}^{-1}$) and highest threshold ($>40 \text{ mm h}^{-1}$). This example shows a single event consisting of three processes, seven sequences, and 11 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. 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, CA). 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 importing rainfall data into GIS formats and developing algorithms to build zones, sequences, processes, and events from the rainfall 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 centre. 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

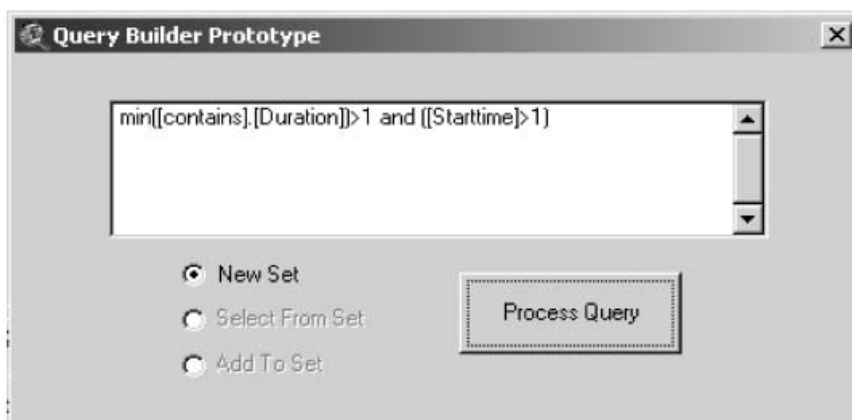


Figure 6. Query Builder Prototype. Applied to the sequence table, this query would return sequences that start after time 1 and contain sequences of a higher threshold with a minimum duration of two or more time intervals. Applying this query to the sequence table in figure 4 would return sequences 1 and 2.

important for exploratory analysis of these objects, including minimum, maximum, mean, variance, standard deviation, range, count, and sum (figure 6). 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 6 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 according to the following algorithm:

1. For each record in the sequence table
 - a. Get the contained sequence IDs from the 'contains' field
 - b. Count the number of contained sequences that have a duration greater than 1 time unit (in this case study, an hour)
 - c. Get the start time of the current record
 - d. Determine if the start time greater than 1 and the number of contained sequences from step b is greater than 1
2. Return all sequence records that meet the conditions of 1d in a new table.

Summary statistics 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 7). Furthermore, it allows the user to explore the relationship between the various zones, sequences, processes, and thresholds used to derive these objects.

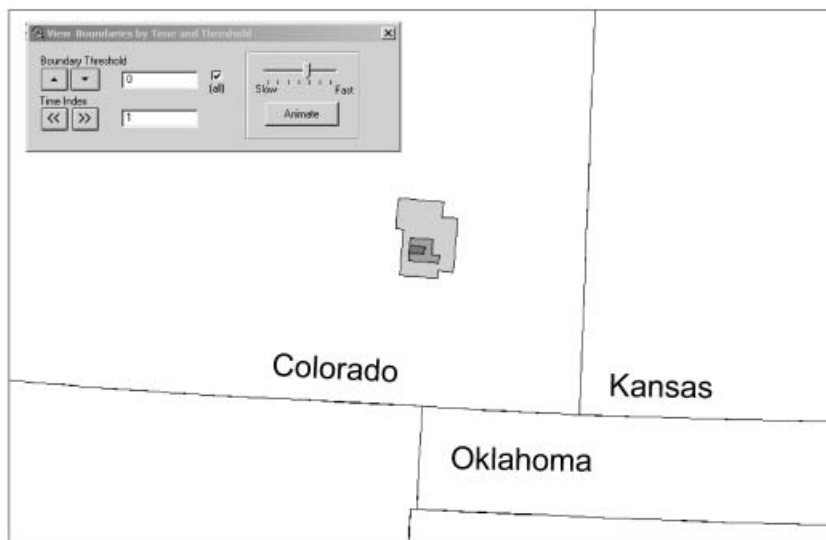


Figure 7. Viewer Dialog. This figure shows the viewer dialog with data included in figure 4. The viewer allows the user to step through the data changing the boundary thresholds and time.

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;
- to determine if the prototype is capable of handling multiple boundaries for object-like features and maintaining their topological and temporal relationships;
- to 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 search manually or interactively 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. Schiesser *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 that follow along a line, and linear aligned thunderstorms that move roughly perpendicular to the major axis. In the US, 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 *et al.* (1990), Schiesser *et al.* (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 are being studied to improve the efficiency of fertilizers or pesticides, the appropriate boundary might be the presence of rainfall (i.e. >0 mm).

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 8). The DPAs are in a raster format and consist of approximately $4\text{ km} \times 4\text{ km}$ 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

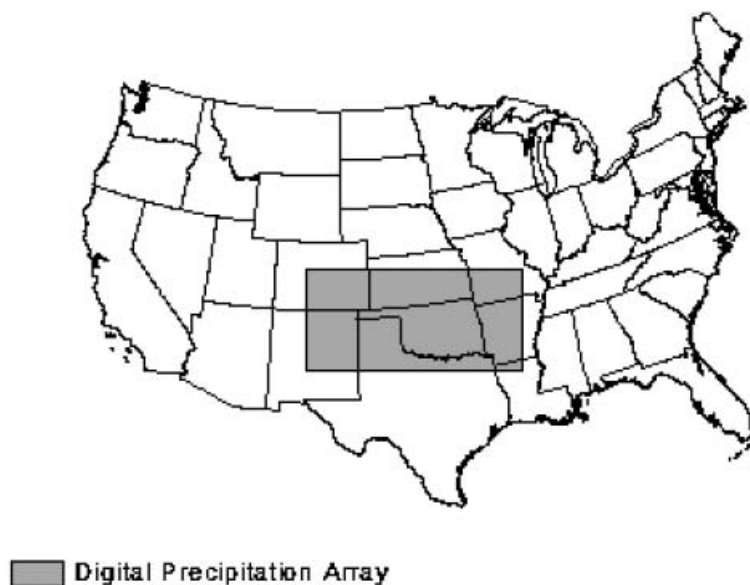


Figure 8. Arkansas Red Basin River Forecast Centre Digital Precipitation Array (DPA).

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 or agronomy. DPAs are in essence field representations of accumulated rainfall, ideal for hydrologic modelling and flood forecasting where the estimates of hourly rainfall accumulation are needed for discrete locations within the modelled domain. For other types of analysis, object-like characteristics within DPAs will be needed, such as spatiotemporal patterns and structures of rainstorms.

The framework was implemented using rainfall data from 15 March 2000 to 15 June 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' ($>0 \text{ mm h}^{-1}$, $>20 \text{ mm h}^{-1}$, and $>40 \text{ mm h}^{-1}$). 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 descendants. Rainfall events represent consecutive periods with rainfall somewhere in the modelled domain, with the understanding that rainfall processes within the same rainfall event may or may not be 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. As mentioned above, an important pattern in the Southern Plains is high-intensity convective precipitation imbedded in large rainfall areas of low-intensity stratiform precipitation (Houze *et al.* 1990). The *contains* and *part_of* relations can be used to identify processes or events in the database that correspond to such patterns. 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.

$$(\min([\text{contains}].[Threshold]) > 20) \text{ and } ([\text{Starttime}] < 46007)$$

In our implementation, 1 January 1995 at 0000 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 1 April 2000. Figure 9 shows one of the zones returned by the query (this zone is part of a rainfall event over the border of Texas and Oklahoma on 27 March 2000). With the 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 Schiesser *et al.* (1995). 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. Table 1 illustrates enhanced support of the proposed framework over a standard raster representation for a

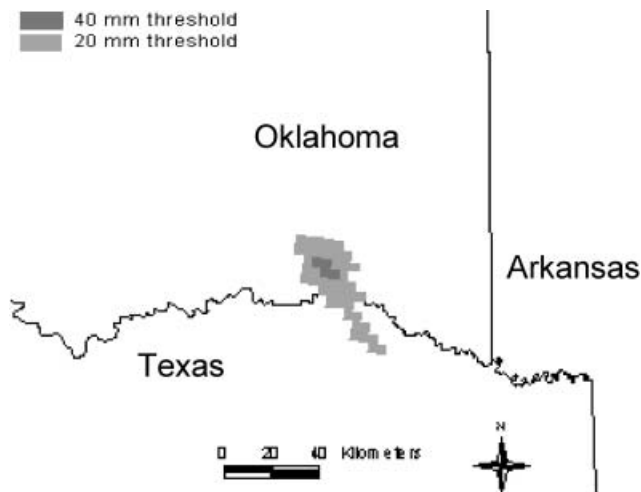


Figure 9. One of the seven sequences returned by the query described in the text. This sequence is part of a rainfall event that occurred over the border of Oklahoma and Texas on 26 March 2000 and 27 March 2000. The 40 mm h^{-1} threshold sequence shown in the figure had a duration of 1 h and occurred on 27 March 2000 at 0300 GMT.

Table 1. Supported query types.

Query type	Examples	Raster snapshot	Proposed framework
Attribute	What is the maximum rainfall value?	Yes	Yes
Simple spatial	Where is the rainfall greater than 100 mm h^{-1} ?	Yes	Yes
Spatial range	What is the mean rainfall within a rainfall area?	No, the geometry of zones is not stored	Yes, geometry of zones is explicitly stored, and spatial characteristics can be stored as attributes
Spatial relationship	What rainfall areas contain zones of rainfall in excess of 2 cm h^{-1} ?	No	Yes, certain spatial relationships such as contains and part_of are explicitly stored and can form the basis of such queries
Simple temporal	What was the average rainfall from 1:00 p.m. to 2:00 p.m. on 3 June 2000?	Supported at the individual cell level. Not supported for zones within raster layers	Yes, for individual cells and for zones within raster layers
Temporal range	What is the change in average rainfall of over the past hour?	Supported at the individual cell level. Not supported for zones within raster layers	Yes, for individual cells and for zones within raster layers
Temporal relationship	What rainfall events have a single rainfall area in one time interval followed by two rainfall areas in the next time interval?	No	Yes, certain temporal relationships such as next and previous ID are explicitly stored and can form the basis of temporal relationship queries
Simple spatiotemporal	Where was rainfall greater than 2 mm h^{-1} at 2:00 p.m. on 3 June 2000?	Supported at the individual cell level. Not supported for zones within raster layers	Yes for individual cells and for zones within raster layers
Spatiotemporal range	Where did the average rainfall rate decrease by more than 1 cm h^{-1} between 1:00 p.m. and 3:00 p.m. on 3 June 2000?	Supported at the individual cell level. Not supported for zones within raster layers	Yes for individual cells and for zones within raster layers
Spatiotemporal behaviour	How did the velocity of a rainfall process change over the life of the storm?	No	Yes, the behaviour of objects can be stored as attributes and used as a basis for behavioural queries.
Spatiotemporal relationship	How did entities of multiple kinds relate to each other in space and time?	No	No

variety of query types. Objects selected by the queries based on the spatiotemporal objects 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 trade-offs 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 modelled in each snapshot. A worst-case scenario in terms of incremental data costs occurs with maximum spatial and temporal variability in the conceptual objects being modelled. 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

$$S = O(3n_z^*(n_{thr} - 1) + c) \quad (5)$$

where n_z is the number of zones in all modelled time intervals, n_{thr} is the number of thresholds modelled, 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

$$Z_{max} = 1/2n_c n_{ts} n_{thr} \quad (6)$$

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 $O(3/2[n_{thr}^2 + cn_{thr}])$.

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 modelling rainfall data with three levels of boundaries based on the thresholds described above, the incremental cost is about 3.3%, 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 analysed 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 h with an average area of about 672 km² (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 with an average of 4.88 KB 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 behaviour 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 behaviour of the object-like parts as a basis for queries and analysis. If this assumption is not met, 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.

References

- ARKANSAS-RED BASIN RIVER FORECAST CENTER 2002, ABRFC precipitation products. Available online at: <http://www.srh.noaa.gov/abrfc/pcpnpage.html> (accessed 2002).
- 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*, A. Cremers and K. Greve (Eds), pp. 555–570 (Marlburg, Germany: Metropolis).
- BURROUGH, P.A., 1996, Natural Objects with Indeterminate Boundaries. In *Geographic Objects with Indeterminate Boundaries, GISDATA 2*, P.A. Burrough and A. Frank (Eds), pp. 3–28 (London: Taylor & Francis).
- BURROUGH, P.A. and FRANK, A., 1996, *Geographic Objects with Indeterminate Boundaries, GISDATA 2* (London: Taylor & Francis).
- BURROUGH, P.A. and McDONNELL, R., 1998, *Principles of Geographical Information Systems* (Oxford: Oxford University Press).
- BUTTENFIELD, B.P., 1995, Object-oriented map generalization: modeling and cartographic considerations. In *GIS and Generalization: Methodology and Practice*, J.C. Muller, J.P. Lagrange and R. Weibel (Eds), pp. 91–105 (London: Taylor & Francis).
- 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*, A.U. Frank, I. Campari and U. Formentini (Eds), pp. 65–77 (Berlin: Springer).
- COVA, T. and GOODCHILD, M., 2002, Extending geographical representation to include fields of spatial objects. *International Journal of Geographical Information Science*, **16**, pp. 509–532.
- DAVIS, C. and LAENDER, A., 1999, Multiple representations in GIS: materialization through map generalization, geometric, and spatial analysis operations. In *ACM-GIS 1999*, pp. 60–65.
- EGENHOFER, M. and MARK, D., 1995, Naive geography. In *Spatial Information Theory: A Theoretical Basis for GIS*, A.U. Frank and W. Kuhn (Eds), pp. 1–15 (Berlin: Springer).
- ERWIG, M. and SCHNEIDER, M., 1997, Vague regions. In *5th International Symposium on Advances in Spatial Databases (SSD'97)*, M. Scholl and A. Voisard (Eds), pp. 298–320 (Berlin: Springer).
- GALTON, A., 1997, Continuous change in spatial regions. In *Spatial Information Theory: A Theoretical Basis for GIS (Proceedings of International Conference COSIT'97)*, S.C. Hirtle and A.U. Frank (Eds), pp. 1–13 (Berlin: Springer).
- HAGEN, M., BARTENSLAGER, 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., AGOURIS, P. and STEFANIDIS, A., 2000, Navigating through hierarchical change propagation in spatiotemporal queries. In *Seventh International Workshop on Temporal Representation and Reasoning*, pp. 123–131 (Nova Scotia: IEEE Press).
- PARENT, C. and SPACCAPIETRA, S., 2000, Database Integration: the Key to Data Interoperability. In *Advances in Object-Oriented Data Modeling*, M.P. Papazoglou, S. Spaccapietra and Z. Tari (Eds), pp. 221–253 (Cambridge, MA: MIT Press).
- PARENT, C., SPACCAPIETRA, S. and ZIMANYI, E., 2000, MurMur: Database management of multiple representations. In *AAAI-2000 Workshop on Spatial and Temporal Granularity*, 30 July 2000, Austin, TX. Available online at: <http://lbdwww.epfl.ch/e/publications/articles.pdf/AAAI-STgranularity.pdf> (accessed 2002).
- 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*, J. Nievergelt, T. Roos, H. Schek, and P. Widmayer (Eds), pp. 59–69 (Berlin: Springer).
- 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. and OLSEN, B., 2000, A comparison of operational precipitation processing methodologies, NOAA technical memorandum NWS SR-205. Available online at: <http://www.srh.noaa.gov/abr/c/p1vol.html> (accessed 2002).
- TIMPF, S., 1997, Cartographic objects in a multi-scale data structure. In *Geographic Information Research: Bridging the Atlantic*, M. Craglia and H. Couclelis (Eds), pp. 224–234 (London: Taylor & Francis).
- 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)*, 6–7 December 1999, Tokyo. Available online at: <http://lbdwww.epfl.ch/e/publications/ADBS99.pdf> (accessed 2002).
- 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*, R. Laurini, K. Makki and N. Pissinou (Eds), pp. 57–62 (Washington, DC: The Association for Computing Machinery Press).
- YUAN, M., 2001, Representing complex geographic phenomena with both object and field-like properties. *Cartography and Geographic Information Science*, **28**, pp. 83–96.

Copyright of International Journal of Geographical Information Science is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.