RESEARCH ARTICLE

Unsupervised regionalization of the United States into landscape pattern types

J. Niesterowicz\textsuperscript{a}, T.F. Stepinski\textsuperscript{a,*} and J. Jasiewicz\textsuperscript{a,b}

\textsuperscript{a}Space Informatics Lab, Department of Geography, University of Cincinnati, OH, USA; \textsuperscript{b}Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznan, Poland

(Received 00 Month 200x; final version received 00 Month 200x)

We present a pattern-based regionalization of the conterminous US — a partitioning of the country into a number of mutually exclusive and exhaustive regions that maximizes the intra-region stationarity of land cover patterns and inter-region disparity between those patterns. The result is a discretization of the land surface into a number of Landscape Pattern Types (LPTs) — spatial units each containing a unique quasi-stationary pattern of land cover classes. To achieve this goal, we use a recently developed method which utilizes machine vision techniques. First, the entire National Land Cover Dataset (NLCD) is partitioned into a grid of square-size blocks of cells, called motifels. The size of a motifel defines the spatial scale of a local landscape. The land cover classes of cells within a motifel form a local landscape pattern which is mathematically represented by a histogram of co-occurrence features. Using the Jensen-Shannon divergence as a dissimilarity function between patterns we group the motifels into several LPTs. The grouping procedure consists of two phases. First, the grid of motifels is partitioned spatially using a region-growing segmentation algorithm. Then the resulting segments of this grid, each represented by its medoid, are clustered using a hierarchical algorithm with Ward’s linkage. The broad-extent maps of progressively more generalized LPTs resulting from this procedure are shown and discussed. Our delineated LPTs agree well with the perceptual patterns seen in the NLCD map.

Keywords: automatic regionalization; landscape; pattern analysis; NLCD; large geodata

*Corresponding author. Email: stepintz@ucmail.uc.edu
1. Introduction

Regionalization is the spatial delimitation of natural, social, economic, cultural, or political spheres of reality (Werlen 2009). In other words, it is the process of aggregating a large number of individual geographical objects into a much smaller number of preferably spatially contiguous and internally uniform units, which are more meaningful and easier to analyze (Chorley and Haggett 1967, Johnston 1970). It has been applied to many diverse domains including political science (George et al. 1997), geomorphology (Chai et al. 2009, Jasiewicz et al. 2014), hydrology (Peterson et al. 2011), climatology (Fovell and Fovell 1993, Stooksbury and Michaels 1991), biogeography (Kreft and Jetz 2010, Patten and Smith-Patten 2008), agricultural science (Hollander 2012, Lark 1998), and ecology (Hargrove and Hoffman 2004, Omernik and Griffith 2014).

One of the best known examples of regionalization on a large scale is the division of the United States into physiographic regions by Fenneman (1915). Fenneman used a holistic approach to regionalization; his division is based on years of analyzing multiple aspects of the environment. Current regionalizations of the US into ecoregions are likewise based on a combination of surficial factors (Hargrove and Hoffman 2004, Omernik and Griffith 2014). In this paper we focus on a single aspect of the environment – Land Use/Land Cover (LULC) – but, instead of identifying areas occupied by any single LULC class, we regionalize the US with respect to patterns of LULC classes. Algorithmic, automatic (unsupervised) subdivision of the US into Landscape Pattern Types (LPTs) – spatial units containing unique quasi-stationary patterns of LULC classes (Wickham and Norton 1994) – is the major novelty of this work. We refer to pattern as quasi-stationary to indicate that it does not substantially change across a delineated zone.

Identification and delineation of LPTs in a large area leads to the creation of broad-extent maps of LPTs. Such maps are of significant interest for conservation planning (Long et al. 2010, Cain et al. 1997, Powers et al. 2012) and large-scale land management (Partington and Cardille 2013), as well as in academic research whose goal is to find associations between various environmental variables.

Previous pattern-based regionalizations, such as those by Fenneman (1915), Wickham and Norton (1994), and Omernik and Griffith (2014), are based on manual methodologies. This is a tedious process, especially when conducted on a large spatial scale using relatively high-resolution data. The process is also intrinsically subjective and skewed toward the analyst’s perspective – different analysts will divide a study area differently even if they use the same data. The recent availability of high resolution LULC datasets on continental (NLCD) and even global (Chen et al. 2015) scales makes adopting a manual approach increasingly difficult and instead calls for an automated means of performing regionalizations.

From a computational perspective a regionalization is an unsupervised classification which results in the generalization of the original LULC map into an useful overview across a large area (broad-extent map of LPTs). Most of the existing work on unsupervised quantification and regionalization of LULC patterns focuses on forestry and conservation. In this context, LPTs of binary (forest/non-forest) land cover patterns have been identified and delineated (Gustafson and Parker 1992, Long et al. 2010, Andrew et al. 2012). To the best of our knowledge only two previous studies aimed at unsupervised delineation of LPTs from multi-categorical LULC datasets have been undertaken. Cardille and Lambois (2010) divided the NLCD 1992 into a square grid having cells of size 6.5 km from which they identified 17 different LPTs. However, their study was restricted to the identification of exemplars of these LPTs and no actual regionalization.
map was presented. Partington and Cardille (2013) applied a similar methodology to the 25 m resolution Earth Observation for Sustainable Development of Forests (EOSD) land-cover dataset (Wulder et al. 2008) within the Canadian province of Quebec. They identified and mapped 5 LPTs at a spatial scale of 30 km.

All the aforementioned unsupervised regionalizations used the same approach to assessing the value of dissimilarity between landscapes. This approach consists of representing landscapes as feature vectors of landscape metrics (LMs) (Haines-Young and Chopping 1996, Herzog and Lausch 2001, McGarigal et al. 2002) and calculating the Euclidean distance between these feature vectors to quantify the similarity between two landscapes. Hereafter we refer to such method as the standard approach. Our approach, inspired by the methods developed in the context of machine vision (Gevers and Smeulders 2004, Datta et al. 2008), departs from this standard. Based on our earlier work (Jasiewicz and Stepinski 2013a, Niesterowicz and Stepinski 2013, Jasiewicz et al. 2015) we represent landscapes as normalized histograms of co-occurrence features and we calculate the dissimilarity between two landscapes by calculating the Jensen-Shannon divergence between two histograms. A machine vision-inspired approach is robust as there is no need to combine metrics of different meaning, ranges of values, and varying relevance into a single value of similarity between two landscapes. It is also computationally efficient and thus suitable for application to large datasets such as the entire NLCD.

The aim of this paper is to describe a methodology for the pattern-based regionalization of a very large multi-categorical raster and to apply it to the LULC dataset in order to reveal broad-scale geographical information that can not be objectively obtained by other means. We demonstrate the feasibility of our method by regionalizing the entire NLCD 2011, thus revealing the forms and spatial extents of the most prevalent LPTs within the conterminous US. In addition, by performing the regionalization in a hierarchical fashion we also test the degree to which an algorithm can be relied upon to “understand” perceptual similarities between different LPTs.

The remainder of this paper is structured as follows. Section 2 describes data used, our computational methodology, and the software we have developed and implemented. Section 3 presents the results of applying our methodology to the regionalization of the entire NLCD2011. Section 4 contains a discussion on the usage, abilities, limitations and potential applications of the method.

2. Data, methodology and software

2.1. Data

For input into our regionalization algorithm we use the National Land Cover Database 2011 (NLCD2011) provided by the Multi-Resolution Land Characteristics (MRLC) Consortium. It is a LULC multi-categorical raster map covering the entire conterminous US with a spatial resolution of 30 m resulting in 161,190×104,424 cells. The NLCD is the result of a decision-tree classification of Landsat-7 satellite images (Jin et al. 2013) into 16 land cover categories within the conterminous US (see legend to Fig. 6). All the analyses and presented maps are in the Albers Equal Area projection, the original projection of the NLCD.
2.2. **Methodology**

2.2.1. **General framework**

Our methodology employs several techniques from the domains of data science and machine vision. The core concept is Complex Object-Based Image Analysis (COBIA) (Vatsavai 2013, Stepinski et al. 2015). Two different implementations of the COBIA framework have been independently developed, Vatsavai’s (2013) implementation concentrates on images whereas our implementation (Jasiewicz and Stepinski 2013a, Stepinski et al. 2014, Jasiewicz et al. 2015) concentrates on categorical rasters including land cover datasets. Within the COBIA framework an image or, in our case, the NLCD raster, is divided arbitrarily into a regular grid of local blocks of cells at minimal computational cost. These blocks of cells are the basic units of analysis. While such blocks are geometrically simple (they are square arrays of cells) they hold complex content (patterns of different LULC classes), hence the name – complex objects. Hereafter we will refer to such a block of cells as a *motifel* – the smallest processing element of the grid containing local motifs (patterns) of LULC classes.

COBIA fundamentally differs from a better-known image analysis framework called Object-Based Image Analysis or OBIA (Hay and Castilla 2008, Blaschke 2010). OBIA uses clumps of cells (objects) which are geometrically complex but carry simple content (homogeneous values of cells) as its basic units of analysis. The key difference between the two approaches is in the meaning of their basic processing units. While OBIA units represent particular visible features of the earth’s surface – such as a house, a road, or a particular cultivated type of crops – COBIA units represent the local spatial arrangement of various visible features, thus, the COBIA framework is particularly well-suited to regionalization on the basis of landscape patterns. Furthermore, by its very design, a grid of motifels to be analyzed by COBIA is several orders of magnitude smaller than the original raster making the technique uniquely suited for analysis of large datasets and for generation of broad-scale maps. To regionalize a raster using COBIA a numerical description of motifels and a function that calculates the level of dissimilarity between them are needed.

2.2.2. **Numerical representation of a motifel**

For the numerical description of a motifel we use a histogram of color co-occurrence pattern features (Barnsley and Barr 1996, Chang and Krumm 1999). A co-occurrence feature is a pair of classes assigned to two neighboring cells (Fig. 1E). Features are extracted from a motifel by combining co-occurrence matrices calculated for eight different displacement vectors along principal directions. For a raster with \( k \) possible classes the result is a symmetric matrix which we reduce to a histogram with \( d = (k^2 + k)/2 \) bins. Figs. 1A to 1D show examples of co-occurrence histograms stemming from four different motifels. In this hypothetical case \( k = 4 \) resulting in a co-occurrence histograms with 10 bins. In the case of NLCD2011, \( k = 16 \) and the co-occurrence histogram has 136 bins. A bin in a histogram gives a (normalized; divided by the sum of all bins) number of co-occurrences (either horizontal, vertical or diagonal as shown in Fig. 1E) between two LULC classes. For example, before the normalization the first bin in Fig. 1A had 20 co-occurrences between green pixels in motifel A.

The \( k \) bins correspond to co-occurrence of same-class pairs and their values reflect both, the abundance of the class and its spatial arrangement. For example, blue class have the same abundance in motifels shown in Figs. 1A and B, but the blue-blue bin in the histogram for motifel A is higher than the corresponding bin for motifel B because the
area occupied by the blue class in A is more aggregated. The \((k^2 - k)/2\) bins correspond to co-occurrences between different-class pairs and their values reflect a texture of classes pattern in a motifel. For example, motifels C and D have the same abundance of green and orange classes but different textures. This difference is reflected by their corresponding histograms with the green-orange bin in the histogram for motifel D being much higher than the corresponding bin for motifel C reflecting its finer texture. In general, a large number of same-class co-occurrences indicates the domination of large, class-uniform patches in the motifel, while a large number of inter-class co-occurrences indicates high textural complexity of the motifel. Overall, a co-occurrence histogram describes both content and texture of the motifel in one compact form. Application of color co-occurrence histograms for assessing the level of similarity between blocks of land cover pixels was first proposed by Barnsley and Barr (1996).

### 2.2.3. Dissimilarity measure

Dissimilarity measures assess the likeness of patterns in two motifels by calculating the dissimilarity between the histograms which represent them. There exist a large number of histogram dissimilarity functions (Cha 2007), from which we have selected the Jensen-Shannon Divergence (JSD) (Lin 1991) as it was shown to agree well with the human
visual perception of similarity between natural images (Rubner et al. 2001). The JSD expresses the informational distance between two normalized histograms \(A\) and \(B\) as a deviation between the Shannon’s entropy of the mixture of the two histograms \((A + B)/2\) and the mean entropy of individual histograms \(A\) and \(B\). The value of \(\text{JSD}(A, B)\) is given by the following formula:

\[
\text{JSD}(A, B) = \sqrt{H \left( \frac{A + B}{2} \right) - \frac{1}{2} [H(A) + H(B)]},
\]

where \(H(A)\) indicates a value of the Shannon’s entropy of the histogram \(A\):

\[
H(A) = -\sum_{i=1}^{d} \log_2 A_i.
\]

\(A_i\) is the value of \(i^{th}\) bin in the histogram \(A\), and \(d\) is the number of bins (the same for both histograms). The JSD has been proved (Endres and Schindelin 2003) to be a metric and thus it is well-suited to measure “distance” between landscapes. For normalized histograms the JSD measure always takes values from 0 to 1 with a value of 0 indicating that two motifels are identical, and a value of 1 indicating maximum dissimilarity (none of the classes existing in one motifel can be found in the other).

It takes getting use to the concept of an overall, single value dissimilarity between patterns, and, in addition, to the way dissimilarity is assessed by the JSD. Fig. 1F is a table giving the values of JSD between four patterns shown in Figs. 1A to D. Visually, the patterns A and B are similar to each other and dissimilar to patterns C and D, which, in turn, are dissimilar to each other. This is reflected by the values in the table. However, note that according to JSD pattern C and D are overall more similar than patterns C and A (or C and B, or D and A or D and B). The overall measure of dissimilarity is a result of a compromise (encapsulated by the choice of the pattern representation and the dissimilarity measure) between many factors contributing to the final assessment. For example, the JSD assessment regarding dissimilarity between \(C\) and \(D\) versus dissimilarity between \(C\) and \(A\) may be questioned by those who emphasize texture over composition.

Fig. 2 shows how different motifels can have the same degree of dissimilarity to a given motifel despite having different patterns. The focus motifel is located in the center (bull’s eye) of the figure. To match a pattern to the pattern of the focus motifel all features of both patterns must match. Lack of matching in one or more features results in a JSD>0. Fig. 2 shows four motifels for which the level of dissimilarity from the focus pattern is JSD=0.1, eight motifels for which JSD=0.2, and eight motifels for which JSD=0.3. The smaller the values of JSD the more similar the patterns are to the focus pattern. As motifels fail to match an increasing number of features to the focus pattern they diverge from it visually in many different ways. This is illustrated in Fig. 2 where motifels at each level of dissimilarity are grouped by visually similar departures from the focus motifel. The focus motifel is a landscape dominated by pasture (yellow) with some forest (green) and a dash of an urban area (red). Four different ways in which a motifel can become dissimilar from the central motifel are shown: (1) tending toward more urban areas, (2) tending toward more cropland (brown), (3) tending toward more water (blue), and (4) tending toward more forest. Thus, although all motifels in Fig. 2 are dissimilar to the focus motifel at a level of at most 0.3, their mutual dissimilarities may be greater.
Figure 2. Illustration of the concept of dissimilarity between motifels. The focus motifel is located in the center (bull eye), other 20 motifels are arranged radially around the center according to their dissimilarities from the focus motifel and azimuthally according to similarities between themselves. See text for more details.

2.2.4. Segmentation and clustering

With motifel representation and distance function now defined, the most straightforward way to regionalize the NLCD is to perform a clustering of all motifels in the grid and treat the spatial footprints of the resulting clusters as the final LPTs. This approach was used in previous works on automated regionalization (Cardille and Lambois 2010, Partington and Cardille 2013). However, clustering of individual motifels may not preserve the spatial cohesion of LPTs (Niesterowicz and Stepinski 2013), which is a desirable characteristic of LPT regions. Additionally, when working with very large datasets and electing to work with patterns at a small spatial scale (small motifels) the number of motifels to be clustered could be large making some otherwise appealing clustering techniques (for example, the hierarchical clustering) unfeasible. Therefore, we first perform
segmentation of the grid of motifels, and then cluster the segments instead of the motifels themselves. Segmentation is the process of partitioning the grid into local areas (segments) of relatively uniform pattern. In principle, segmentation could be the only operation necessary for regionalization, however, as the segmentation is a local aggregation, it is possible, and indeed likely, that several disjointed segments could correspond to the same LPT. Therefore, clustering of segments is necessary to obtain an unambiguous set of LPTs.

As we anticipate that future regionalizations may be performed on datasets different than the NLCD, and they may use different choices for motifel representation and dissimilarity function, we work with segmentation and clustering algorithms that require only the computation of dissimilarity between motifels. Thus, we don’t use algorithms such as the $K$-means algorithm, which requires the averaging of histograms. As it happens, the co-occurrence histogram and the JSD are compatible with an averaging operation, but some other motifel representations (for example, representations used in Stepinski et al. (2014)) and dissimilarity functions may not support averaging.

For segmentation we use a variant of a non-hierarchical greedy region-growing segmentation algorithm (Levine and Shaheen 1981) that has been customized to work only with dissimilarities. The algorithm has only one free parameter $\text{JSD}_m$ – a merging criterion. The $\text{JSD}_m$ value is a threshold indicating the maximum value of JSD between a motifel and its neighboring segment that allows for merging of this motifel with the segment. Dissimilarity between a single motifel and a segment is defined as the value of JSD between this motifel and the medoid of the segment. A medoid is a motifel which belongs to the segment, and whose average value of JSD to all the other motifels in the segment is the lowest; it is an equivalent of centroid when averaging is not supported. For details of the segmentation algorithm see Stepinski et al. (2015).

For the clustering of segments (represented by their medoids) we use a hierarchical clustering algorithm with Ward’s linkage (Ward 1963). This algorithm uses only dissimilarities between segments as input and is based on the principle of minimization of the total within-segment variance - exactly what we want the segmentation to achieve. In addition, by using a hierarchical algorithm we can trace how an algorithm agglomerates more homogeneous LPTs into less homogeneous, more broadly defined LPTs. In other words, we get some insight into the relationship between the algorithmic and perceptual similarities of landscape patterns.

2.2.5. Determining the number of LPTs

The results of hierarchical clustering can be represented in the form of a dendrogram (see Fig. 3). By cutting the dendrogram at a certain height we obtain a clustering solution. The number of clusters (alternatively, the height at which a dendrogram is cut) is not determined by the hierarchical clustering algorithm itself, instead it can be selected by an analyst or estimated by an ancillary algorithm (Dunn 1973, Davies and Bouldin 1979, Rousseeuw 1987, Salvador and Chan 2004). Here we use the L-method (Salvador and Chan 2004) algorithm to estimate a suggested number of clusters. We stress that the result of such an estimation needs to be taken as a suggestion rather than a final determination.

The L-method first constructs an evaluation graph which is a two-dimensional plot with a number of clusters $x$ plotted on the horizontal axis and a measure of quality of clustering consisting of $x$ clusters plotted on the vertical axis. For hierarchical clustering, the quality of clustering is measured by the minimum merging dissimilarity resulting in $x$ clusters. An evaluation graph can be calculated directly from a dendrogram and has
Figure 3. Heat map of dissimilarity matrix between 6,317 segments. Red-to-white color gradient indicates dissimilarities between segments from low to high. Dendrogram used to construct a heat map is also shown up to the fifteen levels. The regions of heat map corresponding to nine and fifteen clusters are labeled by numbers and by unique colors. See text for more details.
a characteristic L shape (see Fig. 4). One portion of the L-shaped graph corresponds to a regime of many clusters where the minimum merging dissimilarity is small, whereas another portion of the L-shaped graph corresponds to a regime of a few clusters where the minimum merging dissimilarity is large. The L-method calculates the location of a “knee” in the L-shaped graph where the two regimes meet, and outputs it as the recommended number of clusters.

2.3. Computer software

We use the GeoPAT toolbox (Jasiewicz et al. 2015) to divide the NLCD into a grid of motifels, to calculate their co-occurrence histograms, to calculate the JSD, and to perform segmentation. GeoPAT is a set of extension modules to GRASS GIS v.7 (Neteler et al. 2012) and is written in ANSI C utilizing GRASS API. GeoPAT can be downloaded from http://sil.uc.edu. Ward’s hierarchical clustering and the L-method were performed using the R software programming language and environment (R Core Team 2015).

3. Results

For regionalization of the NLCD we have chosen the size of the motifel to be 500×500 cells setting the scale of land cover pattern at 15 km. This choice is arbitrary but similar to
the 9.65 km scale of the US landscape selected by Hammond (1964). Dividing the entire NLCD2011 into 15 km motifs results in a grid having the size of $323 \times 209 = 67,507$.

Application of the segmentation algorithm with the $JSD_m$ value set to 0.3 agglomerates 67,507 motifs into 6,317 irregular segments. This particular value of $JSD_m$ has been selected empirically, smaller values result in a large number of small segments whereas larger values lead to segments containing visibly inhomogeneous patterns. Although the subsequent clustering inevitably increases inhomogeneity of LPTs, it is desirable to start from initially homogeneous segments. In preparation for clustering these segments into LPTs, we calculate a $6,317 \times 6,317$ distance (dissimilarity) matrix which contains dissimilarities between each pair of segments. The hierarchical clustering is calculated using only values of the dissimilarity matrix.

Fig. 3 shows the heat map (Wilkinson and Friendly 2009) of dissimilarities between the segments – a graphical representation of the dissimilarity matrix with rows and columns rearranged in accordance with a dendrogram (so the segments similar to each other are close to each other in the matrix) and the color reflecting a degree of dissimilarity from 0 (red) to 1 (white). Clusters appear on the heat map as square red blocks located on the diagonal. The existence of red block structures on the diagonal of the heat map indicates, even without further mapping, that hierarchical clustering of segments will result in the meaningful regionalization of the US.

Fig. 3 also shows a dendrogram resulting from the hierarchical clustering. Although the dendrogram extends to 6,317 nodes, only fifteen nodes, corresponding to the top fifteen clusters are shown. These fifteen clusters are labeled by unique numbers and colors. Subdivisions of the fifteen clusters into smaller sub-clusters and their agglomeration into larger clusters can be seen in the heat map. The nine larger clusters are also indicated by unique numbers and colors. One immediate observation is that at the very top level of the hierarchy, all segments are divided into just two clearly different LPTs, one consisting of clusters 1 to 9 (for partition into 15 clusters), and another consisting of clusters 10 to 15. Interestingly, this broadest landscape pattern-based division corresponds to spatial division of the US into its western and eastern regions (see below).

Using the dendrogram we can regionalize the US into any number of LPTs from 1 to 6,317. The L-method (Fig. 4) suggests regionalization into eight clusters. After examining the dendrogram, the heat map, and the maps of several regionalizations with different numbers of LPTs, we decided to select division of US into nine LPTs as our primary regionalization. Using nine instead of eight LPTs (as suggested by the L-method) distinguishes between LPTs #3 and #4 which we thought are distinct enough (see Fig. 5) to be worthy mapping as separate LPTs. We also retained regionalizations into four, six and fifteen LPTs (see Fig. 5).

Fig. 5 shows the geographical division of the US into the four, six, nine, and fifteen LPTs. The colors and labels indicating LPTs in regionalizations stemming from nine-partitioning and fifteen-partitioning are the same as those used to indicate clusters in the dendrogram shown in Fig. 3. For each regionalization the set of exemplars characterizing each LPT is given. An exemplar gives an example of a pattern of LULC that is characteristic to a given LPT; we use medoids of clusters used for a given regionalization as exemplars. As the division of the US into LPTs is hierarchical, some of the LPTs in the nine-partitioning are agglomerates of LPTs in the fifteen-partitioning (see Fig. 3 to identify agglomerates). Similarly, some of the LPTs in the six-partitioning are agglomerates of LPTs in the nine-partitioning, and the LPTs in the four-partitioning are agglomerates of LPTs in the six-partitioning.

By cutting a dendrogram at progressively smaller values of minimum merging dissim-
Figure 5. Regionalization maps with different number of LPTs as indicated. For each map the set of exemplars is shown to illustrate characteristic land cover patterns for each LPT. The exemplars are color-coded to correspond to the map.

Similarity we obtain maps indicating progressively more homogeneous landscape patterns. With only two LPTs (not shown) the US divides roughly into its western and eastern constituents (there are some patches of “eastern” LPT in the western part; they correspond to areas covered by cropland and mixed forest). If four LPTs are selected the eastern part breaks into water (Great Lakes) and non-water, and the western part breaks
A regionalization into nine LPTs is shown in details in Fig. 6. The purpose of this figure is to illustrate the degree of homogeneity of land cover patterns within each LPT (panel A). Panel B is the NLCD2011 legend. Panel C shows the regionalization of the US into nine LPTs. This is the same map as shown in Fig. 5 (nine-partitioning) which is included here to provide reference to the meaning of boundaries in panel A. Panel D
is the legend to landscape pattern types. Giving concise names to LPTs is a difficult task because they are complex patterns of land cover classes. Here, we use a simplified nomenclature proposed by Wickham and Norton (1994). The term **matrix** describes a pattern dominated by a single LULC class, and the term **mosaic** describes a pattern to which several LULC classes contribute significantly. We name LPTs using the names of dominant land cover classes and adding either matrix or mosaic to it.

A regionalization into fifteen LPTs (see Fig. 5) leaves four LPTs (#1, #3, #5, and #8) unchanged from the nine-partitioning. Another four LPTs from the nine-partitioning (#2, #4, #6, and #7) are split (see Fig. 3) into pairs of more homogeneous LPTs. The final LPT from the nine-partitioning (#9) is split into three more homogeneous LPTs.

4. Discussion and Conclusions

The novelty of the presented approach is that it regionalizes patterns of a variable (LULC) rather than the variable itself, and that it does this in an automatic, unsupervised fashion. To the best of our knowledge this is the first time patterns in the entire NLCD have been automatically regionalized. Earlier work by Cardille and Lambois (2010) identified exemplars of 17 characteristic LULC patterns in the NLCD 1992 but did not present a map of their regionalization. Without a regionalization map their result cannot be compared to ours because the NLCD 1992 and 2011 use different, not compatible sets of LULC classes which results in incomparable sets of exemplars. The only other study describing a technique similar to the one described here was by Partington and Cardille (2013) who regionalized LULC in the Canadian province of Quebec. Their method differs from ours with respect to the manner in which motifels are represented, in the choice of similarity function, and in the clustering algorithm used. Their regionalization was performed by a non-hierarchical clustering without the prior segmentation step. This was feasible because they clustered, in a non-hierarchical fashion, only 1,398 motifels, the result of using a relatively small size region (Quebec) and large sized (30 km) motifel. Such a method would not be feasible for hierarchical regionalization of the NLCD, especially if one is interested in the regionalization of smaller-scale patterns.

It needs to be pointed out that the result of our regionalization cannot be quantitatively assessed beyond calculating the quality of clustering which has only computational relevance (if one would like to compare two regionalizations calculated by two different clustering algorithms) and does not comment on the usefulness of the resultant map. This is a general property of unsupervised machine learning where the goal is to explore a dataset and to reveal a priori unknown groupings in the data. In our case, the LPTs emerge from the data so to ask whether they are “correct” or “incorrect” is not a well-posed question. However, we can ask whether our regionalization is useful. This can be assessed qualitatively by visually examining Fig. 6 for the stationarity of LPTs patterns and for a subjective assessment of usefulness of the map to a problem at hand. We suspect that the regionalization at the level of nine LPTs, although highly illustrative, may not offer insight beyond what is already known. However, higher level maps such as the map with 15 LPTs or one with even more detailed delineation, may indeed offer new insight.

Visual inspection of Fig. 6A reveals that different LPTs appear to have somewhat different levels of stationarity, which is a feature of hierarchical clustering which may produce “non-spherical” (in dissimilarity space) clusters. Non-spherical clusters tend to have a larger range of dissimilarities between their elements than spherical clusters of
comparable size. Using a non-hierarchical algorithm, such as a \(K\)-medoids algorithm, could result in LPTs with a more similar degree of non-stationarity as such algorithms tend to produce more spherical clusters. However, in this paper we preferred to use a hierarchical algorithm for its ability to explicitly reveal different levels of regionalization (see Fig. 5). In addition, some LULC patterns may be characterized by larger variations than others, so it is not clear whether the same level of variation in all LPTs is a desirable feature.

Methodology presented in this paper focuses on regionalization with respect to a pattern of a single aspect of environment – land cover. However, there is much interest in regionalization on the basis of patterns of multiple environmental factors (topography, land cover, soil properties, climate etc.) leading to the broad scale map of physiographic provinces (Fenneman 1915), or, more recently, to the map of ecoregions (Omernik and Griffith 2014). The most recent edition of the ecoregions map is still delineated manually partially because its creators believed (Omernik and Griffith 2014) that regionalization must be based on the patterns of environmental variables rather than the values of
the variables themselves and because no pattern-based regionalization algorithms were available to them. Attempts at algorithmic regionalization of ecoregions (Hargrove and Hoffman 1999, 2004) relied on clustering the values of environmental variables and did not utilize a pattern-based approach.

Our method, when extended to handle multiple environmental factors, can offer a way for pattern-based, algorithmic regionalization of ecoregions. It has been shown to work well with the land cover (this paper) and with topography (Jasiewicz and Stepinski 2013b, Jasiewicz et al. 2014). It can be applied in an analogous manner to other single environmental factors. Future research will address the issue of how to segment and cluster motifels on the basis of a set of dissimilarity values corresponding to various factors. To show the potential of our method for future delineation of ecoregions Fig. 7 shows a comparison between our map of LPTs (nine-partitioning) and the Level III map of ecoregions (Omernik and Griffith 2014) within a state of New Mexico. There are five LPTs and eight ecoregions present. The two top panels show: (A) the boundaries of the LPTs superimposed on the NLCD2011 and (B) the actual LPTs. The two bottom panels show the same for ecoregions.

To assess quantitatively a degree of spatial association between LPTs and ecoregions we use an information theoretic index called the V-measure (Rosenberg and Hirschberg 2007). The V-measure uses two criteria: homogeneity ($V_H$) and completeness ($V_C$). An association satisfies fully the homogeneity criteria if each LPT contains only locations belonging to a single ecoregion. An association satisfies fully the completeness criteria if each ecoregion contains only locations belonging to a single LPT. Both $V_H$ and $V_C$ have ranges between 0 and 1. For maps shown in Fig. 7 $V_H = 0.33$ and $V_C = 0.53$. The interpretation of these values is as follows. On average, the homogeneity of ecoregions in a single LPT is increased over their homogeneity in the entire site (New Mexico) by 33%. On average, the homogeneity of LPTs in a single ecoregion is increased over their homogeneity in the entire site by 53%. The relatively low value of $V_H$ is because each LPT is found in two or more ecoregions. For example, LPT #8 (evergreen forest mosaic) is found in Southern Rockies and Arizona/New Mexico Mountains ecoregions; they may differ in some other environmental factors but they have the same pattern of land cover. The relatively high value of $V_C$ is because each ecoregion comprises predominantly of a single LPT. For example, Chihuahuan Desert ecoregion comprises of LPT#9 (shrub matrix). This is exactly as we would expect, several ecoregions may be characterized by the same pattern of land cover but differ in other factors. The final V-measure is given by computing the harmonic mean of $V_H$ and $V_C$ and is equal to 0.41. Thus, at least within an area shown in Fig. 7, there is a relatively high spatial association between LPTs and ecoregions. Inclusion of additional environmental factors in our regionalization algorithm will further increase the association between delineated units and ecoregions.

Finally, note that although in this paper we concentrated on unsupervised analysis, the COBIA framework also supports supervised analysis (Jasiewicz et al. 2014, 2015). In this way our method can also find application in conservation planning. For instance, it could be used for identifying unprotected areas which are suitable for protection (Andrew et al. 2012) by identifying unprotected areas with patterns of land cover very similar to those already protected. In the same way it can be used to identify areas suitable for long-term monitoring to establish potential benefits of protection (Cardille et al. 2012). In ecology, it could be helpful in searching for control sites for areas that will experience treatment (for example, road construction or forest harvesting) (Dilts et al. 2010).
Acknowledgments

This work was supported by the University of Cincinnati Space Exploration Institute, the National Aeronautics and Space Administration under grant NNX15AJ47G, and by the National Science Center (NCN) under grant DEC-2012/07/B/ST6/01206.

References


REFERENCES


REFERENCES


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

1419–1426.