Design of a Geodatabase for Efficient Retrieval of Geographic Information

Sahadeb De, Manidipa Dasgupta

Abstract

Spatial data represent a new dimension in the field of information retrieval. Various access levels also need to be defined for sensitive data. In this paper we deal with the design and development issues of a statewide spatial database within an enterprise level GIS environment. The topic compels the need to explore applications and supported data formats available on the server and client platforms, and additional, or multiple retrieval formats to be made available on the server machine. In this study we have used SQL Server and ArcSDE in the back end for data storage, indexing and data access, and ArcGIS suite of applications at the front end for data query and retrieval. A customized user interface has also been developed with ArcObjects for enhancing the process of efficient geographic information retrieval and display.

1. Introduction

Information Retrieval (IR) had made a big advance in the present perspective of tremendously high demand in search and retrieval of data. As more and more volume and variety of data become digital, their storage formats, indexing techniques, retrieval processes and display formats gain more variety and advanced dimensions. In addition to the original Boolean and Vector Space models used in categorizing queries for information retrieval, individual database libraries often use extensions to these models, along with fuzzy sets and probabilistic reasoning and interpretation, to augment the robustness of their information retrieval systems, and suit the variability of their data.

IR is of considerable significance in the field of Geographic Information Systems (GIS). With the advent and rapid proliferation of storing spatial data in digital format, people have to deal with enormous amount of associated attribute data. The general idea is to be able to query and retrieve both spatial and associate attribute data using GIS. GIS allow querying of the spatial data based on queries only on the attribute information stored in the tables. These queries are based on the principle underlying SQL, but much simpler and very specific or restricted. They can only use Boolean operators (and, or, not) and some mathematical operations (>, >=, <, <=, =). The queries are done locally and the query results cannot be stored in any transient format. To store them, we must create additional shapefiles. It can be said that there are two major disadvantages to this kind of IR in GIS - one, the query criteria is restricted, and two, results cannot be stored in transient and readily available form.

In this paper we focus on the design and development issues of spatial databases for easy retrieval of geographic information. We have used ArcInfo™ 8.x as the ‘front end’,

ArcSDE® as the middleware, product of Environmental Systems Research Institute (ESRI) and Microsoft's SQL Server™ as the 'back end' relational database management systems.

The organization of this paper is as follows. Section 2 describes different information retrieval models and their relevance in geographic information system. Existing geographic information retrieval models are briefly described in Section 3. The design of a spatial database has been given in Section 4 for efficient retrieval of geographic information along with some experimental results with ArcGIS 8.1™. Finally, Section 5 depicts the conclusion along with problems related GIR.

2. Information Retrieval and GIS

Information retrieval in general, uses probabilistic models and various statistical inferences to group, index, categorize, and actually provide access to documents. Retrieval algorithms are however deterministic, and use various techniques like Vector Space models, Boolean models and other specific algorithms to retrieve documents matching query results, and weight the query terms to determine ranking of retrieved results. Indexing for non-spatial IR usually involves simple extraction (such as extracting keywords from a text), inferential extraction (such as mapping from text word to thesaurus terms) or it may be intellectual analysis and assignment index items (such as assigning subject headings to a document). In Geographic Information Retrieval we are concerned with both deterministic retrieval (such as finding all data sets that contain information on a particular coordinate) and probabilistic retrieval (such as finding all towns near a major river). Thus the algorithms used in actual retrieval must be more specialized and should span both deterministic and probabilistic algorithms and

techniques. Indexing becomes a major challenge when preparing geospatial data for spatial browsing, query and retrieval. It should be remembered that when querying geospatial data, it is possible that the user is looking for both spatial data and related non-spatial information. The query and retrieval processes, as such, must provide for suitable query interfaces to facilitate both. Indexing in GIR, in addition to the usual indexing methods mentioned above, must allow for specialized forms of intellectual indexing (such as assignment of bounding box coordinates to an aerial photograph), and inferential indexing (assignment of coordinates for places mentioned in a text). Retrieval algorithms and models must be effective and efficient to allow proper mapping and results. Indexing in GIR must be such that it provides approximate, partial and also strictly deterministic matching. Traditionally, Boolean logic has been used to provide strictly deterministic matches to queries. Question then remains, what kind of models will be suitable for providing the other kinds of matches in GIR.

3. Existing GIR Models

Having discussed the possibilities of potential schemes of indexing and kinds of retrieval, which should ideally be made possible for geospatial data, let us proceed to discussing what is currently available in terms of query and retrieval, from the most current GIS environments and query interfaces available.

3.1 Larson's model on GIR

Larson defines that Geographic Information Retrieval, is an applied research area that combines aspects of DBMS research, User Interface Research, GIS research, and Information Retrieval research, and is concerned with indexing, searching, retrieving and

browsing of geo-referenced information sources, and the design of systems to accomplish these tasks effectively and efficiently (Larson, 2002).

The terms *geographic queries* and *spatial queries* imply querying a spatially indexed database based on relationships between particular items in that database within a particular coordinate system (or compatible coordinate systems) (Samet et. al., 1994; Larson, 2002). Spatial querying is the more general term. It can be defined as queries about the spatial relationships (intersection, containment, boundary, adjacency, proximity) of entities geometrically defined and located in space (De Florianio et. al. 1993, *in* Larson, 2002) without regard to the nature of the coordinate system. Geographic querying is a more specialized form of spatial querying and assumes that the space is delineated by the well-defined coordinate systems of the "real world." Spatial relationships may be both geometric and *topological* (spatially related but without measurable distance or absolute direction). Examples of topological relations include such properties as adjacency, connectivity, and containment (Larson, 2002).

Larson considers the following to be examples of spatial queries and discusses their retrieval processes in detail:

1. Point-in-polygon queries (this type of query essentially asks for any geo-referenced object or geographic dataset that contains, surrounds or refers to a particular spot on the surface of the earth. This is one of the most precise of all spatial queries).

2. Region queries (asks for relevant data within a particular region).

3. Distance and Buffer Zone queries (data within a particular distance surrounding spatial entity).

4. Path queries (distance queries, shortest distance, etc).

5. Multimedia queries (a combination of multiple geo-referenced sources in a single query).

In order to meet retrieval requirements for these kinds of queries, geographic or spatial indexing techniques have to use geographic coordinates for indexing data and associated documents, in addition to the other common indexing techniques. Geospatial data indexing usually tries to provide for the following: *identifying geographic place names and phrases, locating pertinent data, overlaying polygons to estimate approximate locations*. Probabilistic algorithms are developed such that query term weighting can be done and relevant matches are derived using that indexing scheme, and results can be retrieved. In order to do this kind of complex work, it would be advisable to develop some kind of automatic indexing technique. GIPSY, an experimental technique for automatic geo-referencing of text, developed at U.C. Berkley is one such. It uses a geographic thesaurus developed by the U.S.G.S to determine, understand and weight all significant content-bearing geographical words in a query or text.

3.2 ArcIMS's Retrieval Methods

Typically, current forms of most common geospatial data are geo-referenced and include topological and imagery data. Most spatial query interfaces depend on some sophisticated application, to derive and administer the data (Harris and Clark, 1999). One of them is ArcIMS. This application helps "serve" geographical data and images on the Internet with the potential for retrieving results for user queries for non-spatial attribute information for the spatial data. The spatial data can be served using the usual *shapefile* format or using *layer* format, and optionally administered by ArcSDE. On an inter-organization or system-specific level we can also use ArcView and ArcInfo to query and

display spatial data and associated non-spatial attribute information. The disadvantages to these kinds of queries are that they are restricted to SQL or SQL-like commands. Additionally, one or more spatial data layers must be first displayed before any queries can be imposed on them. In the event the query yields some results, the matching spatial data is automatically highlighted for recognition. There is no possibility of true spatial queries – the user cannot define a coordinate window, nor query using specific geographic coordinate information to produce required data. There is absolutely no possibility for natural language queries. Even when using ArcIMS, query topics are usually chosen from a drop down list, and these are transformed by the application into SQL formats and processed by the system to produce relevant results. These limitations are based mainly on the fact, that the indexing scheme for geospatial data is not complex and sophisticated enough to support other kinds of queries. The only indexing available for shapefiles and coverages is purely an internal one, where each individual topological feature is assigned an internal id when it is created by the digitization or vectorization process. This id is most often hidden to the user, and is part of the internal structure of the spatial data. It is rarely possible to query on the basis of that, and even when possible must, once again take the form of SQL format. Queries involving intersection, containment, boundary, adjacency and proximity of entities, are quite impossible in ArcView, ArcInfo and ArcIMS, and are in reality undertaken as specific spatial operations in the first two applications.

3.3 ArcGIS's Retrieval Methods

Indexing, retrieval and query of geospatial data has reached a new dimension with the arrival of ArcGIS (8.1). Its structure and setup for creating data layers, use of personal

and enterprise level geodatabase to represent spatial data using ArcCatalog, and its interfacing techniques with Microsoft SQL Server and ArcSDE, make it one of the most distributed and durable spatial data representation and manipulation systems. There are far more enhanced internal indexing techniques available with such spatial data. Each data layer represents a table in a personal or enterprise geodatabase. Unlike ArcView and ArcInfo, it is possible to query and display part of a spatial or topological data set using ArcSDE (Harris and Clark, 1999).

In ArcGIS, there is an in-built searching method for Feature classes. But instead of searching tuples, these processes involve the search for entire table(s)/feature class(es) within a database or group of databases. With this search tool, one can search the entire database or any folder on the basis of Name & Location of the data (in the databases or in folder), Geographic extent, Date, and Metadata elements.

'Name & location' pane asks for 'name', 'type', 'search' in folder or database and 'look in'. 'Geography' pane asks for whether or not to use geographic criteria. If user chooses to use geographic criteria, there are four possible ways for user input. He/She can draw a rectangle on the map, put latitude and longitude in the desired text boxes, pick a continent from the top right drop down box or pick a map from the bottom right drop down box to chose from. The 'Date' pane asks whether or not to use date as a criteria and if so, ask the user to put the desire date or dates. These dates indicate the date of creation of the digital data. In the 'Advance' pane user have the choice to search the spatial data with the help of its metadata. Here, a particular string (value) can be searched in metadata elements to find the spatial data.

3.4 Other GIR Models

Many GIS clearinghouses and government organizations provide a good infrastructure for spatial digital libraries, used mainly for download of geo-referenced data. Department of Natural Resources of South Carolina (SCDNR) is one such agency. They have developed some sort of indexing technique for their digital spatial data library holdings. This organization allows user to query the database using very limited natural language query, and possibly extract the words of importance, by probably using some of the currently popular techniques. Then, some algorithm is to get the best matches against their available holdings, and a list of most possible digital data are displayed. The user gets to download his choice from the list. Still this query interface and techniques are really not up to the mark, as obviously their data holding are indexed by keyword and geographical location category, rather than by true coordinates. One cannot search using true spatial or coordinate windows, nor search by the five spatial query techniques discussed by Larson (2002). Additionally, the user must download or view the entire spectrum of the data holding of his choice.

4. Our Work: Design and Development of a Geodatabase

Objects in a real-world system often have particular associations with other objects in the database. These kinds of associations between objects in the geodatabase are called relationships. Relationships can exist between spatial objects (features in feature classes), between nonspatial objects (rows in a table), or between spatial and nonspatial objects. While spatial objects are stored in the geodatabase in feature classes, and nonspatial objects are stored in object classes, relationships are stored in relationship classes

(MacDonald, 1999). With the help of this relationship class, it is possible to retrieve data (tuples) from a single table/feature class.

In the present study, as a first phase of experiment/research, a small module has been developed with one geospatial data layer (Point locations of Animal Feeding Operations in South Carolina) from a geodatabase. To develop this module, the database has been designed first utilizing the capabilities of relationship class. In the field of GIS, the extent of data is always dependent on user's need. It is not always possible to keep the spatial data in an enterprise level GIS environment in every possible extent because of the space requirements. Though it is possible to keep the data statewide (or largest extent possible), retrieve it in the client application (ArcGIS in this case), and 'clip' it with the desired area ('study area'), often it is time-consuming affair. In the present study, we keep all our data with statewide extent. Keeping in mind that the user may want to query a geographic object on the basis of either political boundary (State, County, School District etc.) or natural boundary (Watershed like Hydrologic Unit Codes) a series of relationship classes have been developed. There are three kinds of geographic objects: point, line and polygon. A point in geographic space may lie on one and only one polygon (Figure 1a). A line though may lie on more than one polygon but we can divide that line into more than one line segments to restrict a single line segment to fall on a single polygon. In this case a node has to be generated at every polygon boundary (Figure 1b). A polygon may lie in more than one polygon (Figure 1c).

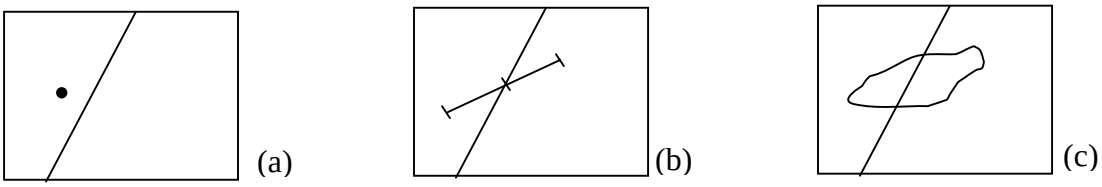
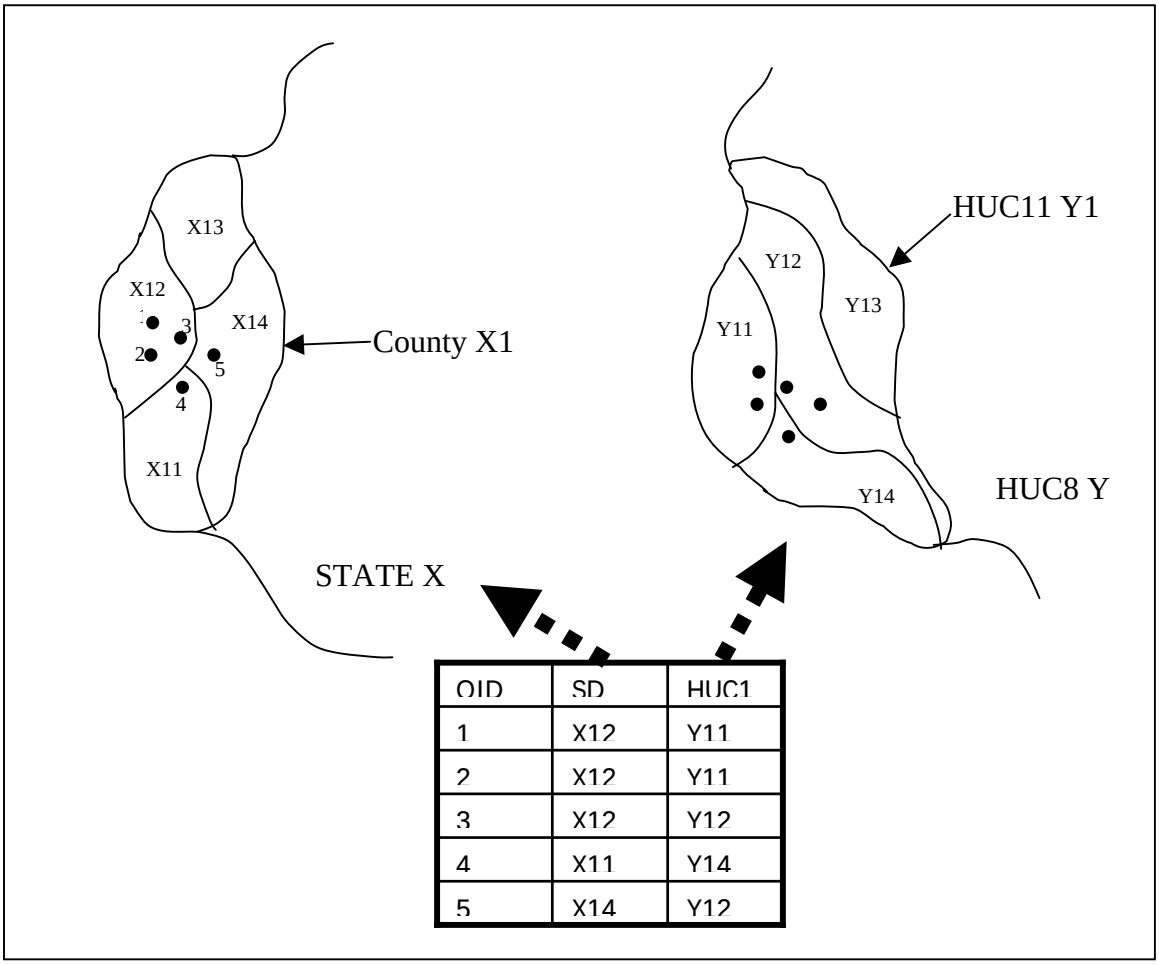


Figure 1: Geographic objects in space



OID	SD	HUC1
1	X12	Y11
2	X12	Y11
3	X12	Y12
4	X11	Y14
5	X14	Y12

Figure 2: Five Animal Feeding Operations on geographic space

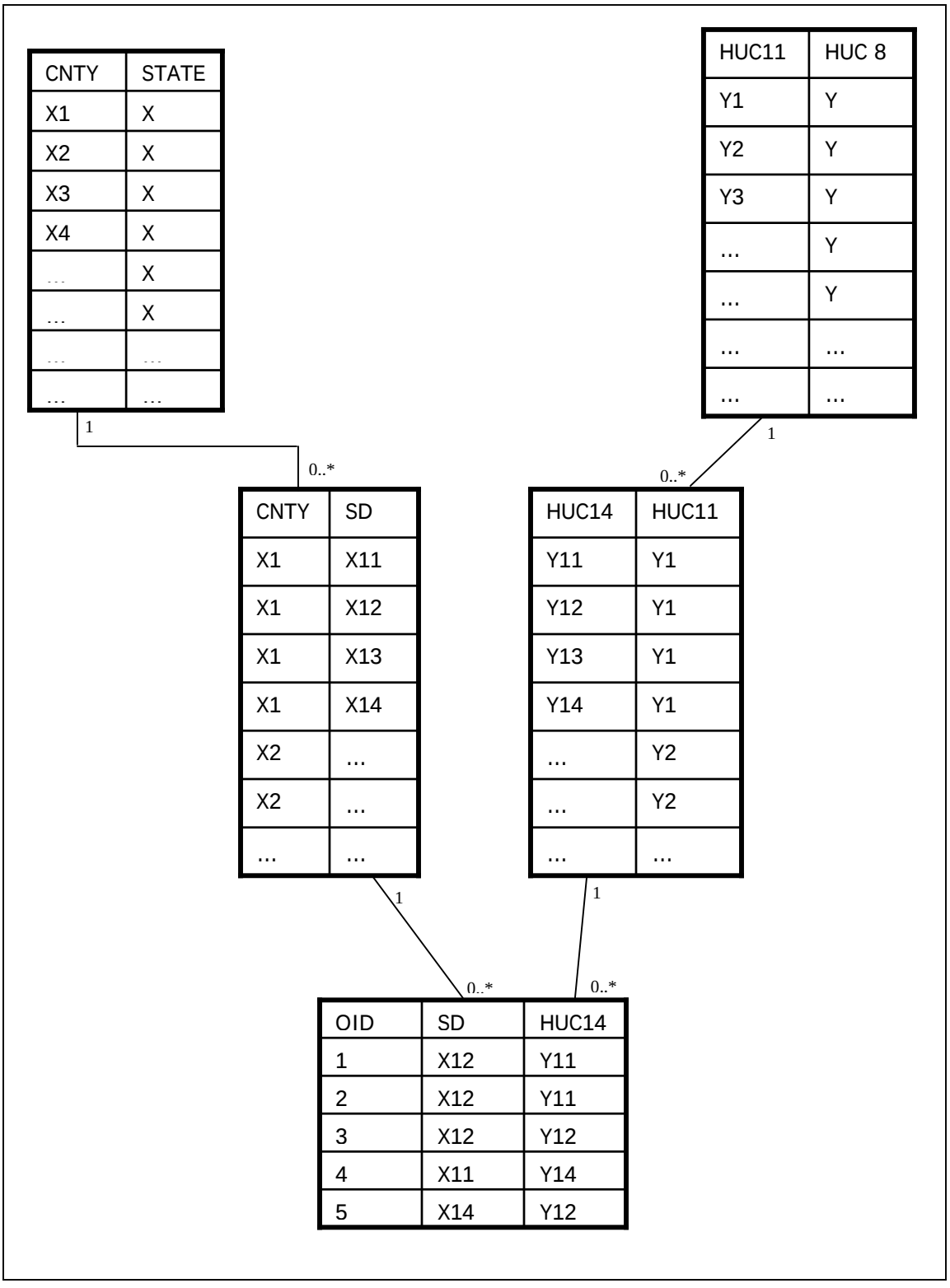


Figure 3: The relationships between the AFO and political or natural boundaries

In the present study, the feature class Animal Feeding Operations in South Carolina is a point class. So each point may lie on a single polygon (either political boundary or natural boundary). Thus each point in geographic space though has a fixed latitude and longitude, it may be contained by one political polygon (the lowest subdivision such as school district) and one natural polygon (the lowest subdivision such as HUC 14) (Figure 2). Again, these political boundaries and natural boundaries have their own hierarchical relationships, such as a state may contain many counties but a county belongs to one and only one state, a county may contain many school districts but a school district must fall on a single county, HUC 8 contains many HUC 11 polygons but a single HUC 11 belongs to a single HUC 8, and so on (Figure 3).

Code has been written in Visual Basic for Application environment to generate these two columns (one smallest political boundary ID and one smallest natural boundary ID as in Figure 2) within the AFO feature class.

As these two columns will be queried most frequently, indexing is necessary. When a feature class is created in ArcGIS using ArcCatalog or ArcSDE, an automatic spatial index is added to every feature in a feature class, whether that class is a table/layer in a personal geodatabase, or in an ArcSDE managed database stored in Sequel Server. Once there is data in a table or feature class, attribute indexes may be created to make queries faster. Spatial indexes increase the selection speed of graphical queries on spatial features. An attribute index is an alternate path used by the RDBMS to retrieve a record from a table.

Attribute indexes have been created for these two fields on the AFO feature class and table property pages. The same property page of the feature class may also be used to both add and delete spatial indexes.

Building a new spatial index for an ArcSDE feature class is a very server-intensive operation, and should not be done on very large feature classes when a large number of users are logged into the server

Once the design and development of the geodatabase is done, it is important to create the user interface to retrieve the data from the database and to display on the ArcGIS view. The user interface will ask a series of question from the user and will take as inputs to query the database (Figure 4). Upon receiving the inputs, it will display the subset of an entire feature class (Figure 5) either on the basis of political (Figure 6) or natural (Figure 7) boundary.

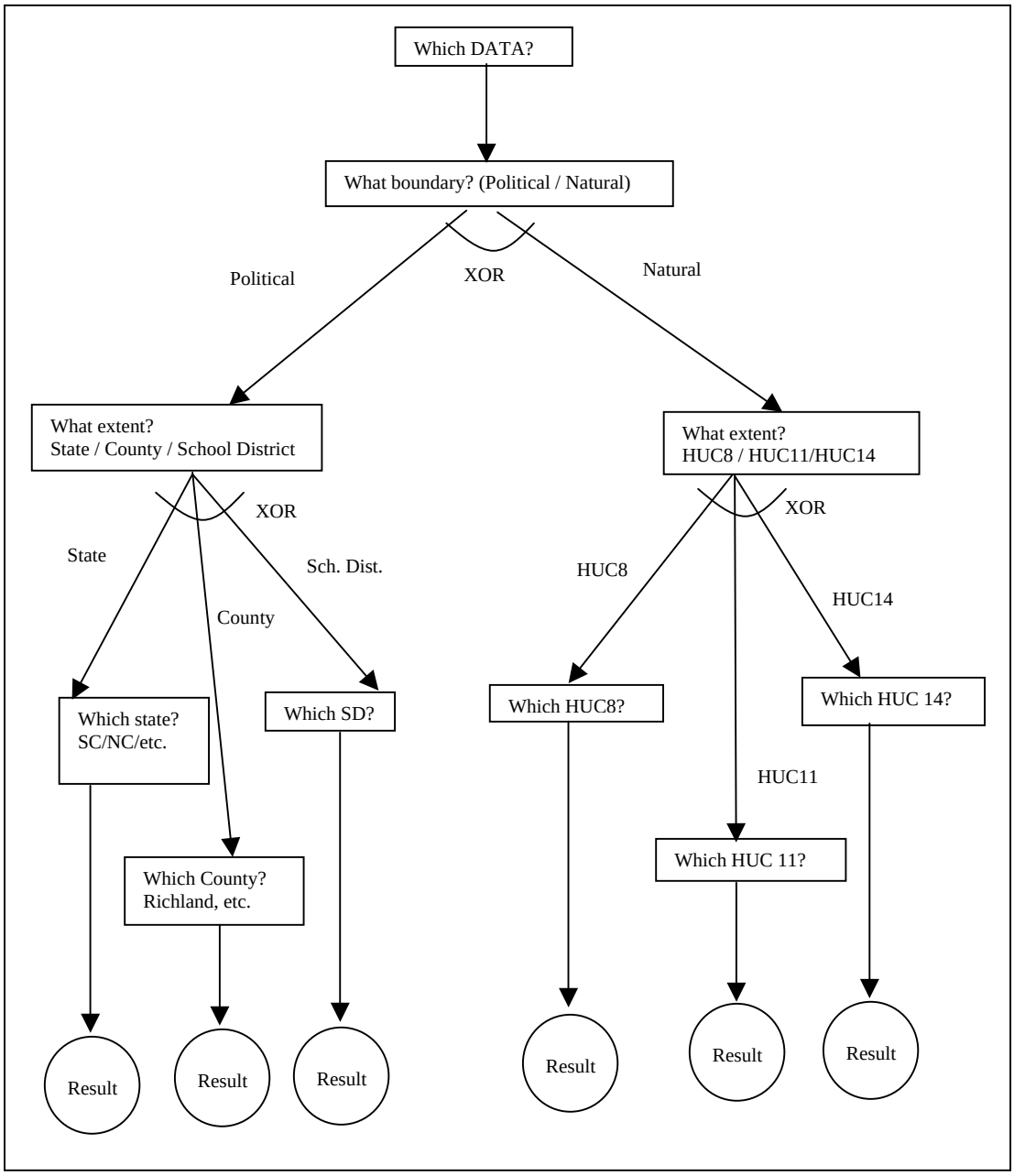


Figure 4: Flow diagram for the user interface attached to the ArcMap

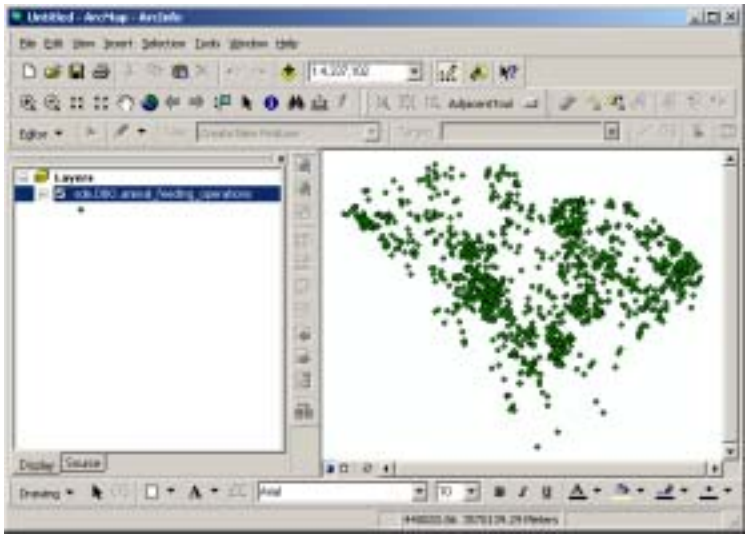


Figure 5: AFO Feature Class for entire South Carolina

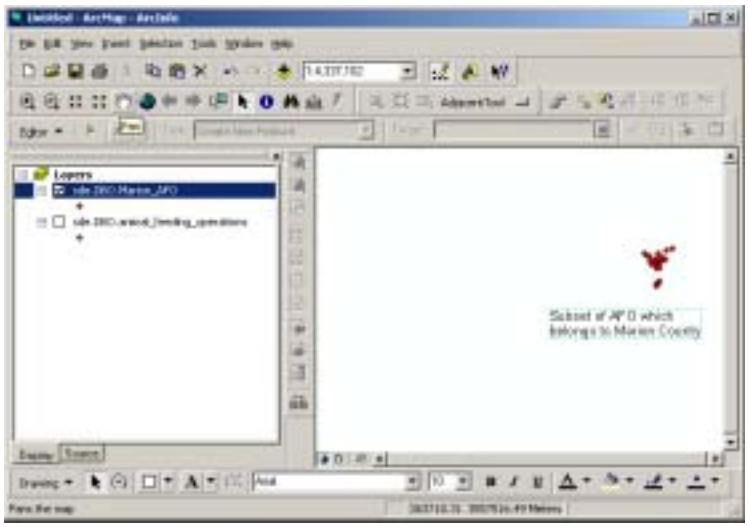


Figure 6: AFO Feature Class for Marion County, South Carolina

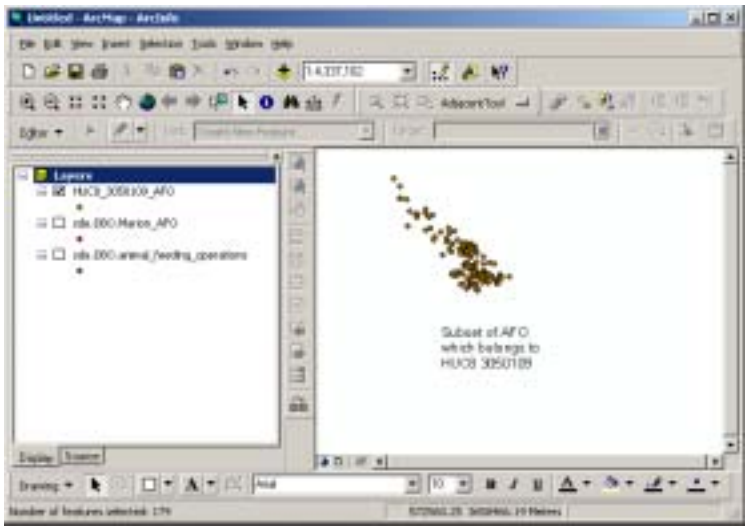


Figure 7: AFO Feature Class for HUC8 3050109

5. Conclusion

It is evident that ArcGIS together with ArcSDE can provide a lot of advanced indexing schemes and deals in SQL, SDE_specific and coordinate queries. These, when applied properly and in proper sequence by the user can be said to match up with Larson's (2002) type 1 and type 2 spatial queries, in terms of information retrieval. The other types are not achievable in this context. Information retrieval can be more successfully achieved than data retrieval using queries in GIS. The creation of queries for data and information retrieval require application-specific knowledge by the user. Natural language queries cannot be done, and ranking of retrieved results is not available, neither for retrieved data nor for retrieved information. More work needs to be done in terms of models and algorithms, to investigate whether user can achieve results using non-sql queries and, additionally do so from a GUI interface other than any GIS environment.

References

Francica, Joseph R., (2002) Large Spatial Databases, Business Geographics, *http://www.geoplace.com/bg/2000/0100/0100data.asp*, March 5, 2002

Harris, M. and Clark, J. (1999) ArcSDE Administration Guide, Environmental Systems Research Institute, Redlands, California, 180 p.

Larson, R. , (2002) Geographic Information Retrieval and Spatial Browsing, University of California, Berkeley, *http://sherlock.berkeley.edu/geo_ir/PART1.html*, March5, 2002

MacDonald, A. (1999) Building A Geodatabase. Environmental Systems Research Institute, Redlands, California, 327 p.

Rigaux, P., Scholl, M., and Voisard, A. (2002) Spatial Databases With Application to GIS. Academic Press, San Diego, California, 410 p.

Samet, H. and Aref, W.G. (1994) Spatial Data Models and Query Processing, Modern Database Systems, *http://citeseer.nj.nec.com/samet94spatial.html*, March 5, 2002

University of North Texas School of Library and Information Sciences, (2002)

Information Organization and Retrieval,

http://people.unt.edu/~skh0001/is/abstracts.HTM, March 5, 2002

Trademark Information

Microsoft, SQL Server are either registered trademarks or trademarks of Microsoft Corporation in the U.S.A. and/or other countries.

ArcInfo, ArcMap, ArcCatalog, ArcToolbox, ArcGIS, ArcSDE are either registered trademarks or trademarks of Environmental Systems Research Institute, Inc.in the U.S.A. and/or other countries.

Author Information

Dr. Sahadeb De, Research Assistant Professor
Earth Sciences and Resources Institute,
University of South Carolina, Byrnes 401,
Columbia, SC 29208
Phone: 803-777-5911
FAX: 803-777-6437
E-mail: sde@esri.sc.edu
Web: http://www.esri.sc.edu/staff/de

Ms Manidipa Dasgupta, Research Assistant
Earth Sciences and Resources Institute,
University of South Carolina, Byrnes 401,
Columbia, SC 29208
Phone: 803 777 6364
E-mail: dasgupta@esri.sc.edu