Comparing GIS Data Types within the Arc Hydro Data Model

Elisabetta T. DeGironimo

Abstract

The advent of an object-oriented GIS data model has facilitated the use of spatiotemporal hydrological event data. This paper presents the results of a master's thesis entitled "Spatiotemporal Queries of Hydrologic Event Data: A Comparison of GIS Data Types." The objective of this thesis study was to determine if there is a computational advantage to using a geodatabase based on the Arc Hydro data model over the coverage or shapefile data types using ArcGIS software. In order to determine if there was a computational advantage to the geodatabase, thirty queries were constructed and programmed in Visual Basic for Applications and run independently of user intervention against the three data types.

This paper describes the findings of a Master's thesis examining spatiotemporal event management (using a GIS) as it relates to watershed hydrological data.

A spatiotemporal event is something that happens at a specific location at a specific point in time. In the context of watershed management, a spatiotemporal event could be a single stream flow reading or the collection of a water quality sample. Each happens at a specific location at a specific moment. The objective of the thesis study was to determine if there is a computational advantage to using a geodatabase over the coverage or shapefile data type when querying watershed hydrological event data.

Discipline-specific data models have been developed for ArcGIS, including one for water resources called Arc Hydro. The time series object class of the Arc Hydro data model has the capacity to store hydrological temporal event data. The advent of the Arc Hydro data model with a time series component is ushering in mainstream usage of hydrological event data within a GIS.

The ability to query by time and location is important to scientists and engineers studying hydrological data. It is of interest, for example, to be able to access all stream gauge, rainfall gauge, and water quality data collected during a particular storm event for a given watershed. There should be an advantage to using a geodatabase (Arc Hydro) to access this hydrological event data because of the object-oriented (less abstraction of geographic phenomena) structure of ArcGIS. The time series object class in Arc Hydro was designed to handle this type of data.

In order to determine if there was a computational advantage to using a geodatabase, a framework watershed GIS (Arc Hydro Framework with Time Series data model), which included subwatersheds, waterbodies, streams, and monitoring points was created for the West Canada Creek watershed in central New York State using each of the three data types (geodatabase, coverage, shapefile). The same source data was used to create each of the three data sets. Thirty (30) queries were constructed and programmed in VBA

(Visual Basic for Applications) to run independently of user intervention. The thirty queries were executed ten times on each of the three data sets (900 total query runs). The time, as measured by elapsed system time (according to the battery-operated computer clock), to complete each individual query was automatically recorded in a text file. The results were then statistically analyzed.

The spatiotemporal event data for the study was downloaded from the National Water Inventory System (NWIS) website. The event data from NWIS (64,285 records) was loaded into a MS Access database (TimeSeries table in Arc Hydro). It contained stream gauge as well as water quality information. Records in the TimeSeries table were related to the MonitoringPoint dataset that contained points representing the 33 event locations in the watershed. The HydroID field in the MonitoringPoint file was related to the FeatureID field of the TimeSeries table (as it is in the Arc Hydro data model).

When querying event data in a GIS, four query outcomes are possible:

1. Features on the map are selected;
2. Records in a related table are selected;
3. Both features and records in a related table are selected;
4. Nothing is selected.

This study considered the first three outcomes listed above. The fourth outcome was not considered because it does not produce a visible result set (to assure that the query worked). In order to insure that either features or records were selected, the event data analyzed and queries were composed that selected varying numbers of records and/or features. The queries were also designed to have varying degrees of complexity. For discussion purposes, the first type of query is considered a spatial query; the second, a tabular query; and the third, a mixed (or compound) query. Of the thirty queries executed against each of the three data types, ten (10) were spatial, ten (10) were tabular, and ten (10) were mixed.

EXAMPLE QUERIES

Query Type	Description	Tables Queried	Records Selected
Spatial	Select all spring or lake/reservoir monitoring points	MonitoringPoint	6
Spatial	Select all stream monitoring points in the Center or Middle West Canada Creek subwatersheds	MonitoringPoint	14
Tabular	Select all gauge readings for July 4 – 5, 1992	TimeSeries	2
Tabular	Select all water quality samples taken 1990 or after with a Total Coliform count >= 1000 colonies/100 ml	TimeSeries	6

Query Type	Description	Tables Queried	Records Selected
Mixed	Select all monitoring points (event sites) with a pH reading < 6 in the Middle West Canada Creek subwatershed in 1994	MonitoringPoint TimeSeries	1 1
Mixed	Select all gauges with readings > 10,000 cfs in March	MonitoringPoint TimeSeries	2 13

By pressing a button in the button bar, the user launches the control program (written in VBA) that calls each of the queries. The queries each have a timer that starts after they are called and stops after the individual query is executed. Once the timer stops, the elapsed time (system time (recorded to a precision of 10^{-7} seconds)), along with the query name and data type, is written to a text file before control is passed back to the control program. The control program terminates after each of the queries has been run 10 times.

Once all of the queries were run for each of the data types the results were statistically analyzed (GLM ANOVA). Of the thirty queries, the geodatabase was the fastest to process twenty-two (22) of the queries. The coverage was fastest for the remaining eight (8). In many cases, the performance of the shapefile was only slightly slower than that of the coverage.

As a group, the spatial queries were quickest to process (none taking longer than 0.3 seconds). The next fastest were the tabular queries (the slowest executing in less than 1 second). The slowest queries were the mixed queries. These were the most complex since they used a relate to select both features and any related tabular records. Five of the mixed queries executed in under 1 second (for all data types). Three of the queries executed between 1 and 3 seconds. The remaining two queries took approximately 64 and 274 seconds, respectively, to execute. Because one of the assumptions of an ANOVA (i.e., equal variance within each group) was not met, a logarithmic transformation of the response variable (i.e., query time) was needed.

Results of this study indicate that the geodatabase is statistically significantly faster than the coverage or shapefile data types. This study showed that there is a computational performance advantage, in terms of elapsed system time, to using the geodatabase data type (Arc Hydro Framework with Time Series data model) over the coverage or shapefile data types when querying spatiotemporal hydrologic event data in ArcGIS.

The geodatabase offers additional advantages over the coverage and shapefile such as storing all data in one computer file and being object-oriented. Because it’s object-oriented, features stored in the geodatabase are objects that can have properties and methods. The ability to have methods (programming code stored within the object) allows the object to exhibit behavior.

Data Type Comparison (for West Canada Creek Watershed) (Arc Hydro Framework with Time Series schema)					
Shapefile		**Coverage**		**Geodatabase**	
Number of Files	**File Size (sum of all files)**	**Number of Files**	**File Size (sum of all files)**	**Number of Files**	**File Size (sum of all files)**
36	7.87 Mb	71	7.56 Mb	1	9.59 Mb

The geodatabase and Arc Hydro data model show promise for managing spatiotemporal hydrological data. As it is currently designed, the Time Series component of the Arc Hydro data model is better suited to handle gauge data than water quality data. Of course, the user always has the option of modifying the Arc Hydro data model to better suit a given project. In fact, the purpose of developing data models such as Arc Hydro is to provide a starting point when developing a discipline-specific geodatabase design.

Author Information

Elisabetta T. DeGironimo
Watershed Coordinator
Mohawk Valley Water Authority
1 Kennedy Plaza
Utica, NY 13502
(315) 792-0353
(315) 792-5201 (fax)
edegironimo@mvwa.us