Decision Support Using Data Mining Methods and Geographic Information

Peter Bajcsy, Peter Groves, Sunayana Saha, and Tyler Alumbaugh
Automated Learning Group (ALG), National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign (UIUC)

ABSTRACT

We present a decision support system using data mining methods and geographic information. The geographic information consists of heterogeneous, raster and vector, data types. While raster data types represent grid-based information collected by camera sensors, e.g., satellite images, vector data types are used for representing boundary information, for example, man-made or naturally defined regions, or point information, for instance, building locations, customer's permanent locations or measurement locations. The occurrence of heterogeneous geographic data types and the information associated with raster and vector data pose challenges to data analyses supporting decision makers in environmental preservation and development planning domains. In this work, we focus on the data analysis problem related to forming geographic regions as aggregations of basic boundaries under a set of decision support constraints. This problem is defined as a search for the best partition of any geographical area that is (a) based on raster or point information, (b) formed by aggregations of known boundaries, (c) constrained by spatial locations of know boundaries and (d) minimizing an error metric. We present our overall approach and describe the proposed optimization processes to the sub-problems, such as, (1) data representation of heterogeneous data types, (2) feature selection and extraction from heterogeneous data, (3) aggregation of boundaries into geographic partitions based on feature similarity, (4) error evaluation of multiple geographic partitions and (5) visualization of heterogeneous data within a geographical context. The proposed system is illustrated with elevation and forest label raster data, US Census Bureau boundary data and FBI crime reports for making decisions about police force deployment.

Keywords: Mining geospatial data, decision support, geospatial clustering.

1 INTRODUCTION

In general, there are many applications of data mining that require incorporating heterogeneous geographic information into analysis, for instance, farmer's yield assessment, water quality analysis, merchandise store distribution, crime or disease control [1], [15], [20], [21]. In one particular case, designing a decision system for police force deployment would require data fusion of historical records about crime rates and criminals, with density of population and available police force, and geographical descriptors about land cover, land use, elevation and so on. Information fusion would be followed by formation of regions with geographic constraints using data mining techniques, then error evaluation of geographic partitions and visualization of the geographic results for the final decision about police force deployment. There are other application scenarios that would involve similar data analysis steps and would lead to geographic partitions derived from, just to name a few, farmer's crop yields or West Nile virus occurrences. We present our overall approach and describe the proposed optimization processes to the sub-problems, such as, (1) data representation of heterogeneous data types, (2) feature selection and extraction from heterogeneous data, (3) aggregation of boundaries into geographic partitions based on feature similarity, (4) error evaluation of multiple geographic partitions and (5) visualization of heterogeneous data within a geographical context.

The problem of geographic partitioning based on typical heterogeneous data types in Geographic Information Systems (GIS), such as, raster and vector data types, is described as follows. Search for the best partition of any geographical area such that the geographic partitions are (a) derived from raster and/or point information, (b) formed by aggregations of known boundaries, (c) constrained or unconstrained by spatial locations of know boundaries and (d) minimizing an error metric. While raster data types represent grid-based information collected by camera sensors, e.g., satellite images, vector data types are used for representing boundary information, for example, man-made or naturally defined regions, or point information, for instance, building locations, customer's permanent locations or measurement locations. The occurrence of heterogeneous geographic data types and the information associated with raster and vector data pose challenges to data analyses supporting decision makers in environmental preservation and development planning domains. In this work, we focus on the data analysis problem related to forming geographic regions as aggregations of basic boundaries under a set of decision support constraints.

Although there are currently available a few Geographic Information Systems (GIS) software packages [3], [24] that can load, georeference and visualize some geographic information, we have not found a suite of software tools that would support all heterogeneous inputs, incorporate several data mining approaches and decision support constraints, and meet various geographic evaluation requirements. The motivation for our work came from the need to meet all application specified requirements (file format support, information fusion, variety of feature extractions, spatially constrained and unconstrained boundary aggregations and error evaluation) and the lack of currently available GIS tools to meet these requirements.

We have approached the problem of geographic partitioning by decomposing the proposed decision support system into six basic components shown in Figure 1. These six components consist of five processing blocks, such as, (1) input information extraction and representation, (2) georeferencing data sets, (3) raster information extraction over defined boundaries, (4) feature driven boundary aggregation and (5) evaluation and decision making, and one visualization block that supports input and output data presentation to the decision making personnel.

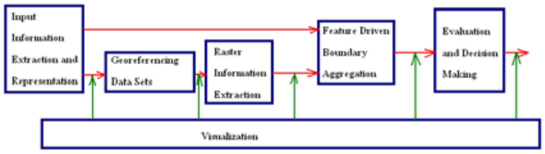

Input Information Extraction and Representation
Georeferencing Data Sets
Raster Information Extraction
Feature Driven Boundary Aggregation
Evaluation and Decision Making
Visualization

Figure 1: The top-level schema of the proposed system for geographic partitioning.

The proposed system for geographic partition was experimentally tested with the data sets including (1) raster data with continuous (digital terrain elevation) and categorical (forest label) variables, (2) vector data with geographical definitions of boundaries (US Census Bureau TIGER files and ESRI Shapefiles), and (3) tabular data defined per geographical boundary (FBI crime reports).

In general, in order to build a system for geographic partitioning, one has to resolve several issues related to (1) the tradeoffs between efficiency of information retrieval and data storage, (2) georeferencing raster and vector data, (3) feature extraction from raster data over georeferenced boundaries, (4) aggregation of boundaries into geographic partitions based on feature similarity, (5) error evaluation and comparison of multiple geographic partitions and (6) visualization of raster and vector data, as well as, features and geographic partitions. The next sections will describe the above issues in more detail.

2 INPUT INFORMATION EXTRACTION AND REPRESENTATION

The overview of pre-processing functionalities and formation of output data types is presented in Figure 2 and corresponds to the "input information extraction and representation" component in Figure 1. While many of pre-processing functionalities are self-explanatory based on Figure 2, we devoted extra attention to the problem of efficient information retrieval and data storage in Section 2.2. This problem also requires understanding of the underlying information that each data source conveys to the end-user as described in Section 2.1, particularly, the information in (1) Census 2000 Tiger/Line files defined by the U.S. Census Bureau and saved in TIGER file data structures and (2) shape files defined by the Environmental Systems Research Institute (ESRI) and stored in a LLS data structure.

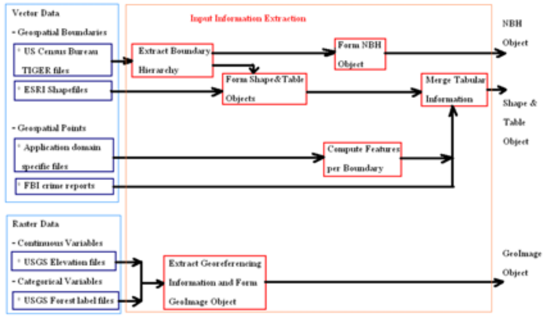


Figure 2: An overview of input information extraction and representation component in Figure 1.

2.1 Raster and Vector Data Information

There are two basic data structures used in Geographic Information System (GIS). They are raster (or cellular) data and vector (or polygon) data structures [1, Chapter 15]. The use of both data structures in GIS applications is basically driven by memory efficient representations of a variety of GIS information that can be classified into a raster and vector based information. Examples of raster-based information would include terrain maps, land use maps, land cover maps or weather maps. Boundaries (or contours or outlines) of parcels, eco-regions, Census tracts or US postal zip codes would be examples of vector data.

2.1.1 TIGER Files

The Census 2000 Tiger/Line Files provide geographical information on the boundaries of counties, zip codes, voting districts, and a geographic hierarchy of census relevant territories, e.g., census tracts, blocks or block groups [2]. The boundaries also form a geographic hierarchy that is encoded in Tiger/Line files, e.g., census blocks form census tracts. It also contains information on roads, rivers, landmarks, airports, etc, including both latitude/longitude coordinates and corresponding addresses [2]. A detailed digital map of the United States, including the ability to look up addresses, could therefore be created through processing of the Tiger/Line files.

Because the density of data in the Tiger/Line files comes at the price of a complex encoding, extracting all available information from Tiger/Line files is a major task. In this work, our focus is primarily on extracting boundary information of regions and hence other available information in Tiger/Line files is not described here.

Tiger/Line files are based on an elaboration of the chain file structure (CFS) [1], where the primary element of information is an edge. Each edge has a unique ID number (Tiger/Line ID or TLID) and is defined by two end points. In addition, each edge then has polygons associated with its left and right sides, which in turn are associated with a county, zip code, census tract, etc. The edge is also associated with a set of shape points, which provide the actual form an edge takes. The use of shape points allows for fewer polygons to be stored.

2.1.2 ESRI Shapefile

A shapefile is a special data file format that stores non-topological geometry and attribute information for the spatial features in a data set. The geometry for a feature is stored as a shape comprising a set of vector coordinates in a location list data structure (LLS). Shapefiles can support point, line, and area features. Area features are represented as closed loop polygons
A shapefile must strictly conform to the ESRI (Environmental Systems Research Institute) specifications [4]. It consists of a main file, an index file, and a dBASE table. The main file is a direct access, variable-record-length file in which each record describes a shape with a list of its vertices. In the index file, each record contains the offset of the corresponding main file record from the beginning of the main file. The dBASE table contains feature attributes with one record per feature. The one-to-one relationship between geometry and attributes is based on record number. Attribute records in the dBASE file must be in the same order as records in the main file.

2.2 The Tradeoffs Between Efficiency Information Retrieval And Data Storage

While the choice of raster type representation was exclusively driven by the efficiency of image analysis algorithms, our selection of the data structure was based on evaluations of the tradeoffs between efficiency of information retrieval and data storage.

Our evaluation of the tradeoffs between efficiency of information retrieval and data storage is based on a comparison of TIGER and Shapefile internal data organizations. Shapefiles do not have the processing overhead of a topological data structure such as TIGER files. They have advantages over other data sources, such as faster drawing speed and edit ability.

Shapefiles handle single features that overlap or are noncontiguous. They also typically are easier to read and write.

The TIGER file internal organization has significant advantages over other organizations in terms of storage space, but requires much processing before it can be retrieved. It is apparent that many boundaries will share the same border edges. These boundaries belong to not only neighboring regions of the same type, but also to different kinds of regions in the geographic hierarchy. A consequence of this is that storing the data contained in the Tiger/Line files in a basic location list data structure (LLS), where every boundary stores its own latitude/longitude points, would introduce a significant amount of redundancy to an already restrictively large data set.

In contrary to apparent storage efficiency, the TIGER vector data representation is very inefficient for boundary information retrieval and requires extensive processing. From a retrieval standpoint, an efficient representation would enable direct recovery of the entire boundary of a region as a list of consecutive points. The conversion between the memory efficient (concise) and retrieval efficient forms of the data is quite laborious in terms of both software development and computation time.

We have designed a new data structure called Hboundary to accommodate an efficient tradeoff between information retrieval and memory usage when a set of polygons/boundaries provide complete, non-overlapping coverage of a region (such as counties covering a state), and/or there are multiple types of boundaries that fall into a hierarchy (such as all state boundaries are also county boundaries, and all county boundaries are also Census Tract boundaries). The efficiency is accomplished by having one master list of boundary points that all boundaries reference by pointers. Given that in the Java programming language, integers, which are used as the pointers, are 32 bits, and double precision floating point numbers, two of which are used for each geographical point (latitude and longitude), are 64 bits, the memory savings for each point that is shared by two county, two census tract, and two block group boundaries is 30 bytes. For the state of Illinois, this optimization translated into a 38% reduction in memory usage, with HBoundary objects requiring 16.45 MB of memory to store its geographic point information while a representation that stores a copy of every point for every boundary would use 26.64 MB of memory.

2.3 Information Representation

Based on our previous considerations and studies, we have chosen BIP representation of raster data type and denoted the data structure as GeoImage Object in Figure 2 and including all georeferencing information associated with the raster data. All extracted information from TIGER files is stored in a novel HBoundary Object and is accessible to the system for additional information retrieval based on user's needs. For the further processing of boundary information and visualization, the HBoundary Object is converted into a location list data structure (LLC) called Shape Object to maximize information access efficiency. Thus, Shape Object and Table Object in Figure 2 denote the data structures that represent geographical boundary information (Shape) and associated features per boundary (Table). In order to represent spatial constraints derived from proximity of boundaries, we formed a NBH Object that contains a list of spatial neighbors with variable length for each boundary and was extracted from the oriented edge list in TIGER files.

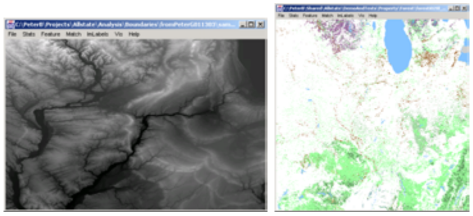


Figure 3: Examples of raster files, elevation (left) and forest labels (right), represented by GeoImage Object.

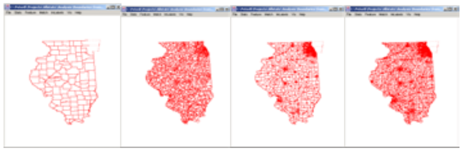
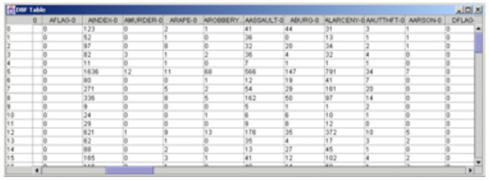


Figure 4: Extracted geographical boundaries of Illinois counties (left), zip code tabulated areas or ZCTAs (second left), census tracts (second right) and census block groups (right) from the US Census Bureau TIGER files. The visualization here shows the information represented by Shape Object.



DBF Table

	0	AFLAG-0	AINDEX-0	AMURDER-0	ARAPE-0	AROBBERY...	AASSAULT-0	ABURG-0	ALARCENY-0	AAUTTHFT-0	AARSON-0	DFLAG-
0		0	123	0	2	1	41	44	31	3	1	0
1		0	52	0	1	0	36	0	13	1	1	0
2		0	97	0	8	0	32	20	34	2	1	0
3		0	82	3	1	2	36	4	32	4	0	0
4		0	11	0	1	0	7	1	1	1	0	0
5		0	1636	12	11	68	566	147	791	34	7	0
6		0	80	0	0	1	12	19	41	7	0	0
7		0	271	0	5	2	54	29	161	20	0	0
8		0	336	0	8	5	162	50	97	14	0	0
9		0	9	0	0	0	5	1	1	2	0	0
10		0	24	0	0	1	6	6	10	1	0	0
11		0	29	0	0	0	9	8	12	0	0	0
12		0	621	1	9	13	176	35	372	10	5	0
13		0	62	0	1	0	35	4	17	3	2	0
14		0	88	0	2	0	13	27	45	1	0	0
15		0	165	0	3	1	41	12	102	4	2	0

Figure 5: Table Object, for example, about FBI crime rate, represents Tabular information.

3 EXTRACTING INFORMATION FROM RASTER DATA OVER DEFINED BOUNDARIES

Given a raster data set and a vector data set defining boundaries, we address the problem of extracting information from raster data over a set of boundaries. The assumption of this problem is that the raster and vector data sets overlap geographically and both data sets are described by sufficient georeferencing information for data fusion. Information extracted from raster data should describe statistics of an ensemble of pixel values inside of each defined boundary. In this work, we extracted descriptive statistics, such as, sample mean, standard deviation, skew and kurtosis, from continuous variable raster data sets and frequencies of label occurrences (or a histogram) from categorical variable raster data sets. The overview of the functionalities for georeferencing and raster information extraction is presented in Figure 6. We will discuss issues related to georeferencing in Section 3.1 and raster information extraction in Section 3.2.

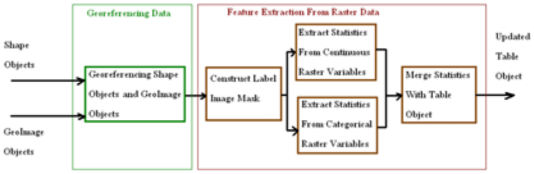

Georeferencing Data
Feature Extraction From Raster Data
Shape Objects
GeoImage Objects
Georeferencing Shape Objects and GeoImage Objects
Construct Label Image Mask
Extract Statistics From Continuous Raster Variables
Extract Statistics From Categorical Raster Variables
Merge Statistics With Table Object
Updated Table Object

Figure 6: An overview of "georefencing data sets" and "raster information extraction" components in Figure 1.

3.1 Georeferencing Vector And Raster Data Sets

Georeferencing, or geographic referencing, is the name given to the process of assigning values of latitude and longitude to features on a map. Latitude (lat) and longitude (long) describe points in three-dimensional (3D) space, while maps are inherently two-dimensional (2D) representations. With the advent of computers, modern maps are usually stored as digital images since they represent raster information similar to 2D image information. The steps involved in georeferencing a digital map image can vary among map image specification types, but the end result is the ability to retrieve the lat/long coordinates for any point on the georeferenced map. This capability is useful because the lat/long coordinates precisely define the position of an object on the Earth.

Lat/long coordinates define a point on a 3D model of the Earth, while map coordinates represent a pixel – a row and column location on a 2D grid obtained from projecting some 3D model of the Earth onto a plane. 2D maps are easy to display and facilitate distance measurement, while 3D coordinates are accurate but cumbersome and have no standard length for different degrees of latitude and longitude. The best way to define the position of an object on a map is with relative horizontal (column) and vertical (row) distances: "Three kilometers north of object A and two kilometers west of object B". In contrast, the best way to specify the position on a 3D sphere is with relative angular offsets: "Five degrees north and six degrees west of A".

A software package that can accept geospecific data and provide geographic referencing is called a *Geographic Information System* (GIS). One example of a GIS system is ArcGIS from ESRI [2], with components that include ArcView and ArcExplorer. ENVI, the Environment for Visualizing Images from Research Systems, Inc. [6], and a suite of applications known as the USGS Mapping Science Software from the U.S. Geological Survey [3] are other GIS systems.

There is nearly limitless information available describing georeferencing (see [8], [11], [12], [13], [14]), but the level of understanding necessary to correctly georeference a single image can be rather daunting. The difficulty is due in part to the complications of projecting a three-dimensional surface, the Earth, onto a two dimensional object, a map. Understanding something about this process goes a long way towards understanding why certain parameters are needed to geographically orient a digital image.

3.1.1 Classes of Georeferencing Transformations

Given 2D and 3D coordinate systems, there are three possible classes of transformations. First, those that go from 2D-to-

2D, bringing into alignment points that represent identical real-world locations. An example of this is multi-modal raster data fusion, for instance, the fusing of synthetic aperture radar (SAR) images with electro-optical (EO) or infrared (IR) images. In our work, we developed a tool for merging multiple digital elevation maps because the USGS data sets came on multiple CDs and any further processing required merging raster files with their associated georeferencing information. An example of this type of fusion is shown in Figure 7, where the fused raster file occupies 1.2 GigaBytes of memory.

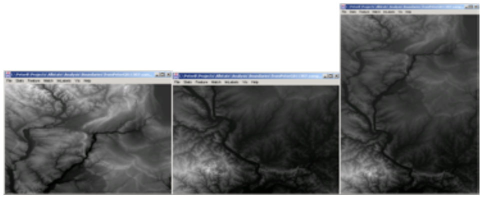


Figure 7: The problem of fusing multiple raster files (left and middle) with elevation information into one file (right).

The second class of transformations, 3D-to-3D, is used for information integration such as the synthesis of road network and river information. The direct use of this type transformation under our development was for fusing boundary information over multiple states (see Figure 8).

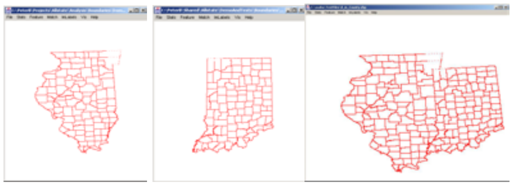


Figure 8: The problem of fusing multiple vector files (left and middle) describing county boundaries of Illinois and Indiana into one file (right).

The third class of transformations operates between 2D and 3D coordinates. Identifying the location of water wells defined in 3D lat/long coordinates on a 2D digital terrain map is an application example in this class of transformations. The transformations across coordinate systems of different dimensions, where one of the coordinate systems represents the Earth, are also called geo-registration. In our work, we focused primarily on the third class of transformations, those that support georeferencing maps (2-D raster data) with contours or boundaries of regions (3-D vector data) in order to extract raster statistics over a defined set of boundaries. An example of this type of georeferencing is shown in Figure 9.

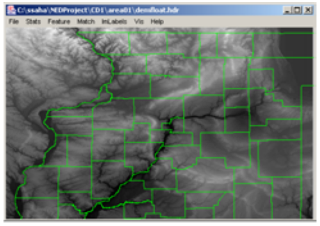

C:\ssaha\NEDProject\CD1\area01\demfloat.hdr
File Stats Feature Match ImLabels Vis Help

Figure 9: An example of georeferencing the Illinois elevation map with Illinois county boundaries and overlaying the boundary information (green) over the elevation map.

3.1.2 Georeferencing Coordinate Transformations

The georeferencing system is built around coordinate transformations to and from three types of location representations: 2D map pixels (column and row), 3D lat/long coordinates, and 2D UTM values. These transformations are shown in Figure 10.

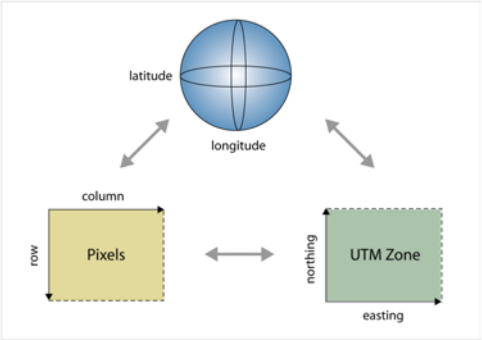


Figure 10: Types of coordinate transformations.

The motivation for the transformations between 2D map and 3D lat/long coordinate systems is fairly obvious, because one would like to know the lat/long values of any pixel in a georeferenced image. The third location representation, the Universal Transverse Mercator (UTM) coordinate system, is a Cartesian coordinate system developed by the oil industry. It is useful for specifying a number of points on a map without having to refer to latitude and longitude. Furthermore, UTM values facilitate metric distance calculations, which are difficult in lat/long coordinates where the distance between two adjacent degrees is dependent upon their location relative to the equator and the prime meridian. In UTM terminology, the horizontal (x) value is called *easting* and the vertical (y) coordinate is called *northing*. A UTM coordinate pair is called a *northing/easting* value. The details about the used set of equations and user's interface can be found in [16] and [17].

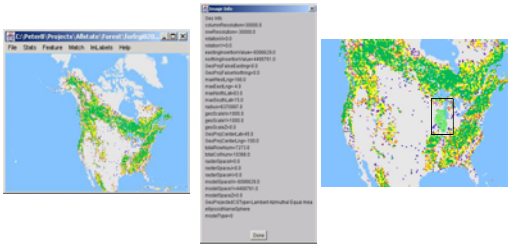
3.1.3 Sources of Georeferencing Information about Raster Files

In order to perform the geographic referencing transformations correctly, a number of map parameters must be given along with the map data. While some file formats were designed to contain georeferencing information, for example the ERDAS IMG file format, the file types of most digital maps were not designed to accommodate this metadata. For these file formats, of which the Tagged Image File Format (TIFF) is one example, the metadata must be provided in some manner for the map to be of any use. Although a number of approaches have been taken, the methods for including the metadata have not yet reached a point of standardization that enables a novice to quickly find the necessary information.

In this section we describe how geographic referencing information can be extracted from TIFF, for example, from the forest label image shown in Figure 11. The georeferencing information extraction from TIFF world files (tfw) and/or private tags is based on the literature available in [6], [7], [9], [10]. The current solution for the TIFF format uses a combination of information extraction approaches that are based on three variants of TIFF files containing georeferencing information. Assuming that sufficient data is provided in some combination of the three sources, we can use the georeferencing interface described earlier. The three TIFF file variants are:

- One or more standardized files are distributed along with TIFF image data as .tfw and/or .txt files.
- The metadata is encoded in the image file using private TIFF tags.
- An extension of the TIFF format called *GeoTIFF* is used.

Given multiple information sources, it is possible to read conflicting values for the same geographic field. For example, tie points and resolution values can be specified in both the TIFF world file and the private TIFF tags. When both a tfw file and the private tags are present, the tfw values have priority. Some georeferencing information may be specified either in deg/min/sec or UTM meters. In particular, if the information in private tag 33922 is specified in deg/min/sec, Image To Knowledge (I2K) tools [5] will perform automatically the necessary transformation. Finally, in the absence of necessary information, some transformations will not be executed. For example, if both the tfw file and the private tags 33550 and 33922 are missing, Transformations 3 and 4 cannot be performed.



```
Geo Info
columnResolution=30000.0
rowResolution=30000.0
rotationX=0.0
rotationY=0.0
eastingInsertionValue=-6086629.0
northingInsertionValue=4488761.0
GeoProjFalseEasting=0.0
GeoProjFalseNorthing=0.0
maxWestLng=180.0
maxEastLng=-4.0
maxNorthLat=83.0
maxSouthLat=15.0
radius=6370997.0
geoScaleX=1000.0
geoScaleY=1000.0
geoScaleZ=0.0
GeoProjCenterLat=45.0
GeoProjCenterLng=-100.0
totalRowNum=7273.0
totalColNum=10388.0
rasterSpaceI=0.0
rasterSpaceJ=0.0
rasterSpaceK=0.0
modelSpaceX=-6086629.0
modelSpaceY=4488761.0
modelSpaceZ=0.0
GeoProjectedCSType=Lambert Azimuthal Equal Area
ellipsoidName=Sphere
modelType=0
```

Figure 11: A TIFF image with forest cover labels (left) and its georeferencing information (middle) can be geo-registered with a shape file containing boundaries of Illinois counties (right image – Illinois falls within indicated rectangle).

3.2 Raster Information Extraction

In order to compute statistics from georeferenced raster files with boundaries, one has to establish a label image first. The label image defines a membership of each pixel (or raster location) to a boundary and is perfectly registered with the raster data. Statistics are then computed from all raster values that are labeled with the same label value in the label image.

3.2.1 Constructing Label Image

Given a set of 2D raster points and another set of boundary points in a pixel coordinate system, the goal of label image construction is to assign a unique label to all raster points surrounded by connected boundary points of each boundary. This problem can be solved theoretically in many ways assuming that (a) internal points of boundaries are known and are extracted from TIGER files, (b) a set of boundary points called a polygon [4] consists of one or more rings, (c) a ring is a connected sequence of four or more points that form a closed, non-intersecting loop, (d) a polygon may contain multiple outer rings and (e) the order of ring points defines whether the ring delineates interior points (outer rings have clockwise order) or exterior points (inner rings or holes have a counter clockwise order).

Although there are many approaches to this problem, the most robust approach is based on painting each ring separately, running connectivity analysis, forming a layered image with all result of connectivity analysis and assigning the same label to all pixels that were marked several times in the layered image. Other approaches, while fully justifiable in a continuous data domain, fail in a discrete data domain. Examples of label images are shown in Figure 12 and Figure 13.

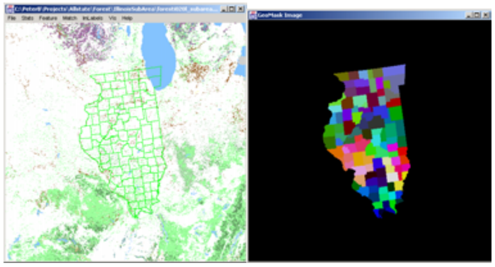


Figure 12: Georeferenced raster file of forest categories with boundaries of Illinois counties (left) and the corresponding label image (right) used for computing statistics.

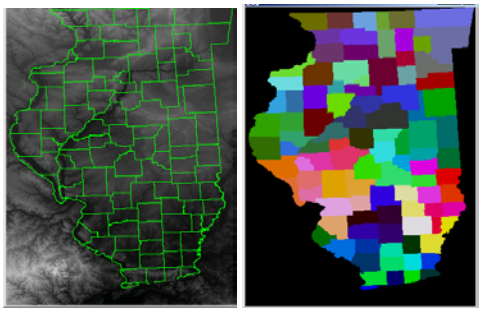
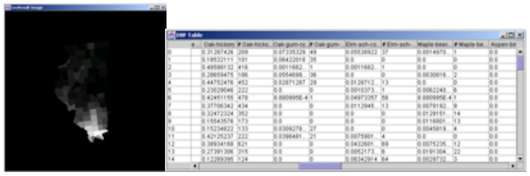


Figure 13: Georeferenced raster file of elevation data with boundaries of Illinois counties (left) and the corresponding label image (right) used for computing statistics.

3.2.2 Computing Statistics from Raster Data Over Label Mask

Given two images of identical size representing raster values and label values, we calculated statistics of all raster values labeled with the same label. In a case of categorical raster data, such as, the forest image with color labels for tree types, we computed occurrence of each label within a boundary (see Figure 14). In a case of continuous raster data, such as, the elevation maps, we computed sample mean, standard deviation, skew and kurtosis per boundary (see Figure 15). All computed statistics were saved in a tabular form (Table Object) and presented to an end user by mapping the range of resulting values into an intensity map of geographical boundaries.



DBF Table

	e...	Oak-hickory	# Oak-hicko...	Oak-gum-cy...	# Oak-gum-...	Elm-ash-co...	# Elm-ash-...	Maple-beec...	# Maple-be...	Aspen-bir...
0		0.31287426	209	0.07335329	49	0.05538922	37	0.0014970...	1	0.0
1		0.18532111	101	0.06422018	35	0.0	0	0.0	0	0.0
2		0.49598132	416	0.0011682...	1	0.0011682...	1	0.0	0	0.0
3		0.28659475	186	0.0554699...	36	0.0	0	0.0030816...	2	0.0
4		0.44752476	452	0.02871287	29	0.0128712...	13	0.0	0	0.0
5		0.23029046	222	0.0	0	0.0010373...	1	0.0062240...	6	0.0
6		0.42451155	478	8.880995E-4	1	0.04973357	56	8.880995E-4	1	0.0
7		0.37706342	434	0.0	0	0.0112945...	13	0.0078192...	9	0.0
8		0.32472324	352	0.0	0	0.0	0	0.0129151...	14	0.0
9		0.15543576	173	0.0	0	0.0	0	0.0116801...	13	0.0
10		0.15234822	133	0.0309278...	27	0.0	0	0.0045819...	4	0.0
11		0.42125237	222	0.0398481...	21	0.0075901...	4	0.0	0	0.0
12		0.38934168	621	0.0	0	0.0432601...	69	0.0075235...	12	0.0
13		0.27391306	315	0.0	0	0.0052173...	6	0.0191304...	22	0.0
14		0.12289395	124	0.0	0	0.06342914	64	0.0029732...	3	0.0

Figure 14: Occurrence statistics computed for the forest image shown in Figure 12 and presented graphically (left) and saved in a tabular form (right).

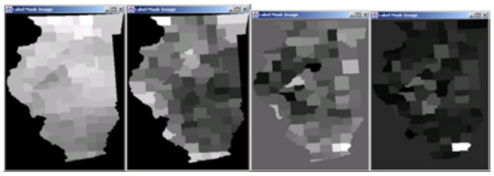
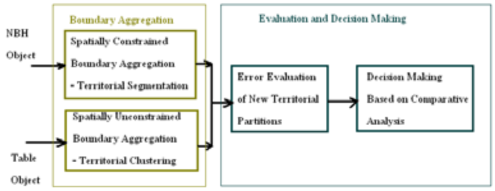


Figure 15: Descriptive statistics computed for the elevation image shown in Figure 13 and presented graphically. The images display (from left to right) the intensity mapping of sample mean, standard deviation, skew and kurtosis values per boundary.

4 BOUNDARY AGGREGATION AND EVALUATION FOR DECISION MAKING

In this section we describe the problems of (1) boundary aggregation given a set of boundary features (also called attributes) and (2) evaluation of boundary aggregations (also called geographic partitions) together with decision-making based on comparative analysis. The schematic overview of this process is illustrated in Figure 16.



NBH Object
Table Object
Boundary Aggregation
Spatially Constrained Boundary Aggregation - Territorial Segmentation
Spatially Unconstrained Boundary Aggregation - Territorial Clustering
Evaluation and Decision Making
Error Evaluation of New Territorial Partitions
Decision Making Based on Comparative Analysis

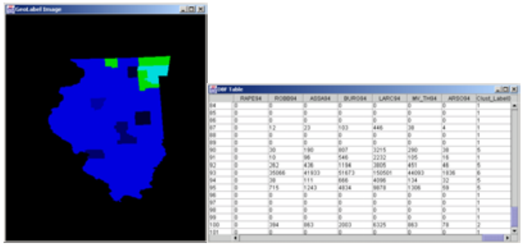
Figure 16: An overview of “feature driven boundary aggregation” and “evaluation and decision making” blocks in Figure 1.

4.1 Boundary Aggregation

In general, boundary aggregation can be formulated as a clustering problem with no spatial constraints [18] or as a segmentation problem with spatial constraints [19] on the final boundary aggregations. Merging spatial boundaries into boundary aggregations (or geographical geographic partitions) is driven by maximizing intra-partition similarity of features and inter-partition dissimilarity of features. Spatial constraints are defined such that the final aggregations must form spatially contiguous geographical partitions.

An aggregation process can stop if (1) a certain number of final partitions have been found or (2) a maximum user-specified intra-partition dissimilarity value has been exceeded. The aggregation of two boundaries (any boundaries or spatially adjacent boundaries) takes place when the Euclidean distance between their features is less than a discretely incremented threshold value until one of the exit criteria is met. The output can be the last partition of the incremental aggregation process or several partitions obtained at multiple threshold values. This bottom-up aggregation process and the multi-partition output explains the method's adjective hierarchical.

We have developed the two hierarchical methods for clustering (no spatial constraint) and segmentation (with spatial constraint). The result of aggregation is stored in a tabular form, for example, Clust_Label0 in Figure 17 (right), and presented to an end-user in a visual form with respect to the geographical location of each boundary (see Figure 17, left). Results from clustering and segmentation can be easily visually compared as it is shown in Figure 18. Hierarchical results obtained as multiple intermediate outputs of the aggregation process are presented as a movie and typical results are shown in Figure 19.



	RAPE94	ROBB94	ASSA94	BURG94	LARC94	MV_TH94	ARSO94	Clust_Label0
84	0	0	0	0	0	0	0	1
85	0	0	0	0	0	0	0	1
86	0	0	0	0	0	0	0	1
87	0	12	23	103	446	38	4	1
88	0	0	0	0	0	0	0	1
89	0	0	0	0	0	0	0	1
90	0	30	190	807	3215	290	38	5
91	0	10	96	546	2232	105	16	1
92	0	262	436	1184	3805	451	48	5
93	0	35066	41933	51673	150501	44093	1836	6
94	0	38	111	666	4096	134	32	5
95	0	715	1243	4834	9878	1306	58	5
96	0	0	0	0	0	0	0	1
97	0	0	0	0	0	0	0	1
98	0	0	0	0	0	0	0	1
99	0	0	0	0	0	0	0	1
100	0	394	863	2003	6325	863	78	2
101	0	0	0	0	0	0	0	1

Figure 17: Hierarchical clustering of FBI crime data with the exit criterion being the number of clusters and the clustered feature being auto theft in 2000 leads to six aggregations that are geographically depicted (left) and the labels are stored in a tabular form (right).

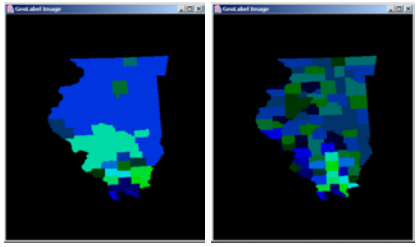


Figure 18: The results of hierarchical segmentation (left) and hierarchical clustering (right) of oak hickory feature with the exit criterion of 18 numbers of connected county aggregations.

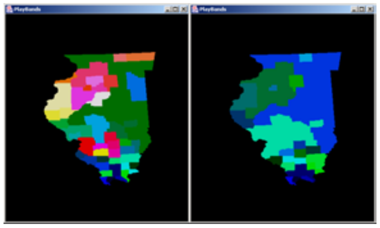


Figure 19: Hierarchical segmentation of extracted forest statistics (oak hickory occurrence) with two output partitions containing 43 (left) and 21 (right) aggregations. The feature for this example was extracted from the image shown in Figure 12.

4.2 Evaluation and Decision Making

Once multiple geographic partitions have been obtained, there is a need to evaluate each partition and make a decision about selecting the best partition according to some criteria. We have addressed this problem by error evaluation of each partition followed by comparative analysis.

4.2.1 Error Evaluation of New Geographic Partitions

In order to evaluate an error of a given geographic partition, it is assumed that the partition without any boundary aggregations had zero error. A partition formed by boundary aggregations will inherently introduce feature errors or deviations from the partition with no aggregations because aggregated boundaries are represented by a new feature value that might differ from the original feature values of un-aggregated boundaries. However, if the cost of having and maintaining a large number of partitions is too high then the trade-off between the number of partitions and the associated feature prediction error must be considered.

An error evaluation can be conducted by using multiple error metrics depending on the application domain. We have developed several metrics defined in Equations below in order to provide a choice of metrics to end-users. The list of metrics includes absolute difference, variance, city block, normalized variance and normalized city block.

Variance metric:

$$Var^k = \sum_{r;label(r)=k} \sum_j (w(r) * f_j(r) - \bar{f}_j^{\,k})^2 = \sum_j Var(\{f_j^k\})$$

City Block metric:

$$CB^k = \sum_{r;label(r)=k} \sum_j \left\| w(r) * f_j(r) - \bar{f}_j^{\,k} \right\| = \sum_j CB(\{f_j^k\})$$

Normalized Variance Metric:

$$NormVar^k = \sum_j \frac{Var(\{f_j^k\})}{\bar{f}_j}$$

Normalized City Block metric:

$$NormCB^k = \sum_j \frac{CB(\{f_j^k\})}{\bar{f}_j}$$

where values are defined as

$$\bar{f}^k = \left(\frac{\sum_{r;label(r)=k} f_1(r) * w(r)}{\sum_{r;label(r)=k} w(r)}, \frac{\sum_{r;label(r)=k} f_2(r) * w(r)}{\sum_{r;label(r)=k} w(r)}, \ldots \right) = (\bar{f}_1^{\,k}, \bar{f}_2^{\,k},)$$

and the notation corresponds to:

$k - aggregation \quad label$

$j - feature \quad index$

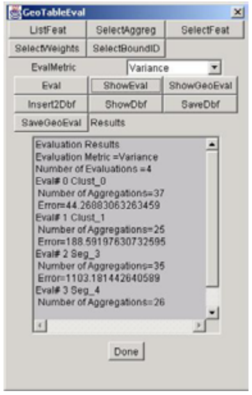
$r - row\ representing\ a\ region$

$\bar{f}_j - sample\ mean\ of\ a\ feature\ j\ over\ all\ regions$

4.2.2 Comparative Analysis

In order to compare multiple partitions, there are two proposed approaches. First, a global prediction error is computed per each partition and the result error values are compared. For example, the results of clustering and segmentation with the sample mean feature of Illinois counties can be evaluated and compared as it is shown in Figure 20.

Second, any prediction error can be viewed as a geographical error distribution since given any set of boundary aggregations the error varies with respect to a geographic location determined by the granularity of the underlying boundaries and their corresponding aggregations. Thus, the error distribution can be viewed as a sequence of geo-image where bright corresponds to a large error while dark means a small error. An example of geographical error distribution is shown in Figure 21.



EvalMetric: Variance

Results

```
Evaluation Results
Evaluation Metric =Variance
Number of Evaluations =4
Eval# 0 Clust_0
 Number of Aggregations=37
 Error=44.26883063263459
Eval# 1 Clust_1
 Number of Aggregations=25
 Error=188.59197630732595
Eval# 2 Seg_3
 Number of Aggregations=35
 Error=1103.181442640589
Eval# 3 Seg_4
 Number of Aggregations=26
```

Figure 20: An example of global errors for multiple partitions obtained by clustering and segmentation of terrain elevation attribute of Illinois counties.

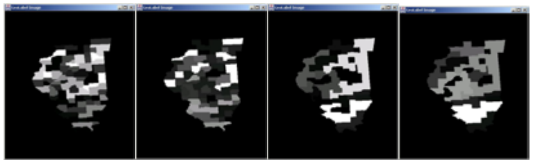


Figure 21: Geographical error distribution for the results presented in Figure 20. From left to right, geographical error distribution for the partitions with 37 (Eval# 0 Clust_0), 25 (Eval# 1 Clust_1), 35 (Eval# 2 Seg_3) and 26 (Eval# 3 Seg_4) aggregations.

In order to compare prediction errors computed for a specific region from several partitions, one could inspect computed errors from all partitions as a function of boundary indices. This type of comparative analysis is facilitated by the appropriate visualization and color-coding each partition differently. An example is illustrated in Figure 22. However, the tacit assumption of this type comparison is that the aggregations are defined over the same set of underlying boundaries. This assumption can be violated, for example, if one would like to compare prediction errors at a defined location by latitude and longitude and the defined location was evaluated with counties and zip codes as the underlying boundaries. In this case, the prediction errors have to be mapped from one type of boundary to another type of boundary. This mapping can be accomplished by a simple image subtraction of images with geographical error distributions.

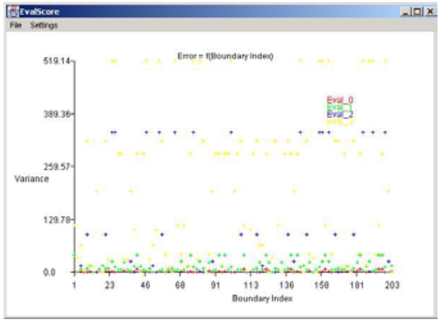


Figure 22: Comparative analysis of boundary prediction errors obtained from multiple partitions.

5 VISUALIZATION

It is apparent that visualization of input data, intermediate results and final geographic partitions plays a very important role in understanding of the input data and the obtained data mining results. We have presented throughout the paper visualizations of a variety of (1) raster (Figure 3), vector (Figure 4) and tabular (Figure 5) input data, (2) intermediate results, for example, overlaid georeferenced data (Figure 9), georeferencing information (Figure 12) and label image of georeferenced raster and boundary data (Figure 13), and (3) final results, for instance, statistics of extracted features (Figure 14, Figure 15), geographical partitions obtained by single output clustering (Figure 17) and segmentation (Figure 18) or multiple output partitioning strategy (Figure 19). We might develop in future a visualization tool for the neighborhood information (NBH Object).

6 SUMMARY

We have presented a decision support system using data mining methods and geographic information that forms geographic regions from GIS raster and vector data. The system development involved solving several research issues including (1) data representation (raster, vector and tabular information), (2) data fusion using key features (multiple tabular information sources) and using georeferencing information (raster and vector data), (3) extraction of statistics from continuous (terrain elevation) and categorical (forest labels) raster data sets, (4) development of boundary aggregation methods (with and without spatial constraints), (5) error evaluation metrics and comparative analysis of multiple resulting geographic aggregations. The described system has been applied to the problem of police force deployment. The developed approaches and software are currently being applied to hydrology [22] and water quality [15] applications to test the generality of the presented approach and the functionality of the developed software.

REFERENCES

[1] Campbell, James B. Introduction to Remote Sensing, Second Edition. The Guilford Press, New York. 1996.

[2] Miller, Catherine L. Tiger/Line Files Technical Documentation. UA 2000. US Department of Commerce, Geography Division, U.S. Census Bureau.http://www.census.gov/geo/www/tiger/tigerua/ua2ktgr.pdf

[3] ArcExplorer product description at ESRI website: http://www.esri.com/

[4] ESRI Shape file, File Format Specification, http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[5] Bajcsy P. et al., "Image To Knowledge" Documentation at web site: http://alg.ncsa.uiuc.edu/tools/docs/i2k/manual/index.html

[6] TIFF Revision 6.0. Aldus Corporation. Copyright 1986-88, 1992.

[7] GeoTIFF Revision 1.0. Ritter, Niles. Ruth, Michael. 2000.

[8] Map Projections – A Reference Manual. L. M. Bugayevsky and J. P. Snyder. Taylor and Francis, 1995.

[9] Information on tfw files: http://www.genaware.com/html/support/faqs/imagis/imagis15.htm

[10] Information on private TIFF tags: http://remotesensing.org/geotiff/spec/geotiff2.6.html

[11] Understanding Map Projections. Kennedy, Melita. Kopp, Steve. Environmental Systems Research Institute, Inc.

[12] Molodensky equations reference http://www.posc.org/Epicentre.2_2/DataModel/ExamplesofUsage/eu_cs34h.html

[13] Lambert projection equations http://mathworld.wolfram.com/LambertAzimuthalEqual-AreaProjection.html

[14] Online reference to transformation equations http://www.posc.org/Epicentre.2_2/DataModel/ExamplesofUsage/eu_cs.html

[15] Novotny V., "Water Quality, Diffuse Pollution and watershed Management,"

second edition, John Wiley & Sons Inc., 2003.

[16] Alumbaugh T.J. and Bajcsy P., "Georeferencing Maps with Contours in I2K," ALG NCSA technical report, alg02-001, October 11 2002.

[17] Bajcsy P. and Alumbaugh T.J., "Georeferencing Maps with Contours," *Journal Proceedings of 7th World Multiconference on Systemics, Cybernetics, and Informatics. Vol X.* July 27-30, 2003. Orlando, Florida.

[18] Han, J., and Kamber, M., 2001. "Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA.

[19] Duda, R., P. Hart and D. Stork, 2001. Pattern Classification, Second Edition, Wiley-Interscience.

[20] Jasani B. and Stein G., "Commercial Satellite Imagery, a tactic in nuclear weapon deterrence," Springer, 2002.

[21] Swain P.H. and Davis S.M. editors, "Remote Sensing: the Quantitative Approach," McGraw-Hill, Inc., 1978.

[22] Maidment D.R., "Arc Hydro, GIS for Water Resources," ESRI Press, 2002.

[23] USGS Mapping Science Software: http://mapping.usgs.gov/www/products/software.html

[24] Research Systems, Inc, The Environment of Visualizing Images (ENVI), http://www.rsinc.com/envi/index.cfm