

DATABASE REDUCTION TECHNIQUES FOR LARGE GIS DATABASES

Michael A. McAdams, Assistant Professor
Geosciences Department, Texas A&M University –Kingsville

and

Sonali Deshpande, Graduate Research Assistant
Computer Engineering, Texas A&M University-Kingsville

A GIS consists of a powerful set of tools for storing, retrieving, transforming and displaying spatial data. GIS technology has revolutionized the entire approach to spatial analysis. Furthermore, the technologies of electronic surveying, remote sensing, air photos, global positioning systems (GPS) and mobile computing/communications has created a powerful alliance with GIS for the improvement of spatial data gathering and analysis. Before GIS technology, spatial analysis was somewhat arduous and dependent upon the ability of the mapmaker. This limitation sometimes prevented extensive research to assist with decision-making and problem solving. However, with GIS, which is inherently linked to computer technology and database management, larger amount of spatial data can be processed quickly to develop numerous alternatives and solutions for better decision-making. One of the highest levels of the use of a GIS is for complex spatial modeling.

Due to the data needs required to adequately model complex physical and human spatial relationships many GIS modeling projects must be performed on large mainframes or supercomputers. Examples of these types of models would be those analyzing climate changes (i.e. El Niño), watershed drainage and degradation due to human activities, leaching of pollutants into sub-surfaces geology, traffic/person trip forecasting, population growth and distribution, employment forecasting and economic models. In many cases, modelers need only a portion or sub-set of the GIS databases or have computer which cannot process such large databases. In addition, the format of the data may be incompatible to be transported from a supercomputer to a mainframe of PC. Thus, there is a need to reduce the amount of GIS databases for users of the data generated by a supercomputer.

A major intrinsic characteristic of a GIS and essential for database reduction is the visualization of the geographic data and the associated non-spatial database. The new generation of visualization methods must cope with the growing data size as well as with data analysis (17.) All GIS programs and the associated models to be effective must be able to visualize geographic and tabular data or they cannot be classified as a GIS or GIS modeling. The visualization of GIS data on a supercomputer to determine the

means of reduction is crucial to delineate the methods of GIS data reduction. Once GIS database visualization and reduction programs are developed or adapted from existing programs such as GMS® or ArcInfo® for supercomputers, they will lead to more efficient and easier manipulation for users accessing the data by mainframe or personal computer.

The purpose of this preliminary literature review is to:

- 1) briefly introduce GIS and spatial modeling to those unfamiliar with the technology
- 2) explore the major methods of GIS database reduction, including database visualization;
- 3) identify possible new methods of GIS database reduction, such as expert systems, fuzzy logic, fractal analysis;
- 4) provide a preliminary outline of the possible integration of GIS and modeling with database reduction or sub-setting techniques such as HDF5; and
- 5) identify key articles, books, URL's for further exploration

The findings of this document are directly related to a project with *the U. S Army Corps of Engineers-Waterways Experimental Station* whose purpose is reduce hydrological spatial data sets so that they can be utilized in the mainframe and PC environment in conjunction in a ground water model (GMS.)

GIS DATABASES and MODELING:

Essentially, a GIS consists of interrelated geographical and attribute databases. The definition of a GIS varies according to the source. David Cowan, an expert in GIS in the Geography Department at the University of South Carolina states, " a Geographic Information System uses geographically referenced data as well as non-spatial data and includes operations which support spatial analysis"[\(3.\)](#) The USGS further states: "In the strictest sense, a GIS is a computer system capable of assembling, storing, manipulating, and displaying geographically referenced information, i.e. data identified according to their locations. Practitioners also regard the total GIS as including operating personnel and the data that go into the system."[\(27.\)](#)

As seen in the above definitions, A GIS is different from other information or database systems. When stored in a GIS, data is divided into geometric data and attribute data. The first refers to geometric aspects (location and dimensions) and has geometric information or topology. The second refers to other, non-geometric characteristics. [Kraak, Ormeling, 4]. An innate quality and defining aspect of a GIS from other database management or information management systems is that a GIS has geographic databases that are linked through unique ID's and database management procedures (relational, networked etc.) to tabular or attribute databases. The geographic databases are either raster or vector indicating the objects' (polygon, lines, points in vector database) or a pixels' (in a raster GIS) spatial address or coordinates (i.e., x, y; latitude, longitude, UTM etc.) (A good URL detailing and illustrating

vector versus raster GIS's can be found at ESRI's site (8.) In the vector model, information about points, node, lines, and polygons is encoded and stored as a collection of x, y or z coordinates. A raster model is used to model continuous features and is comprised of a collection of grid cells rather those found in the cathode ray tube of a monitor or television screen. In a vector GIS, the objects are defined in the database, such that there is information stored about the surrounding objects to conduct complex spatial operations. This geographic characteristic contained in a GIS is referred to as topology (1.) Although a raster GIS is important when discussing GIS, our discussion will primarily be directed toward vector GIS and primarily networks, as the project team will be working data reduction in relation to a network hydrological GIS model (GMS.) However, the structure of raster database could be important if one is interested in layering a remotely sensed image within a vector GIS.

A particularly unique feature of a GIS is that it can store information about the a collection of layers that can be linked together using coordinates and identifiers (See Figure 1 below.) This characteristic of a GIS allows for complex operations and modeling. In a vector GIS, these layers can contain point, lines (i.e., networks) or polygons.

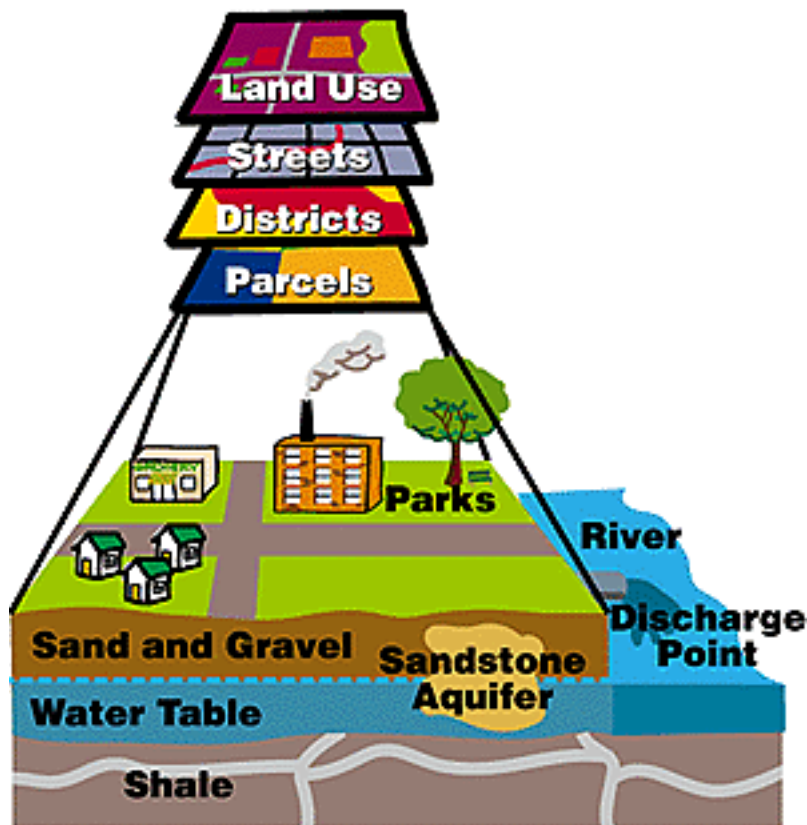


Figure 1

(1)

A GIS combined with mathematical modeling presents some unique situations which are not found within non-spatial modeling. A GIS can be used to display the results of a mathematical model or the mathematical model can be embedded into a GIS. (GMS is an example of a GIS with an embedded

hydrological model.) With a GIS which uses the results of a non-spatial mathematical model, the data resulting from running the model contain unique id's (numeric or alphanumeric) that can be imported into the GIS. Without these linkages the use of non-spatial mathematical models are difficult to use. In a GIS, which has, a mathematical model component linked to the spatial database, the ease of spatial modeling is greatly increased. However, there are further complications in how files are handled within the modeling process.

There are three types of GIS models: 1) dynamic; 2) explanatory, and 3) descriptive. The following presents a good description of these types of GIS models:

Dynamic models include *time* as an explicit element of the model, otherwise the model is *static*. In dynamic models, the *state* of the system at one time, plus the *driving forces*, follow definite *transformation relations* to reach the next state, and so on till the end of the simulation. This is sometimes called the *state-variable* approach.

Explanatory models attempt to explain how a system works, from some first principles. For example, crop growth based on photosynthetic reactions as influenced by temperature, light, vapor pressure etc.

Descriptive models simply attempt to characterize a system for predictive purposes, without pretending to explain. Statistical models are a subclass of these. It would seem that the best explanation would give the best description, in practice the purposes of the models are quite different. In land evaluation we generally are presented with descriptive models with pretensions to explanation. [\(22\)](#)

Although a GIS database can be placed on a supercomputer and reduced or placed in sub-sets with existing software programs such as HDF5 [\(25\)](#), the most effective means to determine the best means of GIS database reduction is via visualization packages incorporated with a GIS. Although there are non-spatially related tools related to database reduction, these functions would function more efficiently within an established GIS uploaded to a supercomputer or via a browser located on a PC or mainframe that view and manipulate the data located on a supercomputer.

METHODS OF GIS DATABASE REDUCTION:

There are several methods for GIS data reduction:

- 1) [GIS Database Visualization](#)
- 2) [Using algorithms to reduce and compress data](#)
- 3) [Querying](#)
- 4) [Buffering](#)
- 5) [Windowing](#)

Each of these methods works in different manner and could be used separately or in various combinations. It should be noted that other disciplines that might have a bearing on database reduction is discussed under ["Chaos, Artificial Intelligence, Fractals, Fuzzy Logic and Systems and Expert Systems."](#)

GIS Data Visualization

When data are entered into the computer, they are stored as files and referred collectively as database. In GIS database language, the items about which the information is gathered are referred to as entities (spatial) and attribute (descriptive) data. A basic difference between these types of information and the information that is collected into a Information Systems is that GIS information has associated with it an underlying geography, or geographic description of locations on the earth. This means a spatial database can be visualized as on-line maps or images and the associated data can be viewed as tables. This data visualization is unique to GIS and is one of its most powerful elements (26).

The spatial data for features may be stored as separate layers or as themes in a geographic database. The attribute data tells what the geographic database in a layer represent. The attribute data can be listed out and visualized in a tabular format. The attribute data is used to determine how specific features will be displayed, or it can be used to select and display specific features. Visualization can be applied in three different situations in the GIS environment. First visualization can be used to explore unknown and raw data. Secondly, visualization is applied in analysis in order to manipulate the unknown data and thirdly visualization is applied to present (e.g. to communicate knowledge of spatial information) Therefore, the aspect of visualization in a GIS is a dynamic process which is dependent upon the needs of the GIS operator.

Visualization transforms raw data into vivid 2D and 3D images which help the scientists reveal important features and trends in the data, convey ideas and communicate their finds. The sheer amount of data to be analyzed is overwhelming and there is not enough time to browse and visually inspect an extremely high-resolution database. According to Bernd Hamann (12), multi-resolution methods can be used to represent and visualize large databases at multiple levels of resolution and automatic methods can be used for extracting features and identifying regions characterized by "unusual" behavior.

Multi-resolution methods help in reducing the amount of time it takes to browse the domain over which a physical phenomenon has been measured or simulated, while automatic feature extraction methods assist in steering the visualization process to those regions in space where a certain interesting behavior has been identified. In GIS terminology, browsing the geographic data is referred to as "zooming" or "panning." In other words, you can view the geographic data at different scales and different areas within the scales. This feature also is peculiar to a GIS particularly when tied to the display of various layers and thematic scenarios.

Although, the new generation of massively parallel computers will have high speeds and handle large databases, their effectiveness depends upon the ability of human experts to interact with their computations and extract useful information. 3D interaction technologies are more common as the humans can interact naturally in a 3D world and the data in important computational problems has a fundamental 3D spatial component. The Center for scientific computing and imaging at University of Utah has developed a problem-solving environment for steering large-scale simulations with integrated interactive visualization called SCIRun. It is a scientific programming environment that allows the interactive construction, debugging and steering of large-scale scientific computations. SCIRun enables to modify geometric models and interactively change numerical parameters and boundary conditions, as well as to modify the level of mesh adaptation needed for an accurate numerical solution.

The visualization of GIS data is key in the process of data reduction. Without this first step, data reduction in a GIS can be preformed only in the crudest of fashions and with great difficulty. While the data may be stored on a supercomputer, the inability to visualize the geographic and the associated attribute data will significantly limit a user that desires to download this information for use in a mainframe or a PC. Further data reduction techniques can be preformed, but they are dependent on the users ability to visualize the GIS data. Database reduction could be done manually once the GIS database is visualized. Geographic and attribute information could be selected and then downloaded to a mainframe or a PC. (The more directed forms of GIS database reduction such as querying and buffering will be discussed in later sections.) The existing database sub-setting programs, such as HDF5, do not have a GIS data visualization component. The concept of maps and data served off the Internet is in its nescient stage and offers great promise to visualize geographic and attribute data. The work that it being presently initiated by ESRI would seem to be a guide for GIS visualization that this project may consider [immolating \(9.\)](#)

Algorithms to Reduce and Compress Data

Once the GIS data is visualized various techniques can be implemented to reduce and compress the data while on supercomputer. This may consist of reducing the amount of polygons, lines, and nodes that are superfluous when the data is downloaded to a mainframe or a PC. When looking at a GIS database, there could be geographically linked data which make very little difference in the operation of a model at a lower level. For example, if there were links in a network model and associated nodes, the combining of the links one and the nodes may have minimal or no effect on the display, operation and manipulation of the data when transported to another platform. This action can not be easily preformed manually in a GIS. In GIS, since there are associated attribute data with the geographic data, these must also be deleted. However, when the links or nodes are numerous this would be a tedious tasks if there was a data visualization package located on the supercomputer. Algorithms to perform this task globally are absolutely necessary. In addition, there has to be processing to reorder the formatting of the geographic and attribute data.

The most promising technique for database reduction is the use of neural networks.

The following is a definition of a neural network and how it functions:

Neural Networks use a set of processing elements (or nodes) loosely analogous to neurons in the brain. (Hence the name, neural networks.) These nodes are interconnected in a network that can then identify patterns in data as it is exposed to the data. In a sense, the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs, that simply follow instructions in a fixed sequential order. [\(28\)](#)

The operation of neural networks are illustrated and in the below figure and quote:

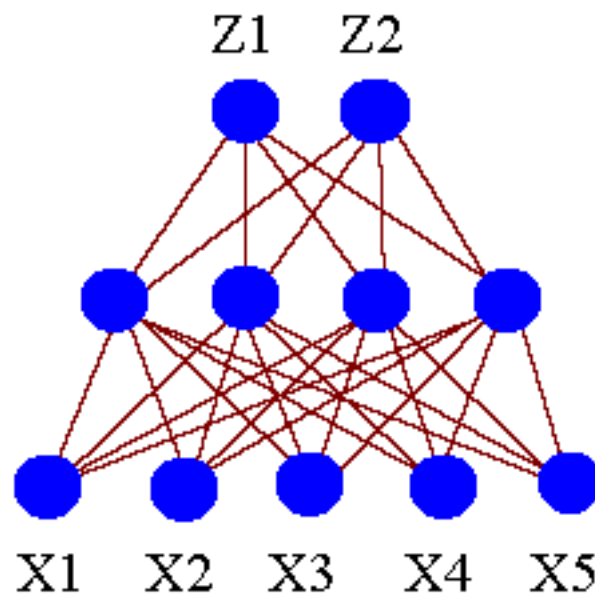


Figure 2

The bottom layer represents the input layer, in this case with 5 inputs labeled X1 through X5. In the middle is something called the hidden layer, with a variable number of nodes. It is the hidden layer that performs much of the work of the network. The output layer in this case has two nodes, Z1 and Z2 representing output values we are trying to determine from the inputs. For example, we may be trying to predict sales (output) based on past sales, price and season (input). [\(28\)](#)

As will be discussed later, neural networks could also be combined with other program and knowledge or expert based systems to perform other duties associated with the processing and reduction of this data.

As GIS relies heavily on the transmission of images for the geographic and attribute database, means of image reduction are pertinent to discuss. In databases used for scientific applications, losing data when an image compressed is unacceptable. There are many techniques for "loss less" image compression. In

sub band coding , an image is filtered to create a set of images, each of which contains a limited range of spatial frequencies. These sub bands are then downsampled, "quantized" and coded. Vector quantized is another technique, which achieves very high compression ratios. A vector "quantizer" is a system for mapping a stream of very high rate or volume discrete data into a sequence of low volume and rate data suitable for storage in memory and communication over digital channel . However, both these techniques require much computation propose a neural network methodology to compress/decompress images (10) . This methodology has a compression ratio of 16:1. One of the common neural approaches in image compression is to train a network to encode and decode the input data so that the resulting difference between input and output is minimized. The network consists of the input layer, output layer and an intermediate layer of smaller size. The ratio of the size of the input layer to the size of the intermediate layer is the compression ratio. The image is subdivided into equal-sized blocks and compression is carried out block by block. This reduces the network learning time. Each block is scaled from an integer with a maximum value of 255 to a real number with a maximum value of 1.0. These scaled blocks are used as the inputs to the neural network. The network is trained to minimize the squared error between the output and input values, thus maximizing the signal-to-noise ratio. The neural networks are trained using the 'learning algorithm' described by (10). The means to compress and decompress images might be useful in the visualization of geographic and tabular data. In addition, since the image compression algorithms are concerning the transmission of data some of these techniques might be valid for data compression and decompression.

Although it is unknown how these different programs will be effective on a supercomputer for use in GIS database reduction, they hold promise at a means to "automate" the process. Neural networks are now a well developed sub-discipline of database management and research . The issue of image compression, while still in the exploratory stage, appear to be pertinent to the main subject of this document. Although not directly related to database reduction is the ability of the program housed on the supercomputer to describe the data in terms of its characteristics (i.e. mean, mode, variation etc.) The following methods of querying, buffering and windowing as a means of database reduction exist within the context a GIS, but have not been applied in a supercomputer environment. Also, the querying, buffering and windowing are dependent on GIS database visualization.

Querying:

Information stored in the GIS database has to be retrieved for various purposes. Querying is one method of retrieving the required data. This resulting data from a query can be downloaded as an associated geographic related sub-sets. One of the most common languages for querying geographic related data is the Spatial Query Language (SQL). According to Egenhofer (7), spatial SQL consists of two components: a query language to describe what information to retrieve and a presentation language to specify how to query results. Spatial SQL is based on relational database query language. An SQL query uses the SELECT-FROM-WHERE clause for the three operations of relational algebra projection, Cartesian product and selection respectively. Aggregate functions such as sum, minimum, average may be used to calculate a single value. The SQL-based query language for GIS is a combination of syntax extension in the SELECT-FROM-WHERE clauses and a command set outside of SQL. The spatial relationships distance, overlay and adjacent are added to the WHERE clauses and spatial relations are

extended by the attributes area, perimeter and length. Interactive communication with drawings is enabled with a PICK qualifier, which allows the user to formulate queries with reference to spatial objects visible on the screen.

The Graphical Presentation language (GPL) provides tools for the manipulation of the graphical presentation of query results. Display environment is a concept, which handles the information about how to display the query result. During query processing, this information is integrated with user query so that the result is rendered according to the display description. There are six types of graphical specifications: 1) Display mode 2) Visual variables 3) Scale of drawing 4) window to be shown 5) Spatial context 6) examination of the content.

Another approach to querying GIS databases is perspective querying (14). As the data is in the hierarchical structure, spatial queries are also organized in a hierarchical way. Searching for an object in a 3D space means traversing the R-tree. R-trees are one of the most important data structures for indexing spatial data. R-trees consist of overlapping rectangles that contain geometric objects. In the 3D model, R-trees can be used to store 3D bounding boxes. As the whole tree is based on bounding boxes, a few small queries in each level are sufficient to find objects in the 3D space. In the perspective querying method, the pyramid of vision is split into axially parallel boxes. In this way, queries for objects within the resulting boxes can be done much faster. In addition, it is possible to process queries for different boxes in different ways. The first query would only consider the box nearest to the point of view. The objects for boxes far away from the point of view can also be queried. .

Buffering/Overlay:

Buffering is another method of data retrieval/reduction that allows the user to select features according to their proximity to a point, a line or an area. Buffering is a matter of measuring a distance in any direction from another object, whether a point, a line or other polygon (5). There may even be a requirement for a buffer around a second buffer around still another buffer to produce doughnut buffer. There are various methods of buffering. Some buffers are designed to indicate that around a given entity, lays a region that needs to be protected, studied, guarded or require special treatment. These are known as arbitrary buffers. A buffer can also be based on intervisibility measures. This is a measurable buffer. Thus, the buffer is selected based on definable, measurable value. Another method of buffering is mandated buffering. Mandated buffers are buffers whose distances are dictated by legal mandate (e.g. frontage along homes dictated by zoning ordinances). Buffers are useful methods of data retrieval. The fundamental problem with buffers is that they frequently require knowing more about the interactions of landscape's elements than the user does.

Map overlay consists of manually overlaying transparent “maps” such that one can see different thematic layers in the same geographic extent (6.) This process involves drawing maps of different thematic information on transparent media at the same scale and over the same geographic extent. Once the maps are drawn, they can be laid on top of one another in various combinations to examine the spatial relationships between different themes. In addition to manual map overlay, there is a computerized map overlay process, which allows users to perform other analytical operations. This is known as computational overlays. A typical computational overlay is the creation of a buffer, described above, around a certain area. This kind of operation takes as its input, the layer containing the spatial information and a distance parameter specifying the size of the buffer zone. It returns a layer specifying the spatial extent of the area within the specified buffer distance. This method of access of geographic information is exploratory because: 1) users need only little knowledge about particular values of data 2) users examine visually the integration of different kinds of spatial data, looking for interesting patterns, neighborhoods or coincidence.

Windowing:

Spatial and attribute data can also be retrieved from a GIS database by selecting a portion of the geographic and attribute database by "drawing a square, rectangle or polygon around the area would like to "cut-out." The window can be determined either by two pairs of coordinates, two diagonal points selected on a screen drawing, or the minimal bounding rectangles from the result of a Spatial SQL query. This is a simple operation, but very effective in selecting a portion or "slice of the GIS database.

[Chaos, Artificial Intelligence, Fractals,](#)

[Fuzzy Logic/Mathmatics, and Expert Systems](#)

The areas of chaos theory (also know as the theories of complexity and simplicity), artificial intelligence fractal analysis, fuzzy logic and fuzzy system and expert systems deserve mention in relation to GIS database reduction. Chaos and the related theories of complexity and simplicity could be perceived as paradigm shifts which may be in direct opposition to linear mathematical modeling. While the purpose of this document is not to explain fully all these elements, there will be an attempt to introduce them and how they may impact GIS database reduction and modeling. Although one might think that these are stretching the purpose of this project, all of the above mentioned areas are fully developed disciplines which are on the forefront of exploring alternative methods to analyze complex processes.

Chaos theory seems to be the overall umbrella discipline to artificial intelligence, fractal analysis, and fuzzy logic/systems. Chaos examines the overall phenomena and attempts to let the data "speak for itself" instead of imposing "Laws" to the modeling process. Chaos theory can be described as such "A

scientific discipline which is based on the study of nonlinear systems. The terms complexity theory and complex systems theory describe the theory more adequately, however, chaos theory is more widely accepted" (21.) The concept of groundwater modeling is certainly a complex system. While not attempting to go to another direction, there is a question of whether all modeling will be subject to increased scrutiny as to its clumsiness in light of chaos theory. Some of the complexity of the data could be due to the inadequacy of the present Newtonian paradigm to model complex systems. However, it is not really the purpose of this paper to question the present model or linear mathematical modeling. The complexity and simplicity concepts of chaos could be used in database reduction. In natural phenomena, many things develop from a simple actions. For example, a tree can grow from a seed because it has implanted in it the essentially elements to direct the cells to multiply, specialize and grow into a tree. In the same light, DNA in a human embryo can grow into a human. The idea that would be applicable to database reduction would be the possible reduction of the essence of the data so that it could be "reproduced" at a lower computing level (i.e., mainframe, PC.) However, there would have to be some sort of reproduction program at the lower level to understand the instructions that were being given at a higher level or supercomputer.

Artificial intelligence, while fairly new, might have some applicability in database reduction. Artificial intelligence (AI) is the idea of the "learning computer. " Neural networks, as discussed earlier, are a part of the discipline of AI. A sub-set of AI that could be adaptable to database reduction is "automated reasoning." Automated reasoning is explained in the following text:

To understand what automated reasoning is, we must first understand what reasoning is. *Reasoning* is the process of drawing conclusions from facts. For the reasoning to be sound, these conclusions must follow inevitably from the facts from which they are drawn. In other words, reasoning [...] is *not* concerned with some conclusion that has a good chance of being true when the facts are true. Indeed, reasoning as used here refers to logical reasoning, not of common-sense reasoning or probabilistic reasoning. The only conclusions that are acceptable are those that follow *logically* from the supplied facts. The object of *automated reasoning* is to write computer programs that assist in solving problems and in answering questions requiring reasoning. The assistance provided by an automated reasoning program is available in two different modes. You can use such a program in an iterative fashion; that is, you can instruct it to draw some conclusions and present them to you, and then, based on your analysis of the conclusions, it can in the next run execute your new set of instructions. Or you can use such a program in a batch mode; that is, you can assign it an entire reasoning task and await the final result (15)

Automated reasoning could be applied to exploring the database or in the querying process. In addition, using automated reasoning, errors in database extraction could be avoided.

Fractal Analysis is a very broad and robust discipline that is now being used for a variety of functions. It is used for repetitive functions and for understanding complex processes: The following gives the reader some understanding of the use of fractals:

For the most part, when the word *fractal* is mentioned, you immediately think of the stunning pictures you have seen that

were called fractals. But just what exactly is a fractal? Basically, it is a rough geometric figure that has two properties: First, most magnified images of fractals are essentially indistinguishable from the unmagnified version. This property of invariance under a change of scale is called self-similarity. Second, fractals have fractal dimensions, as were described above. The word fractal was invented by Benoit Mandelbrot, "I coined fractal from the Latin adjective fractus. The corresponding Latin verb frangere means to break to create irregular fragments. It is therefore sensible and how appropriate for our needs! - that, in addition to fragmented, fractus should also mean irregular, both meanings being preserved in fragment. [\(21\)](#)

Fractals can be used to investigate and replicate complex processes such as the growth of an organism or even a physical or human phenomena that have some regularities. Fractals have been used to examine the growth of cities and dynamic organisms. In fractal analysis, the researcher is looking at the process alone and not analyzing the underlying assumptions. The images are converted into fractals and mathematical equations then associated with the growth of the fractals.

Fuzzy logic and systems is a fully developed sub-discipline of mathematics. Instead of using absolutes for equations, one uses a range of numbers. This is also called "fuzzy mathematics." One definition is that: "Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth -- truth values between 'completely true' and 'completely false' [\(20\)](#)." These are a set of tools that could be used in combination with database reduction techniques. One is often not looking at an absolute set, but a range of sets. In addition, one may want to do operations with these sets. Fuzzy mathematics or logic also works well with logical expressions similar to those used in querying GIS databases. An example, if a researcher was looking asking the GIS to give the property of all "low income" persons, this could be considered a "fuzzy set."

Expert systems may be applicable when examining database reduction techniques.

Expert systems are a set of instructions based on the knowledge of the workings of a particular phenomena and be defined as such:

An expert system tool, or shell, is a software development environment containing the basic components of expert systems. Associated with a shell is a prescribed method for building applications by configuring and instantiating these components. Some of the generic components of a shell are shown in (the below figure) and described below. The core components of expert systems are the knowledge base and the reasoning engine. [\(19\)](#)

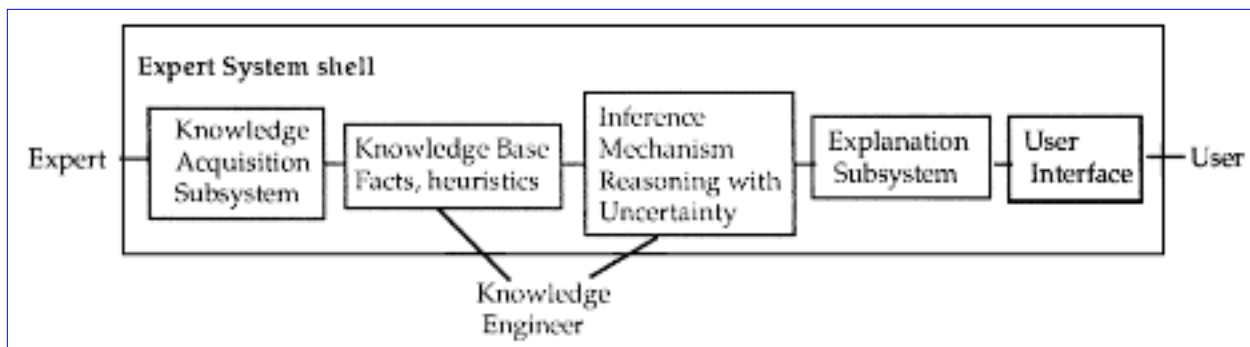


Figure 3

In another source, it is stated:

Conventional programming languages, such as FORTRAN and C, are designed and optimized for the procedural manipulation of data (such as numbers and arrays). Humans, however, often solve complex problems using very abstract, symbolic approaches which are not well suited for implementation in conventional languages. Although abstract information can be modeled in these languages, considerable programming effort is required to transform the information to a format usable with procedural programming paradigms [\(11\)](#)

Expert Systems could be considered a set of programs to accomplish certain tasks, such as GIS data reduction. It could also encompass such tools as fuzzy logic, neural networks and a variety of other AI tools.

It should be noted that chaos, including complexity and simplicity, fractal analysis, artificial intelligence, fuzzy logic/systems and expert systems are presented here as possible areas to explore for GIS data reduction techniques. The purpose of this report is to introduce these areas and not to go into full explanation of their viability for use in database reduction. This will be developed at a latter date.

DISCUSSION AND PROPOSED FRAMEWORK

This document is a preliminary view into to the subject of database reduction. However the authors believe that they have shown some of the basis issues in GIS database reduction. The issue of database reduction and visualization are extremely broad areas. However, when GIS database reduction and modeling is introduced these areas become even more complex. The areas mentioned in the last section are robust, but appear to not have been applied to database reduction and certainly not in the context of database reduction on a supercomputer and transfer to a mainframe or PC. The authors have only given here a cursory look at GIS and modeling and the subsequent methods of database reduction.

From viewing the literature and from the authors knowledge, it would appear that the process of database reduction would be as follows:

- 1) Processing of GIS data would be performed on a supercomputer;
- 2) A GIS program (i.e, GMS, ArcInfo, ArcView) would visualize geographic and attribute data while located on the mainframe.

- 3) Appropriate software (i.e. HDF5) would work with GIS database reduction techniques such as querying, buffering, neural network algorithms, fuzzy querying, windowing within the provided GIS visualization tools.. Expert or knowledge based systems may give further robustness to the tools..
- 4) The selected GIS data would be placed in sub-sets for downloading to a mainframe or PC. Before it was downloaded, the supercomputer would communicate with the mainframe or the PC to determine if there was enough hard drive space. If there was enough, the computer would download the database in appropriate levels to be processed by the mainframe or PC.
- 5) The selected GIS data would be downloaded and translated into a form that could be imported so that the selected software could process (e.g., GMS.)

In the next phase of the project, the research team at Texas A&M University-Kingsville will be working with various experts to determine the suitability of the methods of database reduction that have been outline in this document.

Bibliography

- 1) Buckley, David J., "Fundamental Concepts" in *The GIS Primer*: , <http://felix.geog.mcgill.ca/courses/geo201/primer/concepts.html>, (January 1999).
- 2) Clarke, Keith C., *Getting Started with Geographic Information Systems*, 2nd ed. (New Jersey: Prentice-Hall, Inc, 1999), 34.
- 3) Cowan, David, "UNIT 1 - WHAT IS GIS?", <http://www.geog.ubc.ca/courses/klink/gis.notes/ncgia/u01.html#SEC1.1.2> (30 August 1997).
- 4) Cramer, Christopher E, "Compression of Still and Moving Images", http://www.ee.duke.edu/~cec/research/neural_compression/still.html (15 March, 2000).
- 5) Demers Michael N., *Fundamentals of geographic Information Systems*, 2nd ed. (New York: Wiley, 1999), 111.
- 6) Egenhofer Max J. and Richards James R, "Exploratory Access to Geographic Data Based on the Map Overlay Metaphor", *Journal of visual Languages and Computing*, 4, no.2 (1993):105-125, <http://www.spatial.maine.edu/~max/RJ9.html> (15 March, 2000).
- 7) Egenhofer, Max J, "Spatial SQL: A query and Presentation Language", *IEEE Transactions on Knowledge and Data Engineering*, 6, no.1 (1994): 86-96, <http://www.spatial.maine.edu/~max/RJ14.html> (8 March, 2000).
- 8) ESRI, "About GIS", http://www.esri.com/library/gis/abtgis/gis_wrk.html (29 February, 2000).
- 9) ESRI, "Internet mapping is a powerful communication tool", <http://www.esri.com/software/internetmaps/netmapping.html> (14 October 1999).
- 10) Gelenbe E., "Learning in the Recurrent Random Neural Network", *Neural Computation*, 5, no. 1, (1993):154-164.
- 11) GHG Corporation, "What are Expert Systems?" <http://www.ghgcorp.com/clips/ExpertSystems.html> (7 January 1997).
- 12) Hamann, Bernd , "Visualizing Large Scale Databases: Challenges and Opportunities," http://www.icase.edu/~kma/s99_panel.html (19 January 2000),

- 13) Keim, Daniel A, Herrmann Annemaria, "The Gridfit Algorithm: An Efficient and Effective Approach to Visualizing Large Amounts of Spatial Data", <http://www.informatik.uni-halle.de/~keim/VisualPoints/index.html> (11 March,2000).
- 14) Kofler, Michael, "3D spatial access, perspective querying", <http://www.icg.tu-graz.ac.at/isprs96/node9.html> (10 March, 2000).
- 15) Michael Kohlhase and Carolyn Talcott. "What is Automated Reasoning", <http://www-formal.stanford.edu/clt/ARS/ars-db.html>, (2 September 1996).
- 16) Kraak M. J. and Ormeling F.J., *Cartography: Visualization of Spatial Data*, (Essex: Longman, 1996), 4.
- 17) Ma, Kwan-Liu, "Visualizing Large-Scale Datasets: Challenges and Opportunities", Siggraph '99, http://www.icase.edu/~kma/s99_panel.html (19 January 2000).
- 18) Kuijpers, Bart, Paredaens Jan and Vandeurzen Luc, "Semantics in Spatial Databases", <http://win-www.uia.ac.be/u/kuijpers/semantics.html> (15 March, 2000).
- 19) Japanese Technology Evaluation Center, "Expert Systems Building Tools: Definitions", http://itri.loyola.edu/kb/c3_s2.htm, (May 1993).
- 20) Pacific Northwest Laboratory, "What is Fuzzy Logic", <http://www.emsl.pnl.gov:2080/proj/neuron/fuzzy/what.html>, (13 March 2000).
- 21) Quentmeyer, Tyeler, "Chaos Theory, Dynamic Systems, And Fractal Geometry", <http://library.thinkquest.org/3493/> (Unknown Date).
- 22) Rossiter, David G. "Lecture Notes: Part 3, <http://www.sbg.ac.at/geo/idrisi/landeval/s494ch3p.htm#3.3> (18 October 1995)
- 23) . Shekhar Shashi and Chawla Sanjay, "Spatial Databases: Concepts, Implementation and Trends", http://www.du.edu/~shick/gis4501/notes/data_mod.htm (8 March, 2000).
- 24) Eick Stephen, Hamann Bernd, Heermann Philip, Johnson Christopher and Krogh Mike, "Visualizing Large-Scale Datasets: Challenges and Opportunities", Siggraph '99, http://www.icase.edu/~kma/s99_panel.html (19 January, 2000).
- 25) University Computing Services, "HDF Hierarchical Data Format", <http://www.ucs.uwa.edu.au/sw/pp/HDF.html> (16 January, 2000).
- 26) Unknown Author, "Lesson 2: Introducing GIS", <http://archi.kyungpook.ac.kr/~www/introai/basicai/Lesson02/content/overview.cfm> (Unknown Date).
- 27) USGS, "Geographic Information Systems", <http://www.usgs.gov/research/gis/title.html> (1 July 1997).
- 28) Z Solutions, "Light Description of Neural Networks", <http://www.zsolutions.com/light.htm> (1999)