GeoPAT: A toolbox for pattern-based information retrieval from large geospatial databases

Jarosław Jasiewicz[a,c,*], Paweł Netzel[b,c], Tomasz Stepinski[c]

[a] Institute of Geoecology and Geoinformation, Adam Mickiewicz University in Poznan, Dziegielowa 27, 60-680 Poznan
[b] Department of Climatology and Atmosphere Protection, University of Wroclaw, Kosiby 6/8, 51-621 Wrocław, Poland
[c] Space Informatics Lab, Department of Geography,University of Cincinnati, Cincinnati, OH 45221-0131, USA

Abstract

Geospatial Pattern Analysis Toolbox (GeoPAT) is a collection of GRASS GIS modules for carrying out pattern-based geospatial analysis of images and other spatial datasets. The need for pattern-based analysis arises when images/rasters contain rich spatial information either because of their very high resolution or their very large spatial extent. Elementary units of pattern-based analysis are scenes – patches of surface consisting of a complex arrangement of individual pixels (patterns). GeoPAT modules implement popular GIS algorithms, such as query, overlay, and segmentation, to operate on the grid of scenes. To achieve these capabilities GeoPAT includes a library of scene signatures – compact numerical descriptors of patterns, and a library of distance functions – providing numerical means of assessing dissimilarity between scenes. Ancillary GeoPAT modules use these functions to construct a grid of scenes or to assign signatures to individual scenes having regular or irregular geometries. Thus GeoPAT combines knowledge retrieval from patterns with mapping tasks within a single integrated GIS environment. GeoPAT is designed to identify and analyze complex, highly generalized classes in spatial datasets. Examples include distinguishing between different styles of urban settlements using VHR images, delineating different landscape types in land cover maps, and mapping physiographic units from DEM. The concept of pattern-based spatial analysis is explained and the roles of all modules and functions are described. A case study example pertaining to delineation of landscape types in a subregion of NLCD is given. Performance evaluation is included to highlight GeoPAT's applicability to very large datasets. The GeoPAT toolbox is available for download from http://sil.uc.edu/.

Keywords: pattern analysis, query-by-example, large geospatial datasets, similarity, image classification, GRASS GIS

1. Introduction

Most spatial datasets in geosciences originate from remote sensing (RS) and are in the form of images. Therefore, there exists a significant body of literature on retrieving information from RS images (Richards, 1999). Image classification - a process of converting an image into a thematic map of semantically meaningful classes - is the most common form of spatial information retrieval from an image (Lu and Weng, 2007). An original approach to image classification utilizes a pixel-based methodology. A pixel is the smallest element of a surface, as depicted in an image, for which a value of a color is stored. A pixel-based classification algorithm assigns class labels to individual pixels. Note that this is fundamentally different from how an analyst interprets an image by perceiving the coherence of colors on multiple scales simultaneously and assigning class labels to multi-pixel tracts on the basis of their textures or patterns. Pixel-based classification algorithms may suffer from poor performance especially if applied to very high resolution (VHR) images, where individual pixels correspond to small elements of real objects and their numerical attributes are not sufficient to recognize the class of an object, or, if applied to very large images where the goal of analysis is to retrieve generalized classes (for example, when the goal is to retrieve landscape types rather than their constituent land cover classes (Graesser et al., 2012; Niesterowicz and Stepinski, 2013; Vatsavai, 2013a; Jasiewicz et al., 2014)).

*Corresponding author
Email addresses: jarekj@amu.edu.pl (Jarosław Jasiewicz), pawel.netzel@uni.wroc.pl (Paweł Netzel), stepintz@ucmail.uc.edu (Tomasz Stepinski)

Object-Based Image Analysis (OBIA) was developed (Blaschke, 2010; Lang, 2008) to alleviate the problems associated with pixel-based classification. In OBIA image is first segmented to simplify it by grouping pixels into meaningful segments (called "objects") which are homogeneous with respect to pixel-based attributes. In the second step information is retrieved by classifying objects into semantically meaningful classes. OBIA algorithms get closer to the way an analyst interprets an image but they still suffer from a number of shortcomings (Vatsavai, 2013b). First, segmentation itself is a complex and computationally expensive process and there is no single method that performs consistently well (does not under-segment or over-segment portions of an image) on different RS images. Second, because objects are, by definition, homogeneous segments of the surface, OBIA cannot be used to classify an image into highly generalized classes. For example, although OBIA can classify an image into land cover classes (low-level generalization) more accurately than a pixel-based classifier can, it still cannot classify it into landscape types (high-level generalization). In other words, OBIA can utilize information about image texture but not information about spatial patterns.

For the purpose of this paper we define a spatial pattern as a perceptual structure, placement, or arrangement of image objects having a geometric quality. We then define texture as a structure of pixels arranged quasi randomly and lacking geometric quality. Thus, a single land cover class in a VHR image (for example, a rooftop) is characterized by texture as it appears on image as a quasi random mosaic of pixels having a range of colors. However, a fragment of a thematic map showing an urban scene consisting of a spatial arrangement of several land cover classes needs to be characterized by its pattern.

The case for classifying an image or image-like spatial dataset, for example a Digital Elevation Model (DEM), on the basis of spatial patterns arises in multiple disciplines where a high level of generalization is desired. In RS, with VHR images containing rich spatial information, the use of a pattern-based classification method makes it possible to distinguish between different urban landscapes, for example, between informal settlements, industrial/commercial structures, and formal residential settlements (Graesser et al., 2012; Vatsavai, 2013a). In landscape ecology, it makes it possible to distinguish between different landscape types (Niesterowicz and Stepinski, 2013; Cardille and Lambois, 2009) as well as between different types of forest structures (Long et al., 2010), and in geomorphology it makes it possible to identify and delineate physiographic units (Jasiewicz et al., 2014).

It is only recently that methodologies for pattern-based information retrieval from images and other raster datasets have been proposed. Vatsavai (2013a) proposed a multi-instance learning (MIL) scheme as a means for the pattern-based classification of images. In this method, an image is divided into regular grid of local blocks of pixels. The data (a set of all multi-dimensional attribute vectors from each pixel) in each block is modeled using a multivariate Gaussian distribution. The distance (dissimilarity) between any two blocks, and thus between the two patterns contained in these blocks, is calculated as the probabilistic distance between their modeled Gaussian distributions using the Kullback-Leibler (KL) divergence. Using supervised learning based on the MIL scheme Vatsavai (2013a) and Graesser et al. (2012) classified RS images of several cities into formal and informal neighborhoods.

Independently, we have proposed a general approach for pattern-based information retrieval from all types of geospatial datasets (Jasiewicz and Stepinski, 2013a; Stepinski et al., 2014). For our method to be broadly applicable and computationally efficient it uses an input (image, DEM etc.) that has been preprocessed using a pixel-based classification and thus already converted into a categorical format. This categorical raster is divided into a regular grid of local blocks of pixels. Because the data is categorical, each block can be compactly represented by a histogram of categories or other attributes derived from these categories. We have successfully applied this methodology to search for and classify land-cover patterns in the National Land Cover Dataset (NLCD) (Jasiewicz and Stepinski, 2013a). We have also used it for an assessment of land cover change over the entire United States using the NLCD (Netzel and Stepinski, 2015), and for the identification and delineation of physiographic units using DEM data (Jasiewicz et al., 2014).

The concept of pattern-based information retrieval from geospatial datasets is at the beginning of its developmental cycle. For this concept to mature much more work is needed, including application to many different datasets in multiple contexts. In this paper we present the Geospatial Pattern Analysis Toolbox (GeoPAT) - a collection of GRASS GIS modules that integrate the various tools necessary for experimenting with pattern-based information retrieval from geospatial data. GeoPAT is intended as a convenient platform for experimentation with the pattern-based analysis of rasters including rasters having giga-cell and larger sizes. It integrates into the GIS system procedures for pattern description, pattern similarity, and the

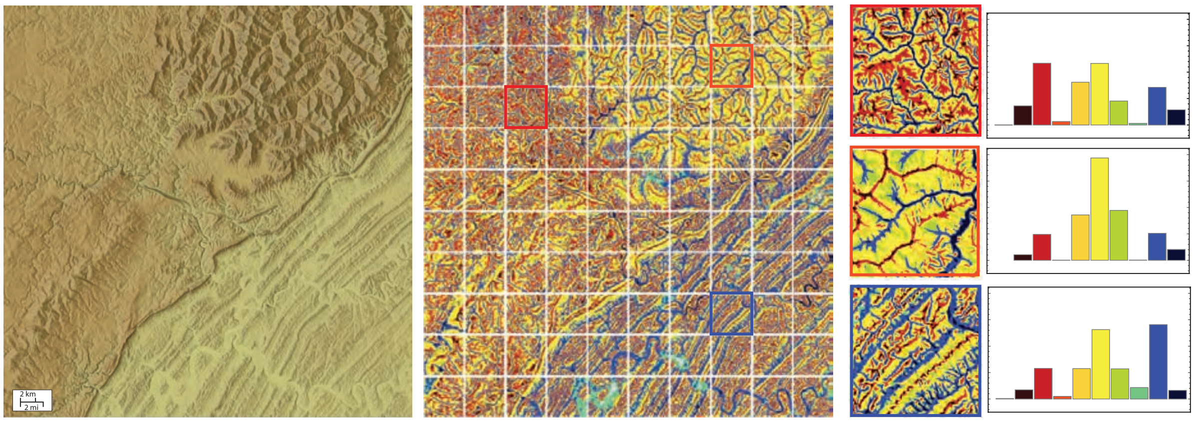

Figure 1: Example introducing a concept of pattern-based analysis of spatial datasets. (Left) Hillshade rendition of DEM over 60×60 km region. (Middle) DEM data classified into ten landform classes and divided into regular blocks. (Right) Close-ups of three sample blocks and histograms of their landform classes.

search and retrieval of similar patterns. These concepts were originally developed for working with natural images in the context of Content-Based Image Retrieval (CBIR) systems (Datta et al., 2008) but are now utilized by GeoPAT for the purpose of geospatial analytics. Such integration allows a user to perform the standard GIS tasks of mapping, map overlay, and segmentation on a grid of pattern-bearing blocks of pixels in a way which is already familiar (from performing similar tasks on standard images). In other words, GeoPAT extends the standard GIS system by adding a new type of attribute – the pattern signature – and a new type of data query – a query-by-pattern-similarity (QBPS). This significantly lowers the cost of entry into experimenting with pattern-based information retrieval, helps to accelerate further development of this concept, and makes possible the assessment of its utility in various domains.

GeoPAT modules are written in ANSI C and are designed to work within the GRASS GIS 7 (GRASS Development Team, 2012) environment. Embedding GeoPAT in GRASS has a number of advantages: (1) GRASS is an open source software available for major computing platforms, (2) GRASS is especially well-suited to work with large datasets, and (3) incorporating a toolbox into an already existing, well-established environment allows for an integrated computational pipeline that provides convenience and boosts efficiency (Körting et al., 2013). GeoPAT is an actively developed solution. The core of the toolbox consists of the seven modules that compute pattern signatures and perform the GIS tasks of comparing, searching, overlaying, and segmenting the rasters on the basis of similarity between local patterns. These modules provide the basic infrastructure for pattern-based information retrieval and are not expected to be modified by a user. In addition, two libraries provide a selection of functions for extracting pattern signatures and for calculation of similarity/distance between two patterns, respectively. As there are no standard means of representing spatial patterns and calculating a measure of similarity between them, we expect users to add to those libraries as they experiment with different datasets.

The rest of this paper is organized as follows: Section 2 presents an overview of our toolbox architecture. Section 3 describes the most important functions in the shared libraries and section 4 describes the seven core geoprocessing modules. A case study (section 5) presents an example on how GeoPAT modules can be utilized to perform regionalization of land cover patterns into landscape types using either unsupervised or supervised approaches. Section 6 gives an assessment of the computational performance of the GeoPAT modules and section 7 contains our discussion and conclusions.

2. Software architecture

As an introduction to GeoPAT we first give an illustration of the basic idea behind the pattern-based analysis of geospatial data. For this we use a DEM with 30 m resolution. The left panel in Fig. 1 shows a hillshade rendition of a 2000×2000 cell DEM (we reserve the term pixel for images and use the more general term *cell* for all raster datasets). The entire spatial extent of the data is referred to as *a region*. Three clearly distinct physiographic units are observed in this region and one of the goals of pattern-based analysis is to delineate these units. In the preprocessing step, which is not a part of GeoPAT, DEM cells are classified into ten landform classes using the geomorphons method (Jasiewicz and Stepinski, 2013b). The result of this classification is shown in the central panel of Fig. 1 with different colors indicating different landforms. This panel also shows a division of the region into a regular grid of blocks, each block containing a large number of cells forming a local, block-bounded pattern of landforms. A block is a particular example of *a scene*; in general, we refer to any subregion of the entire region as a scene.

A grid of regular scenes (as in the middle panel of Fig. 1) is referred to as *a grid-of-scenes*. GeoPAT performs GIS operations on the grid-of-scenes in the same way as the standard GIS system performs similar operations on the grid of cells. Thus, for example, to delineate the three physiographic units as seen in the sample DEM GeoPAT will classify the scenes in a way that is analogous to how a standard pixel-based algorithm would delineate different landform classes. Significant technical differences in performing these operations on scenes vs. cells stem from differences in mathematical representations of patterns vs. numbers, and from differences in the definitions of a distance between patterns vs. distance between vectors.

Close-ups of three sample scenes, labeled by red, orange, and blue frames are shown in the right panel of Fig. 1 with their corresponding histograms of landform classes. GeoPAT uses histograms as concise representations of patterns. Note that the three scenes, each representing a different physiographic unit and exhibiting a different pattern of landform classes, happen to have different histograms of classes. However, in general, visually different patterns may have similar histograms of classes. This is why GeoPAT uses more advanced histograms that encapsulate not only the composition of a pattern (the relative abundance of classes) but also its configuration (spatial arrangement of clumps – contiguous groups of same-class cells).

In general, the input to GeoPAT is a categorical raster (classified original spatial dataset, for example, an image, DEM, etc.) or a set of co-registered categorical rasters. Additional rasters are needed for some tasks, such as, for example, a change detection task, or to provide ancillary information for description of local patterns (see section 3.1). The raster's region is subdivided into a regular grid (grid-of-scenes) having cells (referred to as *s-cells*) with the size equal to or larger than the size of the raster cells. Each s-cell is the center of a square scene containing a local pattern made up of cells with different class labels. It also stores a concise description of this pattern which is referred to as *a signature*. The size of the scene must be equal to or larger than the size of the s-cell. This allows GeoPAT to work with overlapping scenes. If the size of the scene is equal to the size of the s-cell the scenes don't overlap (as in the case shown in Fig. 1). If the size of the scene is larger than the size of the s-cell the scenes overlap. In the example given in Fig. 1 the size of the cell is 30 m and the size of the s-cell is the same as the size of the scene and equal to 6000 m.

The core of the GeoPAT toolbox consists of three modules designed for extracting scene signatures from categorical data, as well as four additional modules for performing geoprocessing tasks on the grid-of-scenes. GeoPAT implements three different signature extraction modules: p.sig.points, p.sig.polygons and p.sig.grid. This is because some geoprocessing tasks require a description of scenes not restricted to those defined by the grid-of-scenes. For example, a search task requires a comparison of scenes in a grid-of-scenes with a scene (a query) defined over a region not aligned with a grid, and a segment classification task requires calculating signatures from irregularly-shaped scenes. The role of geoprocessing modules (p.sim.distmatrix, p.sim.search, p.sim.compare, p.sim.segment) is to perform geoprocessing tasks on scenes generated and described by the signature extraction modules. The names given to the modules adhere to the following convention: p. stands for pattern, sig. stands for signature, and sim. stands for similarity.

In addition to modules, GeoPAT provides two libraries of functions. The first library implements different methods of extracting a signature from a scene. Functions in this library work with signature extraction modules. The second library implements different distance measures between signatures. Functions in this library work with geoprocessing modules. We expect that users may want to add their own functions to both libraries. The overall software architecture of GeoPAT is shown in Fig. 2.

3. Library functions

Library functions implement concepts which facilitate working with spatial patterns in a quantitative fashion. In the domain of geoscience working quantitatively with spatial patterns has been addressed in the fields of remote sensing (Datcu et al., 2003; Daschiel and Datcu, 2005; Li and Narayanan, 2004; Shyu et al., 2007), landscape ecology (Cain and Riitters, 1997; Long et al., 2010; Cardille and Lambois, 2009; Dilts et al., 2010), and cartography (Pontius, 2002; Remmel and Csillag,

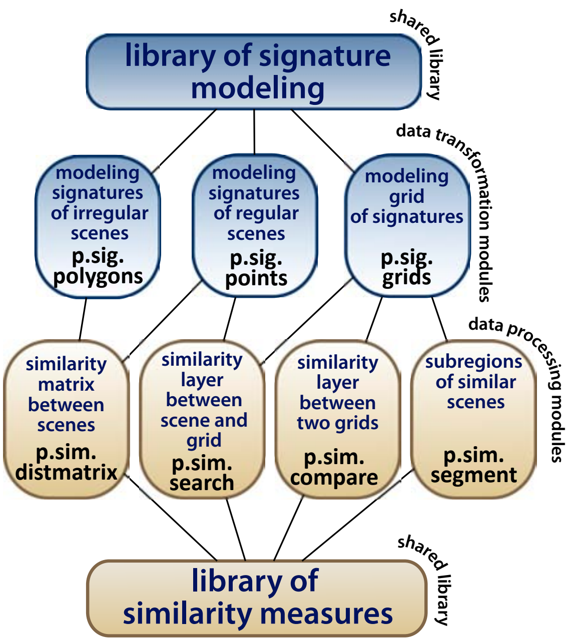

Figure 2: Architecture of GeoPAT toolbox

2006). As GeoPAT uses categorical rasters, only previous approaches developed in landscape ecology and cartography are potentially directly relevant. Because we are interested in pattern similarity measures that are rotationally invariant (for example, a scene and the same scene rotated by 90 degrees must be measured as identical), cartographic approaches, which focus on comparing different maps of the same region for consistency and accuracy, are not directly relevant.

In landscape ecology categorical patterns are described using landscape metrics (Haines-Young and Chopping, 1996; McGarigal et al., 2002; Uuemaa et al., 2009; Steiniger and Hay, 2009) which are rotationally invariant measures of compositional and configurational aspects of a scene. A collection of various landscape metrics forms an attribute vector which potentially can be used as scene signature. Several studies (Long et al., 2010; Kupfer et al., 2012; Cardille and Lambois, 2009; Cardille et al., 2012) used landscape metrics-based attribute vectors and the Euclidean distance to calculate similarities between mostly binary (forest/no forest) scenes, but the validity of such an approach has not been demonstrated. Our own experience with using landscape metrics for assessing the similarity between scenes is negative. We have identified a number of issues for using landscape metrics in GeoPAT including the selection of metrics (this can be overcome by data reduction using PCA (Cushman et al., 2008)), the proper way to normalize metrics, and properly weighting the contribution of composition vs. configuration to the overall similarity value.

Following the principles established in the field of Content-Based Image Retrieval (CBIR) (Gevers and Smeulders, 2004; Datta et al., 2008; Lew et al., 2006) – a non-geoscience domain where the issue of similarity between two rasters (natural images) has been studied extensively – GeoPAT calculates a signature as a (possibly multi-dimensional) histogram of a pattern ”primitive features.” Primitive features are simple local elements of a pattern. For example, the cell's class is a primitive feature. A combination of classes of two neighboring cells is an another example of a primitive feature. Many other such features could be designed. There is no generally preferred choice of primitive features; patterns in different datasets may be best encapsulated by different features. GeoPAT implements several popular methods of representing pattern by a histogram of primitive features, but it is expected that users may want to add their own.

In GeoPAT a similarity between two scenes is calculated as a similarity between two histograms, each representing a pattern contained in its respective scene. Choosing the most appropriate similarity function is largely an empirical decision which depends on the dataset and on the choice of primitive features. GeoPAT implements several histogram similarity functions, but, as with primitive features, we expect users to add their own. Cha (2007) provides a comprehensive review of histogram similarity functions.

3.1. Signature functions

Signature functions define primitive features that characterize a local pattern bounded by an extent of a scene. From among many possible signatures we describe three which are already implemented in GeoPAT.

crossproduct – This method calculates signature as a k-dimensional histogram using k primitive features assigned to each cell. Examples of such features include cell class, the size of the clump to which the cell belongs, the shape of the clump and its spatial orientation (Williams and Wentz, 2008). Because all features must be categorical (so the histogram can be formed), numerical features need to be categorized. For example, clump sizes need to be categorized into size categories from the smallest to the largest. The number of bins in the crossproduct histogram is $N_1 \times N_2 \times \ldots \times N_k$, where N_i is the number of categories of i-th feature. Fig. 3A shows schematically a construction of crossproduct histogram from two features, cell class (N_1=4 categories depicted

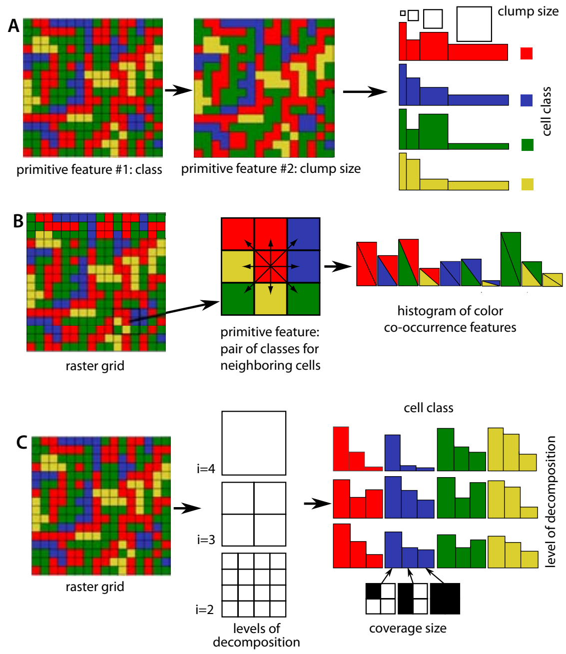

Figure 3: Three of the signature methods implemented in GeoPAT: (A) crossproduct, (B) co-occurrence, (C) decomposition.

as different colors) and clump size (N_2=4 categories depicted as increasing size squares). In this example the crossproduct histogram has 16 bins, the value of each bin is a percentage of cells having specified cell class and specified clump size. Crossproduct signature is designed to be effective in encapsulating spatial structures with clear geometric quality (having relatively low complexity); an example of a dataset with such a structure is the land cover raster. In our pattern-based analysis of the NLCD (Jasiewicz et al., 2013; Stepinski et al., 2014; Netzel and Stepinski, 2015) we used the crossproduct signature with two features (cell class with 16 categories and clump size with 14 categories) paired with the Jensen-Shannon divergence distance function (see section 3.2).

co-occurrence – This method uses a color co-occurrence histogram (Barnsley and Barr, 1996; Chang and Krumm, 1999), a variant of the Gray-Level Co-occurrence Matrix (GLCM) originally introduced by Haralick et al. (1973) to characterize texture in grayscale images. In GeoPAT, color is replaced by cell class and a single cell separation of one pixel is used to calculate a co-occurrence histogram. This results in a single primitive feature - a pair of classes assigned to two neighboring cells; eight-connectivity is assumed for establishing the existence of a neighborhood relationship between the two cells. Thus, eight features are calculated for each cell, but their total number is halved as the same feature is generated twice by the pairs of neighboring cells. For a scene with k cell classes, the co-occurrence histogram has $(k^2 + k)/2$ bins, k of them correspond to same-class pairs, which measure the composition of the classes in the scene, and $(k^2 - k)/2$ bins correspond to different-class pairs, which measure the configuration of the classes in the scene. Fig. 3B shows schematically the construction of a co-occurrence histogram for a scene with k=4 classes resulting in a histogram with ten bins. The co-occurrence signature is designed to be effective in encapsulating spatial structures exhibiting high complexity patterns like the ones resulting from a geomorphons-based classification of a DEM (see Fig. 1). In our pattern-based analysis of DEM data classified to k=10 landform classes (Jasiewicz et al., 2014) we used the co-occurrence signature with 55 bins paired with the Wave Hedges distance function (see section 3.2).

decomposition – This signature method is inspired by the work of Remmel and Csillag (2006) to describe a scene using a set of sub-scenes having a hierarchy of sizes. For the decomposition method to work best the scene should be a square having a linear size of 2^D cells. The scene with k cell classes is scanned without overlap by a series of square moving windows with sizes $w = 2^i$ cells where $i = 2, \ldots, D$ are the decomposition levels. The size of the maximum scanning window, 2^D, is the size of the scene. At the smallest decomposition level $i = 2$ a scene is scanned by a window having a size of $4 \times 4 = 16$ cells. At each scanning position the percentages, $p_1, \ldots, p_k$, of the window's area occupied by cells having classes $1, \ldots, k$, respectively are recorded and a window area is assigned a list of k tags (one for each class) representing those percentages. These tags are classified into one of three categories, 1 if the percentage is below $1/4$, 2 if it is between $1/4$ and $1/2$, and 3 if it is above $1/2$. Tallying all tags results in a histogram with $3 \times k$ bins (three bins for each class).

For example, for a scene having a size of 16×16 cells and k=4 classes (see Fig. 3C) the number of tags for decomposition level $i = 2$ is $16 \times 4 = 64$ (number of sub-windows×number of classes). These tags are histogrammed into 12 bins (number of classes×number of tag categories). If, for example, the entire scene is occupied by only one class (say, red), eight bins are equal to 0 and 4 bins (red-3, blue-1, green-1, and yellow-1) have 16 tags each. In this method tags are the primitive features. Repeating the same procedure for remaining decomposition levels results in $D-1$ histograms each having $3 \times k$ bins. All these histograms can be concatenated into a histogram of length equal to $3 \times k \times (D-1)$. Fig. 3C shows schematically a construction of the decomposition histogram for a scene of size 16×16 cells and k=4 classes. The size of the scenes dictates the maximum level of decomposition D=4 and the histogram length equal to 36. The decomposition signature is designed to be effective for patterns of all levels of complexity, however, we have not yet accumulated sufficient experience working with this signature to offer definitive advice on the types of datasets to which it can be best applied.

3.2. Distance functions

Distance, which assesses the degree of dissimilarity between two scenes, is the opposite of similarity. The input to all distance functions implemented in GeoPAT is a pair of normalized (the sum of all bins adds to 1) signature histograms P and Q and the output is a real number assessing the dissimilarity (distance) between those histograms. When the value of distance function is equal to zero identical histograms are indicated, and thus scenes have identical or very similar patterns, whereas large values of the distance function indicate very different histograms and scenes having significantly different patterns. Note that all histogram distance measures are heuristic and no single measure will work well with all signatures. Of over 40 possible histogram distance measures (Cha, 2007) GeoPAT implements the three methods described below which work well with signatures described in the previous subsection. All three measures have a range of possible values limited to an interval between 0 and 1.

Jensen-Shannon divergence – This measure (Lin, 1991) expresses the informational distance between two histograms P and Q by calculating a deviation between the Shannon entropy of the mixture of the two histograms $(P+Q)/2$ (the second term in eq.(1) below) and the mean of their individual entropies (the second term in eq.(1)). The value of the Jensen-Shannon divergence is given by the following formula,

$$d_{JSD} = \sqrt{\sum_{i=1}^{d} \left[\frac{P_i log_2 P_i + Q_i log_2 Q_i}{2} - \left(\frac{P_i + Q_i}{2} \right) log_2 \left(\frac{P_i + Q_i}{2} \right) \right]} \qquad (1)$$

where d is the number of bins (the same for both histograms) and P_i and Q_i are the values of ith bin in the two histograms. We have found Jensen-Shannon divergence works well (yields dissimilarity values in agreement with human visual perception) for comparison of land cover patterns as encapsulated by crossproduct signatures (Jasiewicz et al., 2013; Stepinski et al., 2014; Netzel and Stepinski, 2015).

Wave Hedges – This measure is designed to work with co-occurrence signature histograms that tend to be dominated by bins corresponding to adjacent cells having the same class. The value of the Wave Hedges distance is given by the following formula (Cha, 2007),

$$d_{WH} = \sum_{i=0}^{d} e_i \frac{|P_i - Q_i|}{max(P_i, Q_i)} \qquad (2)$$

where d is the maximum number of possible bins and $e_i = 1$ if $max(P_i, Q_i) > 0$ or $e_i = 0$ otherwise. In other words, only pattern features present in at least one of the two scenes contribute to the value of the distance. In Wave Hedges distance formula all present features contribute to the overall value of distance with the same weights regardless of feature abundance in the scenes. In the case of the co-occurrence histogram (Fig. 3B) this means that composition-related features and configuration-related features contribute equally to the distance value despite the heavy dominance of composition-related features in histograms stemming from all realistic scenes. This makes the Wave Hedges distance particularly suitable for comparison of terrain scenes as encapsulated by co-occurrence signatures (Jasiewicz et al., 2014)

Jaccard– This measure is an extension of the Jaccard similarity coefficient (Jaccard, 1908), originally developed to assess a similarity between two sets, but used here for assessing the dissimilarity between two histograms. The value of the Jaccard distance is given by the following formula (Cha, 2007),

$$d_J = 1 - \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} \qquad (3)$$

We have found that the Jaccard distance works well for comparison of land cover patterns as encapsulated by decomposition signatures.

4. Core modules

The seven core modules in the GeoPAT toolbox provide the infrastructure for pattern-based analysis of spatial datasets. There are two types of core modules: signature extraction modules and geoprocessing modules. Fig. (4) illustrates different possible pipelines of data processing using these modules.

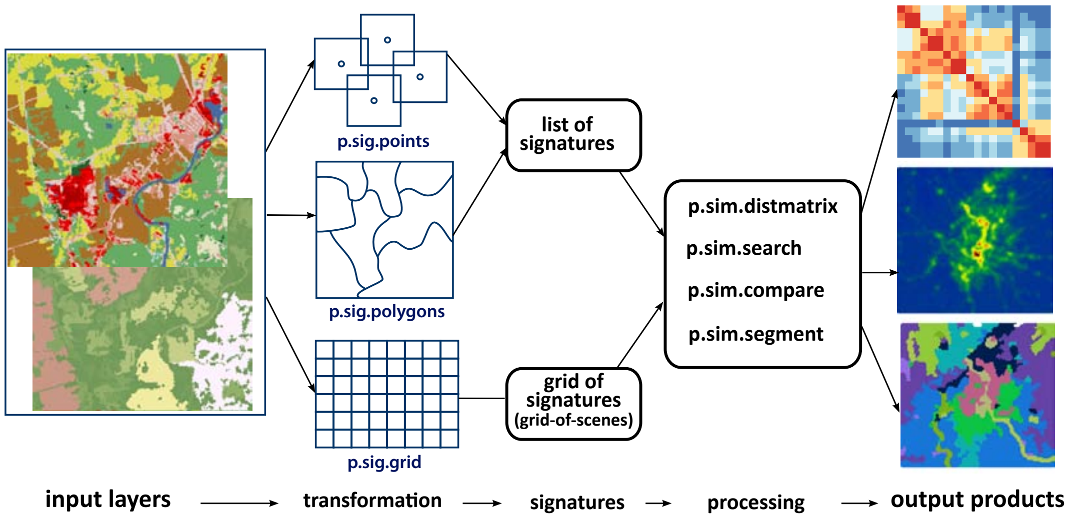

Figure 4: Different data processing pipelines possible using modules from the GeoPAT toolbox. Input consists of raster layers and output consists of raster layers or text tables.

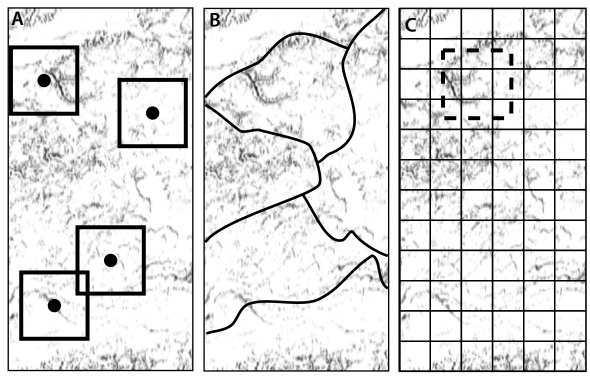

Figure 5: Methods of scene definition: A – scenes defined by points; B – scenes defined by polygons; C – scenes defined by a grid.

4.1. Signature extraction modules

The input to all three signature extraction modules is a set of categorical rasters containing all information layers needed to construct scenes signatures. The output is a set of signatures; each module outputs these signatures in a different data structure depending on its definition of a scene or a set of scenes. Fig. 5 shows three possible scenarios for scene definition which are addressed by the three modules.

p.sig.points. This module extracts signatures for a collection of individual scenes having a square geometry (see Fig. 5A). The user provides the coordinates of the center of each scene (point file) and the size of the scene. The module outputs a list of scene-labeled signatures. Note that p.sig.points can be used to extract a signature for the entire region if needed. There are several typical uses for this module. Examples include generating a query scene to be compared with a grid-of-scenes for the search task, comparison of several scenes in a single raster (like in a comparison of different cities on the basis of their patterns of land cover classes), and comparison of two different co-registered rasters (like in comparison of a natural scene with a scene resulting from a computer simulation aimed at recreating a pattern observed in the natural scene).

p.sig.polygons. This module extracts signatures for a collection of individual scenes having a polygonal geometry (see Fig. 5B). The user provides as input a categorical raster layer which defines the division of a region into polygonal scenes and the module outputs a list of polygon-labeled signatures. A typical use for p.sig.polygons is for comparing irregular scenes resulting from a segmentation of the region (using the segmentation module p.sim.segment, see the next subsection).

p.sig.grid. This module extracts a grid-of-scenes (see Fig. 5C) – a grid of the same spatial extent as the region defined by the input data but having larger cells (s-cells). Each s-cell has only one attribute - a signature of the scene centered on it. The module outputs a header file containing the topology of the grid-of-scenes and a binary file containing signatures ordered row by row. The grid-of-scenes is an input to three geoprocessing modules: p.sim.search, p.sim.segment and p.sim.compare.

4.2. Geoprocessing modules

The four geoprocessing modules use scene signatures and distance functions to perform four popular analysis tasks including: comparison of individual scenes (p.sim.distmatrix), comparison between a single scene (a query) and grid-of-scenes (p.sim.search), comparison between two grids-of-scenes (p.sim.compare), and a segmentation of a grid-of-scenes (p.sim.segment). Fig. (6) illustrates the tasks performed by these modules.

p.sim.distmatrix. This module computes a distance matrix between a collection of scenes. It uses signatures output by p.sig.points or p.sig.polygons modules and an appropriate distance function from the library. The resultant distance matrix is typically used as an input for scene clustering (Fig. 6A) which results in discovering structures in the data without guidance from an analyst. Clustering itself is not implemented in GeoPAT nor is it implemented directly in GRASS; we recommend using the hierarchical clustering algorithm implemented in R (R Core Team, 2013) as GRASS is designed to work together with R (Bivand, 2000). The distance matrix generated by p.sim.distmatrix is the only required input to the hierarchical clustering algorithm. An example of p.sim.distmatrix usage would be the clustering of a collection of cities on the basis of the patterns of land cover classes within their boundaries.

p.sim.search. This module performs a query-by-pattern-similarity (QBPS). The input is a query scene (or a list of query scenes) output by p.sig.points or p.sig.polygons modules and a grid-of-scenes (database to be queried) output by the p.sig.grid module. The signature of each query and the signatures in a grid-of-scenes must have the same structure. The module compares query/queries with the database in a scene-by-scene fashion and outputs a layer(s) having the same topology as the grid-of-scenes but containing values of similarity between a query and each scene in the grid-of-scenes. The results of QBPS can be visualized as a similarity map (Fig. 6B).

QBPS provides a knowledge discovery tool that is qualitatively different from retrieval of top-matches-to-a-query (Câmara et al., 1996; Datcu et al., 2002; Koperski et al., 2002; Aksoy et al., 2005; Barb and Shyu, 2010) – a standard approach to searching for similar scenes. QBPS is a GIS tool inasmuch as it performs spatial processing resulting in a map that shows geographical distribution of degree of similarity to the query scene. Such map provides much more information than a non-spatial list of top matches to a query. By utilizing spatial organization it simultaneously shows similarity relations between the query and *all* scenes in the database. Thus it allows an analyst to concentrate on revealed geospatial phenomena rather than on similarity between specific scenes. QBPS has been used to query the NLCD 2006 dataset (Jasiewicz and Stepinski, 2013a; Stepinski et al., 2014) for similarity between land cover scenes and to query topography of the country of Poland (Jasiewicz et al., 2014) for similarity between landscapes. Note that QBPS can be used as an element of supervised classification of a region into different pattern types such as, for example, urban structures, landscape types, or physiographic units (see the next section).

p.sim.compare. This module compares two grids-of-scenes in a scene-by-scene fashion. The grids must have the same topologies and the output of p.sim.compare is a raster layer having the same topology as the inputs but containing values of similarity between corresponding pairs of scenes (Fig. (6C)). This is equivalent to the GIS *overlay* function, but is performed on signature attributes. Applying p.sim.compare to two land cover datasets of the same region but pertaining to two different time steps enables pattern-based change detection (Netzel and Stepinski, 2015). Unlike traditional land cover change detection which tracks cell-by-cell transitions of land cover categories, pattern-based change detection assesses change in local patterns of land cover; it is especially useful for continental-scale or global-scale assessments of land cover change. Another application of p.sim.compare is for comparison of two layers created using different parameters. For example, the module can be used for comparison of two classifications of a DEM using the same method and the same target landscape classes but different extraction parameters (Jasiewicz and Stepinski, 2013b). Such a comparison allows a user to see what landscape types are most sensitive to the values of free parameters in their mapping algorithm.

p.sim.segment. This module segments a grid-of-scenes into regions of uniform patterns (Niesterowicz and Stepinski, 2013) (Fig. 6D) in the same fashion as traditional segmentation algorithms segment image (or other rasters) into segments of uniform color and texture. The difference is that p.sim.segment segments a grid-of-scenes rather than an ordinary grid, and uses scene signature as attribute rather than color or image texture to decide how to delineate the segments. The module uses a variant of the region growing algorithm (Zucker, 1976; Câmara et al., 1996; Li and Narayanan, 2004; Blaschke, 2010). It has two free parameters: a similarity threshold that defines the minimum similarity between two scenes to be treated as “similar” and, optionally, a minimum number of scenes to constitute a

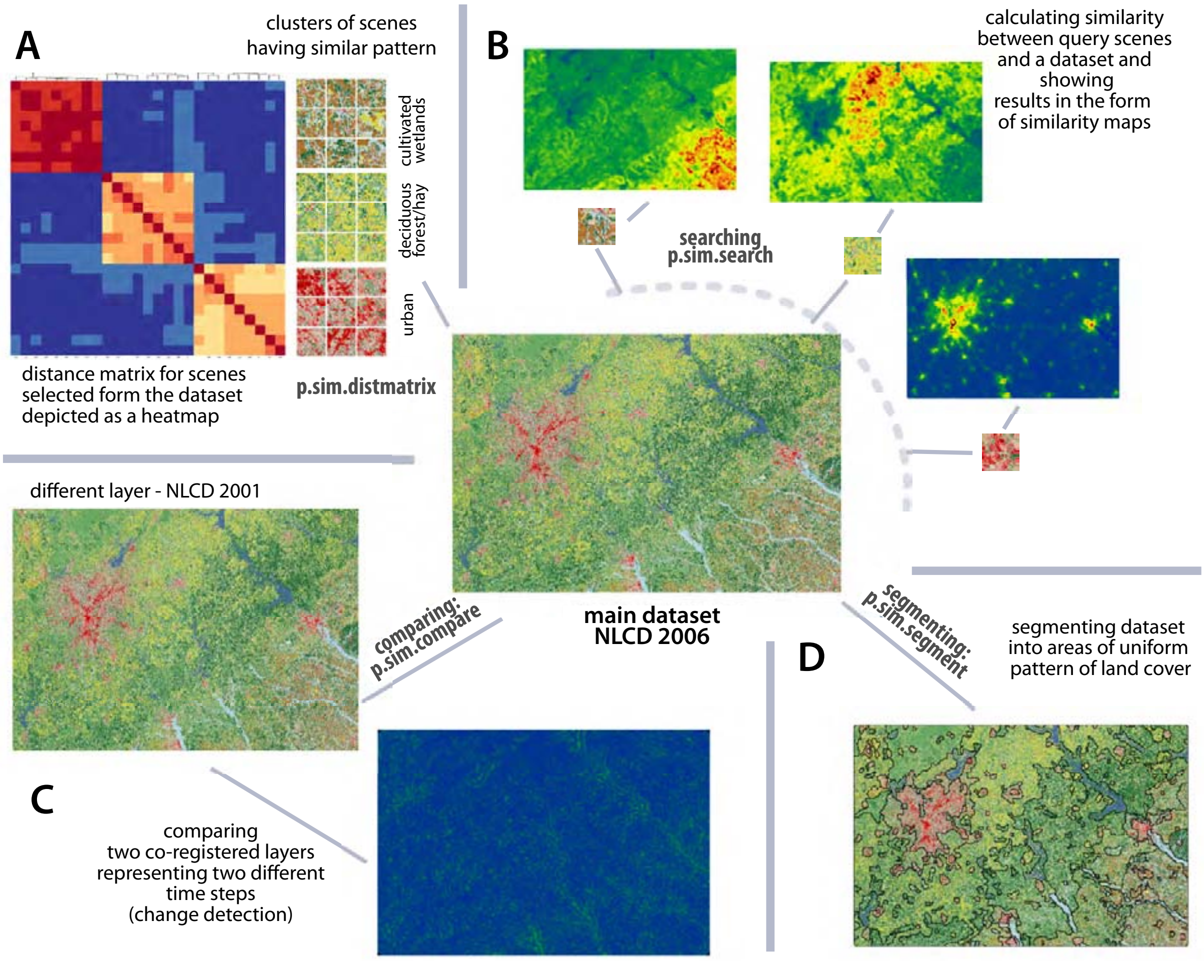

Figure 6: Illustration of data processing, see text for details

separate segment.

The segmentation provided by the p.sim.segment module is typically used as an intermediate step in regionalization of the dataset. The goal of regionalization is to generalize and thus simplify spatial representation of data so it is more meaningful and easier to analyze. Examples of regionalization include delineation of landscape types (Niesterowicz and Stepinski, 2013) or delineation of physiographic units (Jasiewicz et al., 2014). Regionalization is achieved by clustering segments output by p.sim.segment into a number of distinct pattern type classes.

5. Case study

To demonstrate GeoPAT's capabilities in application to a specific dataset we use a region extracted from the NLCD 2006 referred to as "Atlanta." Atlanta covers a region in the northern part of the U.S. state of Georgia (see Fig. 8A) that includes the city of Atlanta. The grid has a size of 11300 x 7500 cells and a resolution of 30m/cell. It is a categorical grid with 16 land cover classes. The data can be downloaded from http://sil.uc.edu. NLCD is the result of classification of Landsat images, so this example pertains to image data. Working with the NLCD avoids performing the pre-processing step of pixel-based classification, which is not a part of GeoPAT toolbox.

The purpose of this case study is to perform unsupervised and supervised regionalizations of the Atlanta site into landscape types (characteristic patterns of land cover) using GeoPAT modules and R. Two different modes of machine learning (unsupervised and supervised) are demonstrated to show the range of tasks that can be achieved using GeoPAT. The schema of the two procedures are shown in Fig. 7. These procedures can be run as a single routine due to the full integration between GRASS 7 and R (Bivand et al., 2008). These routines are available as supplementary material to this paper (http://sil.uc.edu). Based on our earlier experience in working with the NLCD (Jasiewicz and Stepinski, 2013a; Stepinski et al., 2014; Netzel and Stepinski, 2015) we use the crossproduct signature for scene representation and the Jensen-Shannon divergence as a distance function to measure dissimilarity between the scenes. The first step in both procedures is to generate the grid-of-scenes using the p.sig.grid module. We have selected the grid-of-scenes to have a s-cell equal to 900m (30 times the size of cell in the Atlanta grid) and we define scenes as square regions having the size of 4.5 km×4.5 km. Thus, the scenes overlap significantly.

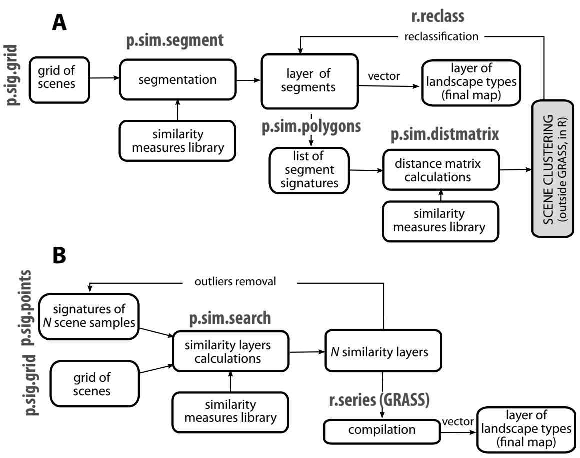

Figure 7: Schemes of processing pipelines for case study calculations: A) unsupervised regionalization; B) supervised regionalization

5.1. *Unsupervised regionalization*

We start by running the p.sim.segment module with the following parameters: similarity threshold=0.75, minimum number of scenes=10. This yielded 434 individual segments. Signatures of the segments were calculated using the p.sim.polygons module. Using these signatures and the Jensen-Shannon divergence we cluster the 434 segments into seven “landscape types” using hierarchical clustering algorithm with ”Ward” linkage (available in R “stats” package). The choice of a number of clusters is arbitrary as is always the case in hierarchical clustering. The result is a map of landscape types in the Atlanta site (Fig. 8B). Note that these landscape types emerged from the data and we assign labels to them (see caption to Fig. 8) only a posteriori, after reviewing the results of hierarchical clustering.

In principle, an unsupervised classification could be performed without the intermediate segmentation step. The signatures calculated by the p.sig.grid module could be clustered to yield landscape types, however this would involve clustering over 90,000 scenes and the results would exhibit salt and pepper noise. The procedure demonstrated here employs the concept of object-based analysis – first segment then classify but it is the grid-of-scenes rather than an image which is segmented.

5.2. *Supervised regionalization*

To begin, we selected seven individual scenes (shown as white squares in Fig. 8C) as landscape type archetypes or samples of the seven landscape types we have chosen for mapping; 1 – urban, 2 – wet croplands, 3 – pasture-dominated, 4 – deciduous forest-dominated, 5 – evergreen forest-dominated, 6 – wetlands and surroundings, 7 – waters and surroundings. Signatures for the seven sample scenes were calculated using the r.sig.points module. These signatures were used as queries over the Atlanta grid-of-scenes using the p.sim.search module resulting in seven similarity maps, one for each query. The maps were overlaid (using GRASS capabilities) and each s-cell was assigned a landscape type label corresponding to the largest value of similarity resulting in a single map of landscape types (Fig. 8D).

It is interesting to observe similarities and differences between the two maps. Differences are expected because the two maps were obtained following very different principles. The unsupervised map on Fig. 8B reflects the natural grouping of landscapes subject to a restriction on the total number (in our case – seven) of the groups. The supervised map on Fig. 8D reflects the preferences of an analyst who has selected a priori specific landscape types to be mapped. The similarities between the two maps stem from the fact that both of them reflects the same physical reality.

The two processing schemes demonstrated here closely resemble standard schemes for unsupervised and supervised image classifications. This is intentional as GeoPAT is designed to work like a standard GIS system but with scenes rather than pixels. However, the output of GeoPAT (see, for example, Figs. 8B and D) is fundamentally different from the output of an image classification algorithm as it yields pattern-based generalization of pixel-based classification.

The unsupervised example illustrates the concept of generalization that has been previously applied (using different methods) to land cover datasets in the context

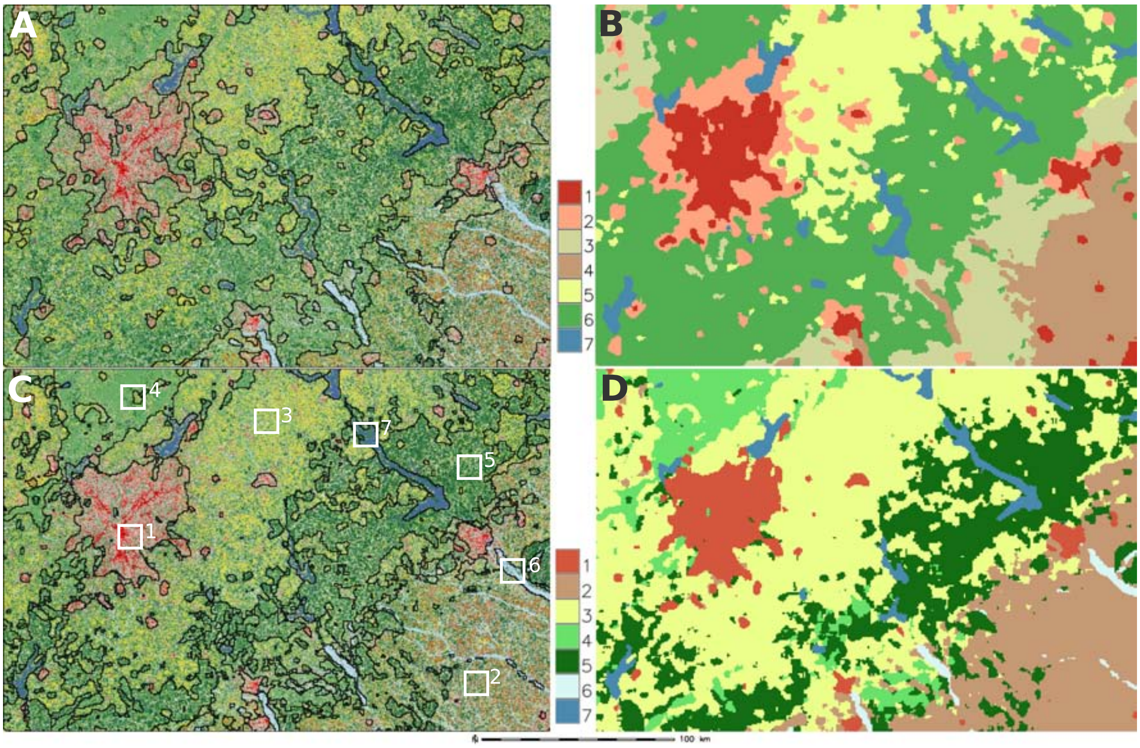

Figure 8: Results of the case study. Boundaries of delineated landscape types obtained using unsupervised (A) and supervised (C) regionalization procedures superimposed on the NLCD map (see http://www.mrlc.gov/ for the land cover legend). Regionalization maps resulting from unsupervised (B) and supervised (D) classification, respectively. Classes of unsupervised regionalization are: 1 – urban, 2 – suburban, 3 – pasture-forest mixture, 4 – wet croplands, 5 – pasture-dominated, 6 – forests dominated, and 7 – waters and surroundings. Classes of supervised classification are: 1 – urban, 2 – wet croplands 3 – pasture-dominated, 4 – deciduous forest-dominated, 5 – evergreen forest-dominated, 6 – wetlands and surroundings, and 7 – waters and surroundings. Scene samples utilized in supervised approach are marked by white squares.

of landscape ecology (Long et al., 2010; Cardille and Lambois, 2009) and the supervised example illustrates a generalization concept that has been previously applied (using different methods) to RS images in the context of mapping urban landscapes (Moller-Jensen et al., 2005; Vatsavai, 2013a; Graesser et al., 2012). With GeoPAT these kinds of generalizations can be performed with relative ease and on much larger datasets by taking advantage of GRASS' ability to handle very large datasets.

6. Performance

GeoPAT has been optimized to work efficiently with big data where it is most effective as a knowledge discovery tool. It can be applied to giga-cell rasters when running on either servers or workstations. Table 1 lists execution times for GeoPAT modules as applied to several large datasets. This gives a rough idea about GeoPAT's level of performance. All calculations were run on a double-CPU XEON machine (8 cores each) with 20 GB of RAM running Linux. Three datasets were used: (1) The POLAND dataset ($24,000 \times 27,000$ cells) is a 10-classes map of landform elements calculated from a 30m/cell DEM (Jasiewicz and Stepinski, 2013b); (2) The CHINA dataset ($84,000 \times 64,000$ cells) is a 10-class map of landform elements calculated from a 90m/cell DEM (SRTM); (3) The USA dataset ($164,000 \times 104,000$ cells) is a 16-class, 30m/cell map of land cover/land use (NLCD 2006) covering the entire conterminous United States.

Table 1 is divided into two parts, part A pertains to the performance of the signature extraction modules and part B pertains to the performance of the geoprocessing modules. In general, the signature extraction modules are significantly more computationally expensive than geoprocessing modules. However, in a typical application signature extraction needs to be performed only once, whereas geoprocessing computation may require multiple runs in order to get a satisfactory result.

Note that the p.sim.search module is fast enough to enable real-time search. Indeed, this module provides a computational engine for our two GeoWeb pattern

module	input	output	processing time
A. Signature extraction modules			
p.sig.points with co-occurrence function	single scene 300×300 cells with 10 categories	single signature with 55 bins	0s 73ms
p.sig.grid with co-occurrence function	POLAND	800×900 grid-of-scenes	2h 26m
p.sig.grid with co-occurrence function	CHINA	1680×1280 grid-of-scenes	20h 39m
p.sig.grid with crossproduct function parallelized with 10 threats	USA	1640×1040 grid-of-scenes	4h 46m
B. Geoprocessing modules			
p.sim.distmatrix with Wave-Hedges function	1084 histograms with 136 bins each	1084×1084 distance matrix	48s
p.sim.search, 55-bins histograms, Wave-Hedges function	64 queries, POLAND grid-of-scenes	64 800×900 similarity layers	5s 33ms
p.sim.search, 192-bins histograms, Jensen-Shannon function	1 query, USA grid-of-scenes	one 1640×1040 similarity layer	6s 11ms
p.sim.compare, 55-bins histograms, Wave-Hedges function	two POLAND grids-of-scenes	one 800×900 similarity layer	0s 97ms
p.sim.segment, 55-bins histograms, Wave-Hedges function	one POLAND grid-of-scenes	one 800×900 layer containing segments	1s 28ms

Table 1: Examples of computation times for different modules of GeoPAT and using different datasets.

search applications, LandEx-USA – which enables discovery of similar land cover patterns over the extent of the United States and TerraEx-PL – which enables discovery of similar landscapes over the extent of the country of Poland. These online applications are available at http://sil.uc.edu/. The real-time response to a query in these applications is achieved by pre-calculating grids-of-scenes and storing them in the RAM. Once a query is submitted by a user the application runs p.sim.search and returns the generated similarity map.

7. Discussion and conclusions

GeoPAT is a toolbox which implements our method (Jasiewicz and Stepinski, 2013b; Stepinski et al., 2014; Jasiewicz et al., 2014; Netzel and Stepinski, 2015) for pattern-based information retrieval from images and other rasters. The recent interest in such methods stems from the need to consider spatial patches larger than a pixel to adequately reflect local content. To the best of our knowledge, the only current methodological (but not software) alternative to GeoPAT is the Complex Object-Based Image Analysis (COBIA) (Vatsavai, 2013a,b) which relies on the MIL concept. At the conceptual level GeoPAT and COBIA are similar inasmuch as both use scenes, both describe a scene as a Probability Distribution Function (PDF) of image features, and both use measures of similarity between PDFs to asses a degree of similarity between the scenes. The differences are in an the implementation of this concepts. COBIA works directly with images and assumes that the PDF of image features can be modeled by a multi-variate Gaussian. GeoPAT works with previously classified images and models PDFs as histograms. Both of these methods differ from earlier approaches (Moller-Jensen et al. (2005) or Lucieer and Stein (2005)) to improve classification and segmentation of images by incorporating texture descriptors as additional features in pixel-based (single-instance learning) algorithms.

The decision to design GeoPAT to operate on categorical rasters was dictated by two considerations: (a) the flexibility of input data, so different modalities of raster data could be analyzed by GeoPAT once classified, and (b) the use of categorical rasters increases the performance of an algorithm and make it easier to work with large datasets. Note that many large datasets of interest are already available in the categorized (land cover) form. These include 30m U.S.-wide NLCD, 30m global GLC30 (http://www.globallandcover.com/), 300m global GlobCover (http://due.esrin.esa.int/globcover/), and 100m Europe-wide CORINE. In addition, topographic datasets (DEMs) can be categorized with relative ease by using robust techniques, such as the geomorphons method (Jasiewicz and Stepinski, 2013b) for which an open source code as well as an online application are available at http://sil.uc.edu/. Additionally, RGB VHR images can be categorized by clustering colors with Euclidean distance in CIE-Lab color space (Rubner et al., 2000).

Given a dataset, what combination of signature/distance function are most appropriate? There is no quantitative means to assist in answering this question directly. Each combination is based on different set of heuristics and will yield different values of similarities for the same pairs of scenes. These values can be compared (Stepinski and Cohen, 2014) by using their corresponding percentiles (calculated from the empirical cumulative distribution functions of the sets of similarities between all scenes in the dataset), but it is not possible to determine objectively which similarity value is "better" as this invokes subjective human perception of similarity between spatial patterns.

However, in applications of GeoPAT to classification tasks, the utility of a particular signature/distance function combination can be assessed indirectly by assessing quality of the classification. When using the GeoPAT for supervised learning, the quality of a classifier can be assessed by applying the classifier to a test set of pre-labeled scenes and calculating the standard metrics of performance. The results will depend on the selection of a particular signature/distance function combination. Note that when a region under consideration has clearly visible divisions between different pattern types, like in the case of the DEM shown in Fig. 1 or the image used by Vatsavai (2013a), all possible classifiers will perform very well as the problem presents little challenge to an algorithm. On the other hand, regions on which an analyst would have difficulty in delineating the boundaries between different types of patterns, like in the case of Atlanta region, present a challenge. First, it is difficult-to-impossible to manually delineate different pattern types in such regions (to construct a test set), second, different classifiers (different combinations of signature/similarity function) will yield different results without an objective means to assess their quality due to lack of a reliable test set. However, in our opinion, such "difficult" regions are where GeoPAT is most useful as it retrieves information which could be difficult to retrieve by any other means. To make this point clearer consider the Atlanta region as shown in the middle of Fig. 6. It would be very difficult for an analyst to manually delineate pattern types in this region and no two manual delineations would be the same, but GeoPAT delineations (Fig. 8) are objective, repeatable, and they make perfect sense to an analyst once the computed boundaries are superimposed on the land cover map (Fig. 8 A and C).

Finally, GeoPAT is an actively developed software and we expect users to contribute to it by adding to the shared library of functions. GeoPAT is available for download from http://sil.uc.edu/. Currently, it has been tested on and made available for the Linux operating system; it requires the development version of GRASS 7.

Acknowledgments

The authors wish to thank J. Niesterowicz and A. Dmowska for helpful comments and discussions. The work was supported by the National Science Centre (NCN) under Grant DEC-2012/07/B/ST6/01206, and by the National Science Foundation (NSF) under Grant BCS- 621 1147702, and by the University of Cincinnati Space Exploration Institute.

References

Aksoy, S., Koperski, K., Tusk, C., Marchisio, G., Tilton, J., Mar. 2005. Learning bayesian classifiers for scene classification with a visual grammar. IEEE Trans. Geosci. Remote Sens. 43 (3), 581–589.

Barb, A., Shyu, C., 2010. Visual-Semantic Modeling in Content-Based Geospatial Information Retrieval Using Associative Mining Techniques. IEEE Geosci. Remote Sens. Lett. 7 (1), 38–42.

Barnsley, M., Barr, S., 1996. Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. Photogrammetric engineering and remote sensing 62 (8), 949–958.

Bivand, R. S., 2000. Using the r statistical data analysis language on grass 5.0 gis database files. Computers & Geosciences 26 (9), 1043–1052.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., 2008. Applied spatial data analysis with R. Vol. 747248717. Springer.

Blaschke, T., Jan. 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65 (1), 2–16.

Cain, D., Riitters, K., 1997. A multi-scale analysis of landscape statistics. Landsc. Ecol. (630), 199–212.

Câmara, G., Souza, R. C. M., Freitas, U. M., Garrido, J., May 1996. Spring: Integrating remote sensing and gis by object-oriented data modelling. Comput. Graph. 20 (3), 395–403.

Cardille, J. A., Lambois, M., 2009. From the redwood forest to the Gulf Stream waters: human signature nearly ubiquitous in representative US landscapes. Frontiers in Ecology and the Environment 8(3), 130–134.

Cardille, J. A., White, J. C., Wulder, M. A., Holland, T., 2012. Representative landscapes in the forested area of Canada. Environmental management 49(1), 163–173.

Cha, S., 2007. Comprehensive survey on distance/similarity measures between probability density functions. Int. J. Math. Model. Methods Appled Sci. 1(4) (4), 300–307.

Chang, P., Krumm, J., 1999. Object recognition with color cooccurrence histograms. In: In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, June 23-25, 1999. IEEE Computer Society Conference.

Cushman, S. A., McGarigal, K., Neel, M. C., 2008. Parsimony in landscape metrics: strength, universality, and consistency. Ecological indicators 8(5), 691–703.

Daschiel, H., Datcu, M., Jan. 2005. Information mining in remote sensing image archives: system evaluation. IEEE Trans. Geosci. Remote Sens. 43 (1), 188–199.

Datcu, M., Daschiel, H., Pelizzari, a., Quartulli, M., Galoppo, a., Colapicchioni, a., Pastori, M., Seidel, K., Marchetti, P., D'Elia, S., Dec. 2003. Information mining in remote sensing image archives: system concepts. IEEE Trans. Geosci. Remote Sens. 41 (12), 2923–2936.

Datcu, M., Seidel, K., D'Elia, S., Marchetti, P., 2002. Knowledge-driven information mining in remote-sensing image archives. ESA Bull. 110 (may), 26–33.

Datta, R., Joshi, D., Li, J., Wang, J. Z., Apr. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Comput. Surv. 40 (2), 1–60.

Dilts, T. E., Yang, J., Weisberg, P. J., 2010. The landscape similarity toolbox: new tools for optimizing the location of control sites in experimental studies. Ecography 33(6), 1097–1101.

Gevers, T., Smeulders, A. W., 2004. Content-based image retrieval: An overview. In: Kang, G. M. S. B. (Ed.), Emerging Topics in Computer Vision. Upper Saddle River, NJ: Prentice-Hall, Ch. 8, pp. 333–384.

Graesser, J., Cheriyadat, A., Vatsavai, R., Chandola, V., Long, J., Bright, E., 2012. Image based characterization of formal and informal neighborhoods in an urban landscape. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5(4), 1164–1176.

GRASS Development Team, 2012. Geographic Resources Analysis Support System (GRASS GIS) Software. Open Source Geospatial Foundation, USA.

Haines-Young, R., Chopping, M., 1996. Quantifying landscape structure: a review of landscape indices and their application to forested landscapes. Progress in Physical Geography 20(4), 418–445.

Haralick, R. M., Shanmugam, K., Dinstein, I., Nov. 1973. Textural features for image classification. Syst. Man Cybern. IEEE Trans. 3 (6), 610–621.

Jaccard, P., 1908. Nouvelles recherches sur la distribu- tion florale. Bull. Soc. Vaudoise Sci. Nat. 44, 223–270.

Jasiewicz, J., Netzel, P., Stepinski, T. F., 2014. Landscapes similarity, retrieval, and machine mapping of physiographic units. Geomorphology 221, 104–112.

Jasiewicz, J., Stepinski, T., Netzel, P., 2013. Content-based landscape retrieval using geomorphons. In: Geomorphometry 2013. Nanjing, China, pp. 1–4.

Jasiewicz, J., Stepinski, T. F., 2013a. Example-Based Retrieval of Alike Land-Cover Scenes From NLCS2006 Database. IEEE Geosci. Remote Sens. Lett. 10 (1), 155–159.

Jasiewicz, J., Stepinski, T. F., 2013b. Geomorphons-a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156.

Koperski, K., Marchisio, G., Aksoy, S., Tusk, C., 2002. VisiMine: interactive mining in image databases. In: IEEE Int. Geosci. Remote Sens. Symp. Vol. 3. Ieee, pp. 1810–1812.

Körting, T., Fonseca, L. G., Câmara, G., 2013. GeoDMAGeographic Data Mining Analyst. Comput. Geosci. 57, 133–145.

Kupfer, J. A., Gao, P., Guo, D., May 2012. Regionalization of forest pattern metrics for the continental United States using contiguity constrained clustering and partitioning. Ecol. Inform. 9, 11–18.

Lang, S., 2008. Object-based image analysis for remote sensing applications: modeling reality–dealing with complexity. Object-Based Image Anal., 3–27.

Lew, M., Sebe, N., Lifi, C., Jain, R., 2006. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput., Commun., Appl., 2(1), 1–19.

Li, J., Narayanan, R., 2004. Integrated spectral and spatial information mining in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 42 (3), 673–685.

Lin, J., 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 31(1), 145–151.

Long, J., Nelson, T., Wulder, M., Jul. 2010. Regionalization of landscape pattern indices using multivariate cluster analysis. Environ. Manage. 46 (1), 134–42.

Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28(5), 823–870.

Lucieer, A., Stein, A., 2005. Texture-based landform segmentation of LiDAR imagery. International Journal of Applied Earth Observation and Geoinformation 6(3), 261–270.

McGarigal, K., Cushman, S. A., Neel, M. C., Ene, E., 2002. FRAGSTATS: spatial pattern analysis program for categorical maps. Tech. rep.

Moller-Jensen, L., Kofie, R. Y., Yankson, P., 2005. Large-area urban growth observationsa hierarchical kernel approach based on image texture. Geografisk Tidsskrift-Danish Journal of Geography 105, no. 2 (2005): 105(2), 39–47.

Netzel, P., Stepinski, T. F., 2015. Pattern-based assessment of land cover change on continental scale with application to NLCD 2001-2006. IEEE Transactions on Geoscience and Remote Sensing 53(4), 1773–1781.

Niesterowicz, J., Stepinski, T. F., Dec. 2013. Regionalization of multi-categorical landscapes using machine vision methods. Appl. Geogr. 45, 250–258.

Pontius, R. G., 2002. Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. Photogrammetric Engineering & Remote Sensing 68(10), 10411049.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Remmel, T. K., Csillag, F., 2006. Mutual information spectra for comparing categorical maps. International Journal of Remote Sensing 27(7), 14251452.

Richards, J. A., 1999. Remote sensing digital image analysis. Springer-Verlag, Berlin.

Rubner, Y., Tomasi, C., Guibas, L. J., 2000. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40(2), 99–121.

Shyu, C., Klaric, M., Scott, G., 2007. GeoIRIS: Geospatial information retrieval and indexing system - Content mining, semantics modeling, and complex queries. IEEE Trans. Geosci. Remote Sens. 45 (4), 839–852.

Steiniger, S., Hay, G. J., 2009. Free and open source geographic information tools for landscape ecology. Ecological Informatics 4(4), 183–195.

Stepinski, T., Netzel, P., Jasiewicz, J., 2014. LandEx - A GeoWeb tool for query and retrieval of spatial patterns in land cover datasets. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(1), 257–266.

Stepinski, T. F., Cohen, J. P., 2014. Comparing semantically-blind and semantically-aware landscape similarity measures with application to query-by-content and regionalization. Ecological Informatics 24, 69–77.

Uuemaa, E., Antrop, M., Roosaare, J., Marja, R., Mander, b., 2009. Landscape metrics and indices: an overview of their use in landscape research. Living Rev. Landsc. Res. 3 (1), 1–28.

Vatsavai, R. R., 2013a. Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery. In: In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 1419–1426.

Vatsavai, R. R., 2013b. Object based image classification: state of the art and computational challenges. In: In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. ACM, pp. 73–80.

Williams, E. A., Wentz, E. A., Apr. 2008. Pattern Analysis Based on Type, Orientation, Size, and Shape. Geogr. Anal. 40 (2), 97–122.

Zucker, S. W., 1976. Region growing: Childhood and adolescence. Computer graphics and image processing 5 (3), 382–399.