RESEARCH ARTICLE

Pattern-based identification and mapping of landscape types using multi-thematic data

Jakub Nowosad[a], Tomasz F. Stepinski[b]

[a]Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznan, Poland; [b]Space Informatics Lab, Department of Geography and GIS, University of Cincinnati, Cincinnati, OH, USA

ARTICLE HISTORY
Compiled March 7, 2021

ABSTRACT
Categorical maps of landscape types (LTs) are useful abstractions that simplify spatial and thematic complexity of natural landscapes, thus facilitating land resources management. A local landscape arises from a fusion of patterns of natural themes (such as land cover, landforms, etc.), which makes an unsupervised identification and mapping of LTs difficult. This paper introduces the integrated co-occurrence matrix (INCOMA) – a signature for numerical representation of multi-thematic categorical patterns. INCOMA enables an unsupervised identification and mapping of LTs. The region is tessellated into a large number of local landscapes – patterns of themes over small square-shaped neighborhoods. With local landscapes represented by INCOMA signatures and with dissimilarities between local landscapes calculated using the Jensen-Shannon Divergence (JSD), LTs can be identified and mapped using standard clustering or segmentation techniques. Resultant LTs are typically heterogeneous with respect to categories of contributing themes reflecting the human perception of a landscape. LTs calculated by INCOMA are more faithful abstractions of actual landscapes than LTs obtained by the current method of choice – the map overlay. The concept of INCOMA is described, and its application is demonstrated by an unsupervised mapping of LT zones in Europe based on combined patterns of land cover and landforms.

KEYWORDS
spatial patterns, pattern similarity, regionalization, thematic maps, global categorical datasets

1. Introduction

The term "landscape" means different things in different scientific domains, in this paper, we follow the definition of landscape as a recognizable, although often heterogeneous part of the terrestrial surface showing characteristic pattern of natural themes (Mücher *et al.* 2010). There is a significant interest in the identification and classification of landscape types (LTs) as a foundation for planning and management to develop optimal strategies for sustainable use of land resources (Wascher 2005; Mücher *et al.* 2010). LTs also provides first-order information about the geographical distribution of biodiversity and ecological processes (Heikkinen *et al.* 2004).

CONTACT Jakub Nowosad. Email: nowosad.jakub@gmail.com, Tomasz F. Stepinski. Email: stepintz@uc.edu

The spatial and thematic complexity of landscapes can be better grasped in terms of categorical spatial zones. However, because of their complexity, identifying, and mapping LTs using available data is challenging. First, multiple layers (themes), including topography, land cover, climate, and soil/geology, contribute to the character of a landscape. Global categorical datasets for those themes exist (Karagulle *et al.* 2017; ECMWF 2019; Metzger *et al.* 2013; Hengl *et al.* 2017), but how to best combine them so they reflect a LT remains an open question. Second, a landscape must have a spatial scale larger than the size of cells in global datasets of natural themes ($\sim$ 300 m) and thus, it is characterized by patterns that contributing variables form over its extent rather than by values of these variables (Omernik and Griffith 2014). Using multi-thematic data and pattern-based elementary spatial units is needed for the identification and mapping of landscapes.

Simensen, Halvorsen, and Erikstad (2018) offered a review of 54 different papers devoted to landscape identification, classification, and mapping. The majority of them covered a single country or a single region within a country, while two (Mücher *et al.* 2010; Sayre *et al.* 2014) had a continental or global coverage. This review did not focus on specific technical methodologies employed, but the two papers where landscapes were mapped on a large spatial scale both used multi-thematic datasets but were cell-based instead of being pattern-based. Such an approach has some important shortcomings. First, it identifies LTs at the scale of the size of the cell. Therefore, spatial units of maps stemming from these methods are homogeneous in categories of contributing variables. By not being pattern-based, these methods can only map homogeneous LTs, whereas many LTs, especially those defined on the scale of a few kilometers or more, are heterogeneous.

Wickham and Norton (1994) were the first to consider a pattern-based approach to the identification and mapping of LTs. In such approach, a study area is tessellated into relatively small square blocks of cells carrying values of contributing variables. Each block is characterized by a pattern formed by values of contributing variables over its extent, therefore we refer to it as a local landscape (LL) – a landscape at the scale that is large enough to be called a landscape but small in comparison to the size of the study area. Identification and delineation of LTs are tantamount to grouping similar LLs into distinct clusters (zones). Recent works based on the LL concept can be divided into three groups, those devoted to mapping LTs (Long, Nelson, and Wulder 2010; Cardille and Lambois 2010; Cardille *et al.* 2012; Partington and Cardille 2013; Niesterowicz and Stepinski 2013; Niesterowicz, Stepinski, and Jasiewicz 2016; Niesterowicz and Stepinski 2017; Nowosad and Stepinski 2018), those enabling search using a query-by-example (Stepinski, Netzel, and Jasiewicz 2013; Liu *et al.* 2017), and those which assess temporal change (Netzel and Stepinski 2014; Nowosad, Stepinski, and Netzel 2019). It is important to note that starting from the paper by Wickham and Norton (1994) all cited pattern-based methods used only a single theme – land cover. Therefore, they cannot be used to identify landscape types in the sense we consider them in this paper.

Pattern-based methods have two fundamental ingredients: (a) pattern signature and (b) a dissimilarity function. A pattern signature is a numerical embedding of a categorical pattern. A signature is a vector of numbers quantifying a pattern in a form that can be utilized for the calculation of similarity between different patterns. A dissimilarity function (also called a "distance") is a function that takes signatures of two patterns as an input and outputs a single number that quantifies a degree of dissimilarity between two patterns. A categorical pattern has two major attributes, its composition (how its constituting cells are distributed between different categories) and its configuration (how the cells are spatially arranged to form a motif). Similar

patterns have similar compositions and similar configurations. A dissimilarity function is used for grouping LLs into distinct zones by either clustering or segmentation techniques. In each zone a dissimilarity between any two constituent LLs is small, i.e., same motif approximately repeats itself throughout the zone.

Early works (Long, Nelson, and Wulder 2010; Cardille and Lambois 2010; Cardille *et al.* 2012; Partington and Cardille 2013) used a set of landscape metrics (O'Neill *et al.* 1988; Turner and Gardner 1991) as a pattern signature and the Euclidean distance as a dissimilarity function. However, Niesterowicz and Stepinski (2016) demonstrated that a pattern signature based on landscape metrics suffers from problems of normalization and assigning weights. Moreover, it is not clear how to extend this approach to multi-thematic patterns. Alternatively, previous research used a signature based on a co-occurrence matrix (COMA) that provides a statistical description of the whole pattern without quantifying individual elements (Haralick, Shanmugam, and Dinstein 1973; Barnsley and Barr 1996; Chang and Krumm 1999; Jasiewicz, Netzel, and Stepinski 2015). The COMA tabularizes first order adjacencies – pattern-wide counts of the pairs of the adjacent cells of given categories. As such, they form a pattern's signature that is free from issues of normalization and assigning weights. COMA signatures, which are normalized histograms (discrete probability distributions), are often paired with the Jensen-Shannon Divergence (JSD) (Lin 1991) as a dissimilarity function.

The purpose of this paper is to extend the COMA/JSD method to multi-thematic patterns so we can identify and map LTs as defined at the beginning of this section. For this purpose, only the signature needs a modification as the JSD will work with any signature which is in the form of a normalized histogram, regardless of what this histogram represents. We introduce the integrated co-occurrence matrix (INCOMA) – an extension of a COMA signature for numerical representation of multi-thematic categorical patterns. INCOMA is based on the ideas from the field of pattern recognition in images (Vadivel, Sural, and Majumdar 2007). INCOMA tabularizes intra-thematic as well as inter-thematic adjacencies. By doing so it contains not only information about patterns of all variables but also information about relative positions of different patterns.

To demonstrate the ability of INCOMA to identify and map LTs we apply it to the continent of Europe using two variables – land cover (LC) and landforms (LF). The method identifies twelve LTs. We assess the quality of this delineation, i.e., calculate an average dissimilarity between motifs in a zone (the smaller the value the better), and an average dissimilarity between motifs in different zones (the larger the better). We also compare the resultant map with (a) the map obtained by an overlay of LC and LF maps, which results in 130 LTs, and (b) the method that first performs pattern-based delineations of LC and LF separately (using COMA) and then overlays the two pattern-based maps of LC patterns and LF patterns. This method yields 36 LTs. The comparison reveals that of the three methods, only INCOMA can produce, in an unsupervised fashion and without post-processing, a delineation that includes zones characterized by simple motifs (e.g., collections of LLs with mostly homogeneous land cover and landforms) as well as zones characterized by complex motifs (e.g., collections of LLs with mostly heterogeneous land cover and landforms).

2. Materials

We used two global categorical raster datasets as inputs to an unsupervised algorithm that identifies and maps LTs in Europe, the C3S land cover map for the year 2018

(ECMWF 2019), and the Hammond's landforms map (Karagulle *et al.* 2017). The C3S global land cover map was produced by the European Space Agency. It consists of 64,800 by 129,600 pixels, thus their spatial resolution is 10 arc-seconds (about 300 meters at the equator). Each pixel is assigned to one of 37 land cover categories, with some existing only for regions with more accurate information available. Hammond's landforms map has sixteen categories of landforms based on global 250-meter resolution GMTED2010 elevation data. It consists of 67,020 by 172,800 pixels. We reprojected both datasets into the interrupted Goode homolosine projection and resampled them to the resolution of 300 meters. Furthermore, in C3S, we grouped 37 original categories into nine broader IPCC (Intergovernmental Panel on Climate Change) categories that represent major classes of land cover: agriculture, forest, grass, wetland, settlement, shrubland, sparse, bare, and water.

3. Methods

3.1. *Co-occurrence matrix*

A COMA is constructed by counting all pairs of the adjacent cells in the categorical raster representing a single theme over the study area. The adjacency is defined by the Von Neumann neighborhood (four cardinal directions). Thus, COMA is a $k \times k$ square matrix, where k is a number of unique categories present in the theme. The normalized COMA (the sum of all matrix components adds to 1) is a joint probability distribution of two variables, "a value of categorical label of the focus cell" and "a value of categorical label of the adjacent cell." COMA serves as a pattern's "signature" – a statistical summary of the pattern, which is invariant to rotation and reflection of the pattern and can be used to assess a degree of dissimilarity between two patterns.

For convenience, we flatten the COMA row-wise to transform it (without any loss of information) into a histogram (right panels in Figure 1). Because we consider unordered adjacent pairs, the COMA transforms into a histogram with $(k^2 + k)/2$ bins (symmetric off-diagonal terms are added to a single bin). Both normalized matrix and histogram derived from it represent the bi-variate probability distribution of adjacent pairs occurrence in the pattern, the histogram is just a format more convenient to pass to a dissimilarity function. Note that the dimensions of the matrix or the length of the histogram depend only on the number of categories. As the distinction between normalized COMA, non-normalized COMA, and the normalized histogram is clear from the context, we use COMA abbreviation for all.

The concept of COMA is illustrated in Figure 1. This figure shows an LL with a land cover (LC) pattern shown in the upper row and a landforms (LF) pattern shown in the lower row; both thematic layers have just two categories. COMAs and corresponding histograms are also shown. Because there are only two categories, COMAs are 4×4 matrices, and histograms have only three bins. The first and the last bins correspond to the intra-thematic co-occurrence of categories, if those bins are high, the patterns are integrated - cells of different categories form large spatial patches. The middle bin corresponds to the inter-thematic co-occurrence of categories. In an integrated pattern, where different categories rarely co-occur, these bins are low. Note that the middle bin for LF is higher than the middle bin for LC, reflecting more co-occurrences between two LF categories than between two categories of LC. Although each histogram characterizes its corresponding pattern, together they are insufficient to unambiguously describe LTs (bi-modal patterns consisting of LC and LF). This is

because they do not carry information about relative layouts of the two patterns.

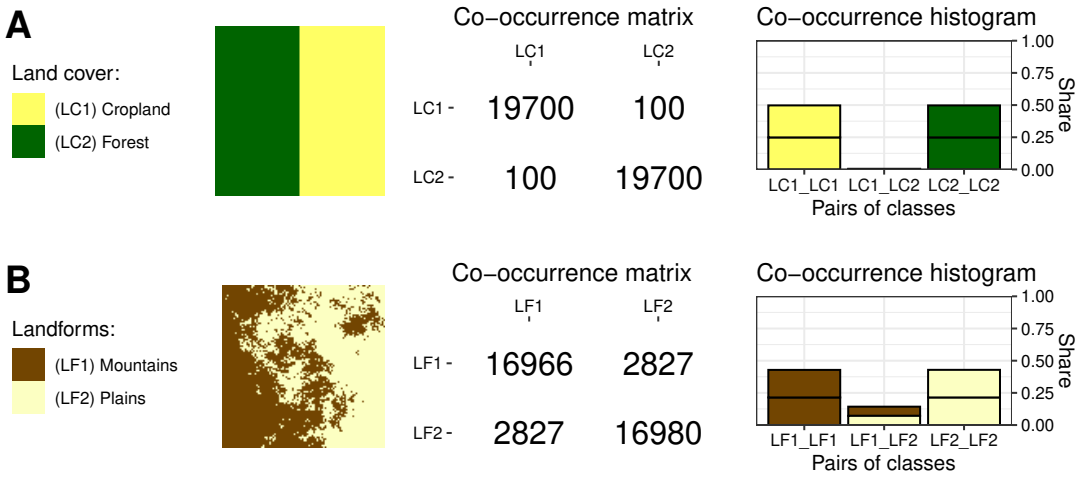

Figure 1. Two illustrative patterns of categorical rasters and their spatial signatures, a non-normalized COMA and a COMA histogram. (A) A land cover data with two classes (LC1) Cropland and (LC2) Forest. (B) A landform data with two classes (LF1) Mountains and (LF2) Plains

3.2. *Dissimilarity function*

A dissimilarity between two patterns is calculated using a dissimilarity function. There is a large selection of possible functions that measure dissimilarity between two histograms (Cha 2007). We use the Jensen-Shannon divergence (Lin 1991) to calculate dissimilarity between two histograms because of its robustness and because it has been shown to have a good performance in the side-by-side comparison with other measures (Rubner *et al.* 2001). The Jenson-Shannon divergence (JSD) is given by

$$JSD(A,B) = H(\frac{A+B}{2}) - \frac{1}{2}[H(A) + H(B)], \tag{1}$$

where A and B are LLs signatures (normalized histograms) and $(A+B)/2$ is a normalized histogram constructed by concatenation of A and B. Symbols $H(A)$, $H(B)$, and $H((A+B)/2)$ are the values of the Shannon entropy (Shannon 1948) for different histograms. The JSD takes values from 0 to 1, with the value of 0 indicating that two COMAs are identical (the two patterns are very similar), and the value of 1 indicating that two COMAs do not share bins (patterns do not share classes and are thus very dissimilar). Intermediate values of JSD, between 0 and 1, represent a degree of dissimilarity between the two patterns. For a discussion of how to interpret the values of JSD see Stepinski, Netzel, and Jasiewicz (2013) and Niesterowicz, Stepinski, and Jasiewicz (2016).

3.3. *Integrated Co-occurrence Matrix*

Vadivel, Sural, and Majumdar (2007) developed a signature of color image called Integrated Color and Intensity Co-occurrence Matrix (ICICM) to improve context-based image search and retrieval. An image is characterized by the color and intensity

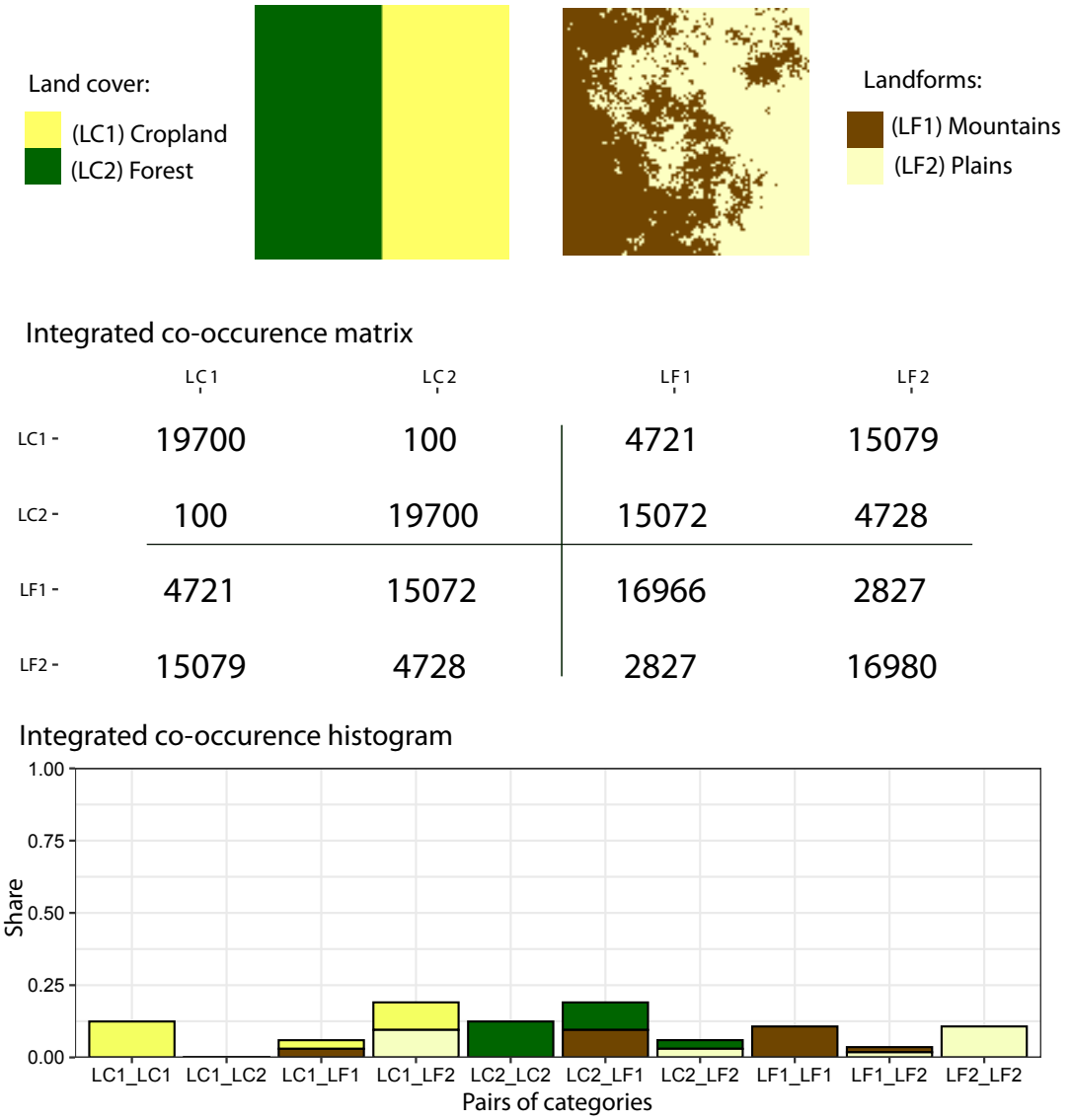

Figure 2. Bi-thematic categorical pattern and two forms of its signature, integrated co-occurrence matrix and normalized histogram. Lines are drawn in the middle panel to indicate a blocky structure of INCOMA.

of its constituent pixels. These authors showed that including information about spatial relations between the raster of colors and the raster of intensities improves the image signature and leads to better image retrieval results. The situation in geographic multi-thematic patterns is analogous, except it may not be restricted to only two patterns.

Based on this insight, we propose a modification of ICICM for application to geographical multi-thematic patterns. We call this modification the INtegrated CO-occurrence MAtrix (INCOMA). In the bi-thematic case, the INCOMA is analogous to ICICM, so a reader interested in the mathematical description is referred to Vadivel, Sural, and Majumdar (2007). However, the key concept of INCOMA can be explained without mathematics by using a simple illustration (Figure 2).

This illustration shows a bi-thematic site having simple patterns of LC and LF (each theme has only two categories). Two adjacent cells in the bi-thematic pattern are characterized by four numbers, x - a category of theme 1 in the focus cell, y - a

category of theme 1 in the adjacent cell, w - a category of theme 2 in the focus cell, and v - a category of theme 2 in the adjacent cell. For the bi-thematic pattern shown in Figure 2, each of these variables has only two possible values. In general, each variable has as many values as there are categories in a corresponding theme.

INCOMA is constructed from blocks, with each block by itself being a COMA. Diagonal blocks are COMAs of theme 1 and theme 2, respectively. Off-diagonal blocks are COMAs calculated from adjacent pairs of cells, where a focus cell uses its theme 1 category label, and an adjacent cell uses its theme 2 category label. For example, the lower off-diagonal block in the middle panel of Figure 2 informs that 4721 LC1 cells are adjacent to LF1 cells, 15072 LC2 cells are adjacent to LF1 cells, 15079 LC1 cells are adjacent to LF2 cells, and 4728 LC2 cells of adjacent to LF2. These off-diagonal blocks provide information about spatial relations between categories from different themes, while diagonal blocks provide information about the spatial relationships between categories of the same theme.

Note that INCOMA is *not* a joint probability distribution of variables x, y, w, and z. It contains less information than the joint probability distribution. However, its advantage is a smaller size. The small size of INCOMA in comparison with the size of joint probability may not be obvious from the simple example in Figure 2, but assuming a more realistic number of classes, $C_1 = 5$ classes for LC and $C_2 = 10$ classes for LF, INCOMA consists of only $(C_1 + C_2)^2 = 225$ numbers, whereas the joint probability consists of $C_1 \times C_2 \times C_1 \times C_2 = 2500$ numbers.

Extending INCOMA to more than two modalities is straightforward, with n themes, the INCOMA has n^2 blocks. If themes have a different number of classes, the diagonal blocks would have different sizes but would maintain square shapes, off-diagonal blacks would have different sizes and rectangular shapes. However, the entire INCOMA would have a square shape with the linear size equal to the sum M of categorical classes in all themes $M = m_1 + m_2 + \cdots + m_n$. We normalize INCOMA so the sum of all components in all of its blocks is equal to 1 and flatten it to a 1D histogram with $(M^2 + M)/2$ bins. This makes it a well-defined input to the JSD.

3.4. *Software*

We wrote an open-source R (R Core Team 2020) package `comat` (Nowosad 2020) allowing to create integrated co-occurrence matrices and integrated co-occurrence histograms based on input data in the form of a categorical matrix. Most functions in this package use computationally fast and memory-efficient C++ code. `comat` depends only on two other R packages, Rcpp and RcppArmadillo (Eddelbuettel and François 2011; Eddelbuettel and Sanderson 2014), and therefore can be used as a building block for other R packages. The package installation instructions and documentation can be found at `https://nowosad.github.io/comat/`.

Additionally, we used the *philentropy* package (Drost 2018) to calculate the distances between the co-occurrence histograms, the *stars* package (Pebesma 2020) to represent raster data, and the `motif` package for (Nowosad 2021) to perform pattern-based analysis.

4. Identifying and mapping landscape types in Europe

To demonstrate how INCOMA works on real-world data, we use it to perform an unsupervised identification and mapping of LTs in Europe based on two themes, LC

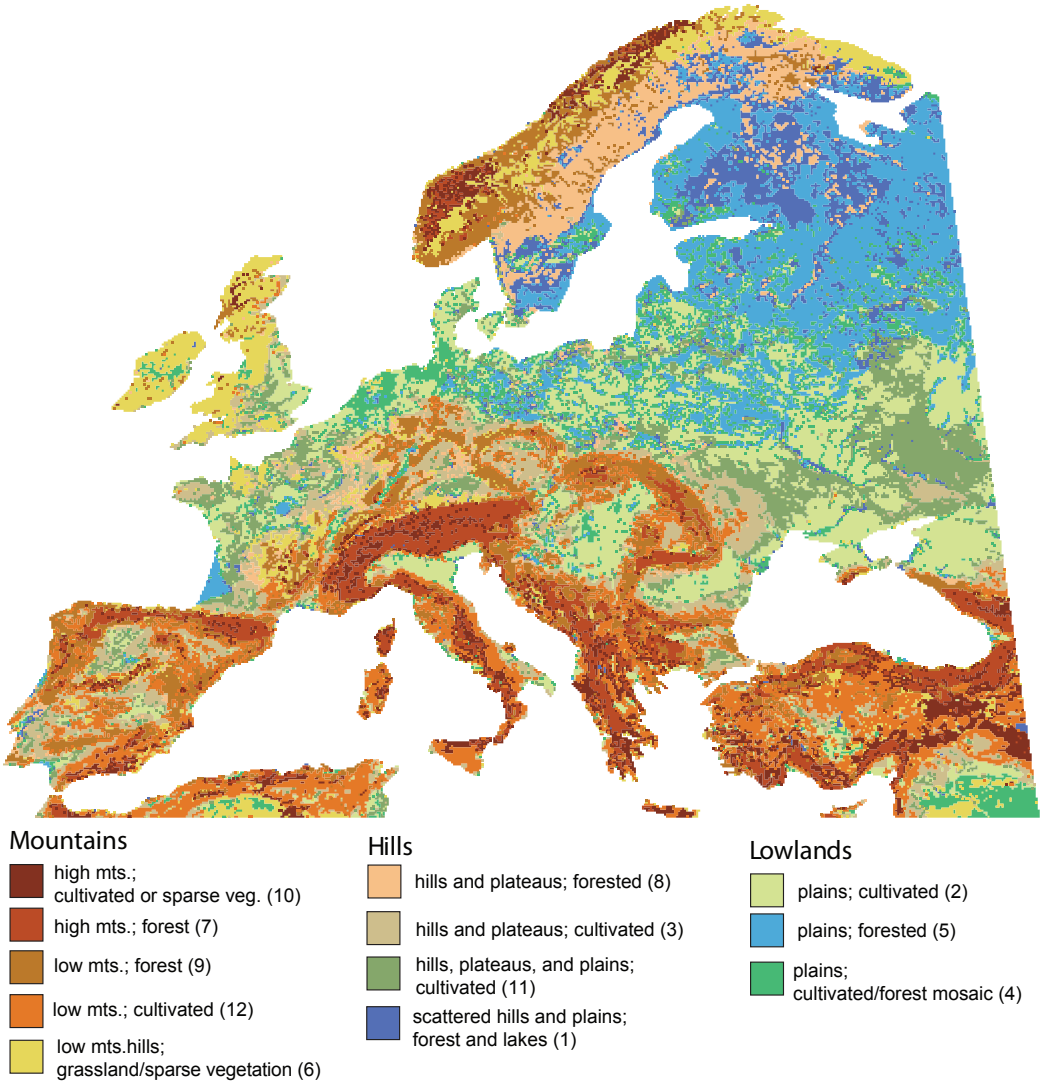

Figure 3. Map of 12 landscape types in Europe obtained by clustering 105,512 9 km×9 km LLs (INCOMA embeddings of patterns of land cover and landforms). Numbers following legend descriptions are numerical labels of LTs.

and LF. The resulting zones enclose characteristic patterns of LC and LF categories.

We used the two datasets described in Section 2 clipped to the coordinates of -13° West to 43° East and 35° North to 73° North to isolate Europe. Next, the study area is divided into 105,512 non-overlapping 9 km×9 km (30×30 cells) square blocks or LLs. Patterns of LC and LF in each LL are used to calculate INCOMA signature of LL. Identification of LTs is achieved via clustering LLs into a given number of clusters. To get insight into the appropriate number of clusters, we performed the t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton 2008) to visualize the structure of the multi-dimensional space of INCOMA signatures in a 2D. This was done to learn whether this space's structure is spatially intermittent (there are distinct clusters) or continuous, meaning that the space of signatures is stratified but without distinct gaps. The result showed a stratified structure, so the number of clusters is somewhat arbitrary. We used the K-means clustering with $K = 12$ and the JSD as a distance between signatures.

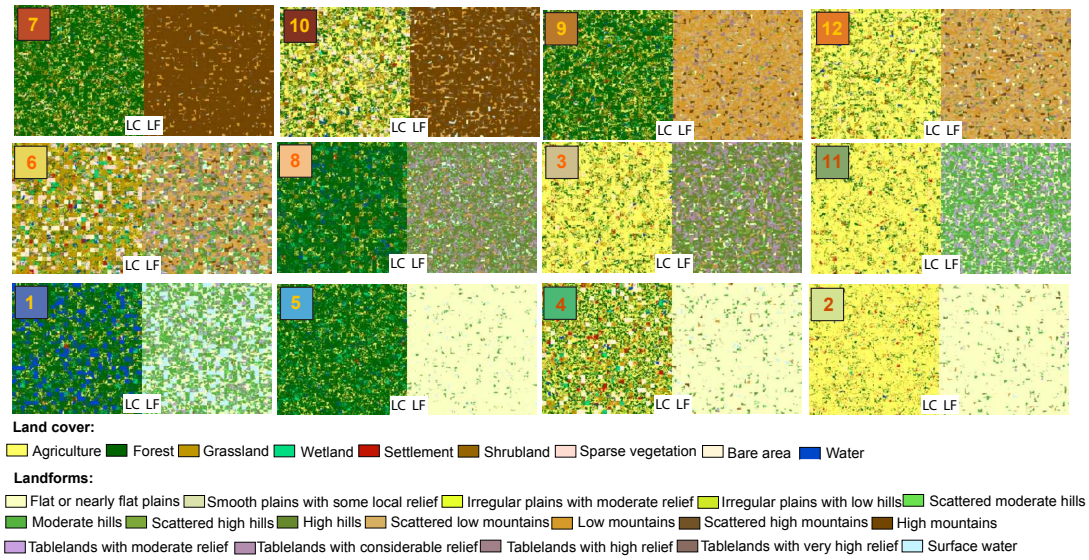

Figure 4. Patterns of land cover (left) and landforms (right) in each of 12 landscape types (LTs). These patterns are formed by 900 LLs randomly selected from a respective LT and organized into 30×30 grid.

Figure 3 shows the resultant map of LTs. The names of LTs in the legend are conceived a posteriori by interpreting LC and LF patterns in each LT. The LTs patterns are shown in Figure 4. This figure has twelve sections, each section shows a pattern of LC (left panel) and a pattern of LF (right panel) for an LT as indicated by a numerical label. These patterns are montages of 900 LLs randomly selected from a given LT and organized into 30×30 grid.

Examining Figure 4 reveals that there are five LTs (10, 7, 9, 12, and 6) that are mountainous to a different degree and covered by mosaics of predominantly forest, predominantly agriculture, or predominantly grassland. There are also four LTs (8, 3, 11, and 1) that are hilly and covered by mosaics of predominantly forest or predominantly agriculture. Finally, there are three LTs (2, 4, and 5) located in plains and covered by either forest or agriculture, or a forest-agriculture mosaic.

As we have stated in the Introduction, the quality of zones delineation depends on how similar are patterns' motifs within a single zone and how dissimilar are patterns' motifs between different zones. The intra-cluster dissimilarity of an LT is calculated using the metric, δ, based on an average dissimilarity between all LLs within a zone

$$\delta(\mathrm{LT}) = \frac{1}{k(k-1)} \sum_{i} \sum_{j \neq i} JSD(\mathrm{LL}_i, \mathrm{LL}_j) \tag{2}$$

where LL_i and LL_j are local landscapes within a given LT, and k is the number of LLs in this LT. Smaller values of δ indicate smaller dissimilarities between LLs in the LT.

The inter-cluster dissimilarity between a given LT and other LTs is calculated using the metric β based on a "distance" between two clusters, LT_1 and LT_2 is calculated using the average linkage (Sokal and Michener 1958).

$$D(\mathrm{LT}_1, \mathrm{LT}_2) = \frac{1}{k1k2} \sum_{i=1}^{k1} \sum_{j=1}^{k2} JSD(\mathrm{LL}_{1,i}, \mathrm{LL}_{2,j}) \tag{3}$$

Table 1. Statistics of landscape types (LTs) in Europe

zone	landscape characterization	% LLs	mean elev.	std. elev.	δ	β
10	high mts.; cultivated or sparse veg.	5.99	1093	739	0.45	0.81
7	high mts.; forested	6.93	1110	552	0.20	0.81
9	low mts.; forested	9.66	555	356	0.34	0.79
12	low mts.; cultivated	9.87	603	454	0.35	0.78
6	low mts./hills; grass/sparse veg.	6.72	451	424	0.71	0.84
8	hills and plateaus; forested	4.98	275	159	0.34	0.80
3	hills and plateaus; cultivated	8.41	322	266	0.36	0.78
11	hills, plateaus, and plains; cultivated	7.74	183	150	0.30	0.78
1	scattered hills and plains; forest and lakes	6.47	151	136	0.45	0.78
5	plains; forested	12.03	127	65	0.15	0.74
4	plains; cultivated/forest mosaic	7.87	133	134	0.33	0.72
2	plains; cultivated	14.34	138	162	0.11	0.80

Abbreviations: elev. – elevation (meters), std. – standard deviation, δ – inhomogeneity, β – distinction, mts. – mountains

where $k1$ and $k2$ are the numbers of LLs in LT_1 and LT_2, respectively. The value of the metric β for a given LT is an average of values of D between this LT and all other LTs. The larger the values of β, the more distinct the LT is from other LTs.

Table 1 lists statistics of LTs, percentage of total LLs, mean elevation, the standard deviation of elevation, δ, and β. Note that elevation was not a theme directly used in the delineation of LTs, values of the mean and standard deviation of elevations in each LT was calculated a posteriori from a digital elevation model (DEM). Values of metric β are high and have a small range. This means that LTs are distinct from each other. Values of metric δ have a large range, from as small as 0.11 for LT 2 to as large as 0.71 for LT 6. This is in agreement with visual evidence in Figure 4.

LTs characterized by small values of δ (2, 5, and 7) are landscapes with relatively simple motifs where LC/LF pattern has a scale smaller than the scale we have assumed for LLs (9 km). In such a case, all LLs in the zone have similar patterns and the value of δ is small. LTs characterized by moderate values of δ (9, 12, 8, 3, 11, and 4) are landscapes where a pattern of landforms or a pattern of land cover has a scale comparable or larger than the scale we have assumed for LLs. LLs in such LT differ in either their land cover patterns or their landform patterns, resulting in a moderate value of δ. However, together, they form a recognizable pattern. An example of such a pattern is LT 3, where a single motif of LC (a simple pattern dominated by an agricultural land) repeats itself throughout the zone, but there are two different motifs of LF each one repeating itself throughout the zone. Visual inspection reveals that despite being characterized by moderate values of δ these zones constitute proper landscape types, with a repeating motif of LF that happens to have a scale larger than our assumed scale of LL. In this case, LLs divide a single motif into two motifs.

LTs characterized by relatively large values of δ (6, 10, and 1) can be divided into two groups. In the case of LT 1, patterns of both LC and LF have structures comparable in size with an LL resulting in a large value of δ, but this LT still forms a cohesive landscape (see discussion in the previous paragraph). In the case of LT 10, and LT 6, patterns of LF are cohesive; LF pattern in LT 10 has a scale smaller than an LL, while LF pattern in LT 6 has a scale comparable to a LL. However, patterns of LC are less cohesive; they depend on geographic location. Most of zone 6 is covered by grassland, but zone 6 in Scandinavia is covered by bare areas and sparse vegetation. Similarly, most of zone 10 is covered by agriculture, but some LLs are covered by bare area. Thus, each of these two zones should be split into two zones.

5. Comparison to other methods

We compare a map of LTs in Europe constructed using the INCOMA-based method (hereafter referred to as M1) with maps of LTs constructed using two alternative methods hereafter referred to as M2 and M3. Method M2 uses COMA to identify and delineate pattern types in each theme separately. We clustered LLs carrying only LC or only LF patterns into 6 pattern zones each. The result is two maps, each having a resolution of 9 km, depicting types of patterns formed by categories of LC, and, separately, categories of LF. In the next step, these maps are overlaid to combine patterns of two themes resulting in 36 LTs. This method is a simplification of the INCOMA method resulting from neglecting all inter-thematic spatial dependencies. To the best of our knowledge, such a method has never been presented in the literature.

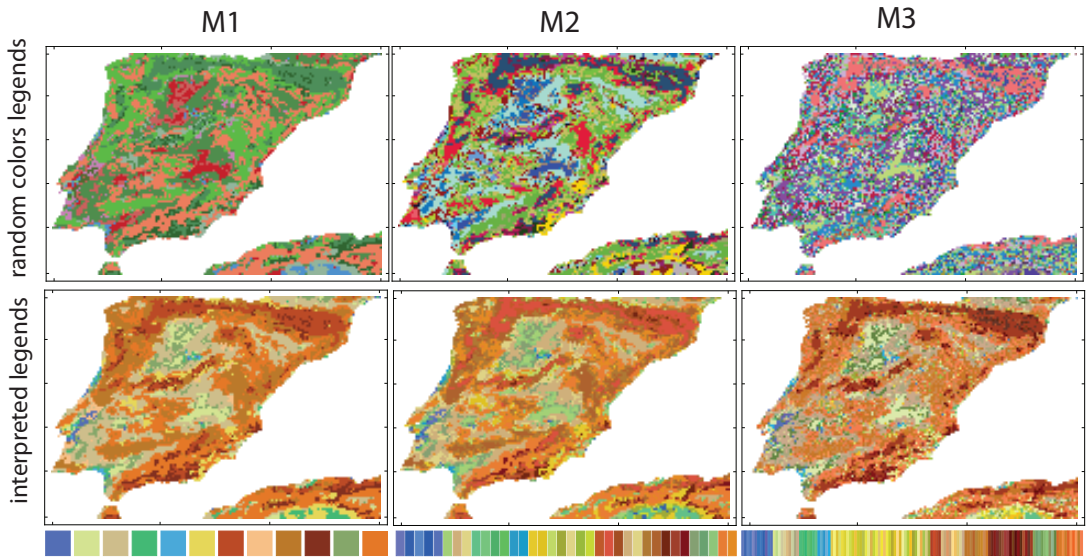

Figure 5. Comparison of landforms types mapping methods M1, M2, and M3 (panels in columns from left to right). The top row shows maps with random legends and the bottom row shows maps with interpreted legends (see main text for details).

Method M3 is an overlay of cell-based maps representing contributing variables. This method is frequently used in literature, in particular, Sayre *et al.* (2014) used it to map "ecophysiographic" zones over the entire terrestrial landmass and Mücher *et al.* (2010) used a variant of such method to map landscape types in Europe. We resampled LC and LF data to the resolution of 9 km (so M1, M2, and M3 have the same resolution). Overlaying LC with 9 categories with LF with 17 categories resulted in the map of LTs in Europe with 130 different LTs present.

Figure 5 shows a comparison of maps obtained by using M1, M2, and M3, respectively. In this figure, we have zoomed the maps to the extent of the Iberian peninsula to show more details. The figure has six panels organized in three columns (corresponding to M1, M2, and M3) and two rows (corresponding to random legends and interpreted legends). To construct an interpreted legend, we combined M2/M3 categories into 12 groups, so the combined spatial extents of categories in a group correspond as best as possible to a single zone in M1. Categories in a given group were assigned different but similar colors roughly corresponding to the hue of the matching M1 category (see legends in Figure 5). For example, the first five categories in M2 and the first 12 categories in M3 (dark bluish colors) correspond to the first category in M1.

Comparing maps with random-colored legends (top row) reveals that the M1 map is the most cohesive and integrated. Maps M2 and M3 are progressively less cohesive and integrated because they have more, possibly superfluous, categories. Comparing maps with interpreted legends shows that maps M2 and M3 are in principle capable of showing the same major placement of LTs in the Iberian peninsula as the map M1 does but only after merging some of their categories. However, such merging needs guidance by map M1 to be automatic. Without map M1 and without the ability to calculate values of dissimilarities between landscapes, the merging would have to be done manually.

6. Discussion

This paper's main contribution is the introduction of INCOMA – a signature of a multi-thematic pattern of natural themes that makes possible identification and delineation of landscape types using a principle of similarity-based aggregation of small local landscapes into large zones – landscape types. Thus, INCOMA brings together an interest in landscape classification with unsupervised machine learning methods like clustering and segmentation; it provides a principled way to make maps of such zones.

The current method of choice for mapping LTs is the map overlay. However, it is debatable whether zones delineated by the overlay method are indeed landscapes. This depends on the character of the input data. The commonly used data (land cover, topography, soil type) consist of variables defined at the cell level, whereas a landscape type is a variable defined over an areal unit much larger than a single cell. Formally, it is not correct to define a landscape type in terms of aggregates of cells having the same categories of variables. For example, in Sayre *et al.* (2014) the variables are land cover, landforms, lithology, and climate, all defined at the 250 m cell. The unit of "landscape", that Sayre *et al.* (2014) call an ecophysigraphic unit, is a homogeneous zone consisting of cells having the same combination of categories of the four variables – an entity not commonly perceived as a landscape type.

Omernik and Griffith (2014) developed maps of ecoregions over the conterminous United States. Because they used variables like geology, landforms, soils, climate, and land cover, technically, their ecoregions are not different from landscapes as defined in this paper. This work involved manual mapping by a large group of experts from diverse disciplines. They concluded that ecoregions/landscapes must be mapped based on patterns of variables rather than values of variables. By such reasoning, zones delineated by the map overlay method are not landscapes. On the other hand, zones delineated by INCOMA are indeed landscapes (see Figure 4).

The key parameter in the INCOMA method is the size of the local landscape. This size must be large enough to encompass a meaningful pattern of variables, but small enough to serve as a basic unit of aggregation into zones. Given the data we use, the minimum LL size could be perhaps as small as 3 km×3 km (10×10 cells). In Europe, LLs larger than 15 km×15 km would be too heterogeneous to lead to useful landscape types. In pattern-based approach landscape types depend on the size of LL, the larger the size, the more heterogeneous is the resultant LTs. Thus, the size of LL needs to be chosen, keeping in mind the resolution of the data and the purpose of mapping.

In our example, there is another important parameter - the number of identified LTs (equal to the number of clusters in K-means clustering). We used 12 LTs after experimentation with several different choices. In a more comprehensive study (our example in this paper is only an illustration of INCOMA application), segmentation rather

then clustering would be used to aggregate LLs (Jasiewicz, Stepinski, and Niesterowicz 2018). Resultant segments must be subsequently clustered to get LTs. Because the number of segments would be much smaller than the number of LLs, it would be feasible to use a hierarchical clustering (which requires storing a distance matrix in a computer memory) instead of K-means, which would alleviate the problem of selecting a "right" number of clusters.

Note that we judged the quality of landscape types delineation by the coherence of LC and LF patterns within them and by uniqueness of each zone. This is a standard method when assessing the quality of clustering. If a zone has a coherent pattern and this pattern changes perceptibly on its boundaries, it represents a recognizable geographical unit. We have found that the standard assessment, which was developed for pixels rather than LLs (units of the pattern), is necessary but insufficient to judge the quality of our zones. This is because the zone may enclose a pattern whose scale is larger than our assumed scale of a LL. Thus, some zones with moderate values of metric δ may nevertheless constitute a coherent pattern. This has been the case for our zones 9, 12, 8, 3, 11, and 4. Therefore, a visual inspection of the zones vis-a-vis individual data layers is recommended.

There were previous attempts to delineate multi-thematic landscapes manually, e.g., Wickham and Norton (1994) or (Omernik and Griffith 2014). Omernik and Griffith presented a large work of dividing the area of the entire conterminous U.S. into ecoregions based on multiple variables, including land cover and landforms. Their definition of the ecoregion is very similar to our definition of a multi-thematic landscape. Thus, using our INCOMA method on data for the US, it would be in principle possible to compare results of automatic and manual delineations.

In our illustrative example, we only used two natural themes, LC and LF, however, in a comprehensive study, LTs should be determined by a minimum of four themes, land cover, landforms, soils/lithology, and climate (Mücher *et al.* 2010; Sayre *et al.* 2014). Because climate changes on scales larger than other themes, it is not necessary to include it in the process of LTs delineation (Mücher *et al.* 2010); LTs can to be delineated on the basis of the remaining three themes and later subdivided by climatic zones. The INCOMA can easily accommodate three themes. Recall from Section 3.3 that INCOMA's size is $(C_1 + \cdots + C_n)^2$, where C_k is the length of the legend of theme k. In our example, this size is $(9 + 16)^2 = 625$, although a large fraction of these counts is equal to 0. If we add soils theme with say 10 categories the size of INCOMA increase to 1225, again with a large fraction of these number being equal to 0. We do not expect the need for INCOMA to encapsulate more than four themes, with the increase of the number of themes, the zones of LTs become small and thus not very useful for a majority of applications.

Finally, INCOMA can find application in tasks other than identification and mapping of landscapes, which also require calculating similarities between multi-thematic patterns. One such application is content-based search and retrieval in spatial databases (Sharifi 1999; Terribile *et al.* 2015). The purpose of search and retrieval is to find LLs most similar to the query local landscape (LL_0). This is achieved by calculating INCOMA-based dissimilarities between LL_0 and each LL in the area of interest and select LLs characterized by the smallest dissimilarities values as the answer to the query. Recently, Peng *et al.* (2019) presented a similarity-based method for query and retrieval of bi-thematic "local landscapes" consisting of a land cover theme and a terrain (digital elevation model or DEM) theme. Their purpose was to retrieve similar Earth images using bi-thematic patterns as proxies for images. INCOMA enables more general queries based on multi-thematic local landscapes.

Data and code availability statement

Three global categorical raster datasets used in this study were derived from the following resources: http://maps.elie.ucl.ac.be/CCI/viewer/, https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover, and https://rmgsc.cr.usgs.gov/outgoing/ecosystems/Global/.

The data that support the findings of this study are available at https://doi.org/10.6084/m9.figshare.13379228.v1.

The R package allowing to create integrated co-occurrence matrices and integrated co-occurrence histograms is available at https://github.com/Nowosad/comat. The R package allowing for identification of landscape types (LTs) is available at https://github.com/Nowosad/motif.

References

Barnsley, M J, and S L Barr. 1996. "Inferring Urban Land Use from Satellite Sensor Images Using Kernel-Based Spatial Reclassification." *Photogrammetric engineering and remote sensing* 62 (8): 949–958.

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

Cardille, Jeffrey A, and Marie Lambois. 2010. "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.

Cha, Sung-Hyuk. 2007. "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions." *International Journal of Mathematical Models and Methods in Applied Sciences* 1 (4): 300–307.

Chang, Peng, and J. Krumm. 1999. "Object Recognition with Color Cooccurrence Histograms." In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, Fort Collins, CO, USA, 498–504. IEEE Comput. Soc.

Drost, HG. 2018. "Philentropy: Information Theory and Distance Quantification with R." *Journal of Open Source Software* 3 (26): 765. http://joss.theoj.org/papers/10.21105/joss.00765.

ECMWF. 2019. *ICDR Land Cover Product User Guide and Specification*. Technical Report.

Eddelbuettel, Dirk, and Romain François. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40 (8): 1–18. http://www.jstatsoft.org/v40/i08/.

Eddelbuettel, Dirk, and Conrad Sanderson. 2014. "RcppArmadillo: Accelerating R with high-performance C++ linear algebra." *Computational Statistics and Data Analysis* 71: 1054–1063. http://dx.doi.org/10.1016/j.csda.2013.02.005.

Haralick, Robert M., K. Shanmugam, and Its'Hak Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3 (6): 610–621.

Heikkinen, R. K., M. Luoto, R. Virkkala, and K. Rainio. 2004. "Effects of habitat cover, landscape structure and spatial variables on the abundance of birds in an agricultural-forest mosaic." *Journal of Applied Ecology* 41: 824–835.

Hengl, T., J. M. deJesus, G. B. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagotic, W. Shangguan, *et al.* 2017. "SoilGrids250m: Global gridded soil information based on machine learning." *PloS One* 12(2): e0169748.

Jasiewicz, Jaroslaw, Paweł Netzel, and Tomasz Stepinski. 2015. "GeoPAT: A Toolbox for Pattern-Based Information Retrieval from Large Geospatial Databases." *Computers & Geosciences* 80: 62–73.

Jasiewicz, Jaroslaw, Tomasz Stepinski, and Jacek Niesterowicz. 2018. "Multi-Scale Segmenta-

tion Algorithm for Pattern-Based Partitioning of Large Categorical Rasters." *Computers & Geosciences* 118: 122–130.

Karagulle, Deniz, Charlie Frye, Roger Sayre, Sean Breyer, Peter Aniello, Randy Vaughan, and Dawn Wright. 2017. "Modeling Global Hammond Landform Regions from 250-m Elevation Data." *Transactions in GIS* 21 (5): 1040–1060.

Lin, Jianhua. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37 (1): 145–151.

Liu, J., B. Luo, Q. Qin, and G. Yang. 2017. "Alike scene retrieval from land-cover products based on the label co-occurrence matrix (LCM)." *Remote Sens.* 9: 912.

Long, Jed, Trisalyn Nelson, and Michael Wulder. 2010. "Regionalization of landscape pattern indices using multivariate cluster analysis." *Environmental Management* 46 (1): 134–142.

Maaten, L. V. D., and G. Hinton. 2008. "Visualizing data using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605.

Metzger, M. J., R. H. Jongman R. G. Bunce, R. Sayre, A. Trabucco, and A. Zomer. 2013. "A high-resolution bioclimate map of the world: A unifying framework for global biodiversity research and monitoring." *Global Ecology and Biogeography* 22 (5): 630–638.

Mücher, C. A., J. A. Klijn, D. M. Wascher, and J. H. Schaminée. 2010. "A new European Landscape Classification (LANMAP): A transparent, flexible and user-oriented methodology to distinguish landscapes." *Ecological indicators* 10(1): 87–103.

Netzel, P., and T. F. Stepinski. 2014. "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., and T. F. Stepinski. 2013. "Regionalization of multi-categorical landscapes using machine vision methods." *Applied Geography* 45: 250–258.

Niesterowicz, J., and T. F. Stepinski. 2016. "On using landscape metrics for landscape similarity search." *Ecological indicators* 64: 20–30.

Niesterowicz, J., and T. F. Stepinski. 2017. "Pattern-based, multi-scale segmentation and regionalization of EOSD land cover." *International journal of Applied Earth Observation and Geoinformation* 62: 192–200.

Niesterowicz, J., T.F. Stepinski, and J. Jasiewicz. 2016. "Unsupervised regionalization of the United States into landscape pattern types." *International Journal of Geographical Information Science* 30(7): 1450–1468.

Nowosad, J., T. F. Stepinski, and P. Netzel. 2019. "Global Assessment and Mapping of Changes in Mesoscale Landscapes: 1992–2015." *International Journal of Applied Earth Observation and Geoinformation* 78: 332–340.

Nowosad, J., and T. GF. Stepinski. 2018. "Global inventory of landscape patterns and latent variables of landscape spatial configuration." *Ecological Indicators* 89: 159–167.

Nowosad, Jakub. 2020. *comat: Co-Occurrence Matrices of Spatial Data*. R package version 0.8.4, `https://nowosad.github.io/comat/`.

Nowosad, Jakub. 2021. "Motif: an open-source R tool for pattern-based spatial analysis." *Landscape Ecology* 36: 26–43.

Omernik, J. M., and G. E. Griffith. 2014. "Ecoregions of the Conterminous United States: Evolution of a Hierarchical Spatial Framework." *Environmental Management* 54 (6): 1249–1266.

O'Neill, R. V., J. R. Krummel, R. H. Gardner, G. Sugihara, B. Jackson, D. L. DeAngelis, B. T. Milne, *et al.* 1988. "Indices of Landscape Pattern." *Landscape Ecology* 1 (3): 153–162.

Partington, Kevin, and Jeffrey Cardille. 2013. "Uncovering Dominant Land-Cover Patterns of Quebec: Representative Landscapes, Spatial Clusters, and Fences." *Land* 2: 756–773. `http://www.mdpi.com/2073-445X/2/4/756/`.

Pebesma, Edzer. 2020. *stars: Spatiotemporal Arrays, Raster and Vector Data Cubes*. Https://r-spatial.github.io/stars/, https://github.com/r-spatial/stars/.

Peng, Feifei, Le Wang, Shengyuan Zou, Jing Luo, Shengsheng Gong, and Xiran Li. 2019. "Content-Based Search of Earth Observation Data Archives Using Open-Access Multitemporal Land Cover and Terrain Products." *International Journal of Applied Earth Observa-*

tion and Geoinformation 81: 13–26.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Rubner, Y., J. Puzicha, C. Tomasi, and J. M. Buhmann. 2001. "Empirical Evaluation of Dissimilarity Measures for Color and Texture." *Computer Vision and Image Understanding* 84: 25–43.

Sayre, R., J. Dangermond, C. Frye, R. Vaughan, P. Aniello, S. Breyer, D. Cribbs, *et al.* 2014. *A new map of global ecological land units – an ecophysiographic stratification approach*. Technical Report. Tech. rep., Washington, DC: Association of American Geographers.

Shannon, C E. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27: 55.

Sharifi, Ali. 1999. "Remote Sensing and Decision Support Systems." In *Spatial Statistics for Remote Sensing*, edited by Freek van der Meer, Alfred Stein, Freek Van der Meer, and Ben Gorte, Vol. 1, 243–260. Dordrecht: Springer Netherlands.

Simensen, T., R. Halvorsen, and L. Erikstad. 2018. "Methods for landscape characterisation and mapping: A systematic review." *Land use policy* 75: 557–569.

Sokal, R. R., and C. Michener. 1958. "A statistical method for evaluating systematic relationships." *Univ. Kansas Sci. Bull.* 38: 1409–1438.

Stepinski, T. F., P. Netzel, and J. Jasiewicz. 2013. "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.

Terribile, F., A. Agrillo, A. Bonfante, G. Buscemi, M. Colandrea, A. D'Antonio, R. De Mascellis, *et al.* 2015. "A Web-Based Spatial Decision Supporting System for Land Management and Soil Conservation." *Solid Earth* 6 (3): 903–928.

Turner, Monica G, and Robert H Gardner. 1991. *Quantitative Methods in Landscape Ecologythe Analysis and Interpretation of Landscape Heterogeneity*.

Vadivel, A., Shamik Sural, and A.K. Majumdar. 2007. "An Integrated Color and Intensity Co-Occurrence Matrix." *Pattern Recognition Letters* 28 (8): 974–983.

Wascher, D. M. 2005. *European landscape character areas: typologies, cartography and indicators for the assessment of sustainable landscapes (No. 1254)*. Technical Report. Landscape Europe.

Wickham, J. D., and D. J. Norton. 1994. "Mapping and analyzing landscape patterns." *Landscape Ecology* 9 (1): 7–23.