motif: an open-source R tool for pattern-based spatial analysis

Jakub Nowosad

2020-10-14

Abstract

Context Pattern-based spatial analysis provides methods to describe and quantitatively compare spatial patterns for categorical raster datasets. It allows for spatial search, change detection, and clustering of areas with similar patterns.
Objectives We developed an R package **motif** as a set of open-source tools for pattern-based spatial analysis.
Methods This package provides most of the functionality of existing software (except spatial segmentation), but also extends the existing ideas through support for multi-layer raster datasets. It accepts larger-than-RAM datasets and works across all of the major operating systems.
Results In this study, we describe the software design of the tool, its capabilities, and present four case studies. They include calculation of spatial signatures based on land cover data for regular and irregular areas, search for regions with similar patterns of geomorphons, detection of changes in land cover patterns, and clustering of areas with similar spatial patterns of land cover and landforms.
Conclusions The methods implemented in **motif** should be useful in a wide range of applications, including land management, sustainable development, environmental protection, forest cover change and urban growth monitoring, and agriculture expansion studies. The **motif** package homepage is https://nowosad.github.io/motif.

Keywords: spatial patterns, multi-layer similarity, query-by-example, similarity search, patterns comparison, patterns clustering

Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Krygowskiego 10, 61-680 Poznan, Poland, email: nowosad.jakub@gmail.com

1 Introduction

Discovering and describing spatial patterns is an important element of many environmental studies, as spatial patterns on different scales are related to ecological processes. With the

ascent of computational methods, identification and characterization of spatial patterns started to be answered with numbers rather than by means of qualitative depiction. Most methods of spatial analysis for remotely sensed data treat single cells as basic units of analysis, and while this standard approach is sufficient for analysis of local areas, it is not well suited for analysis on regional, continental, or global scales. This is because cell-scale information is, in itself, irrelevant for broad-scale analysis. For example, mapping land cover change using cell-based transitions in a large area results in an unsatisfactory salt-and-pepper output due to classification errors, or simply because cell size is smaller than objects whose change we want to detect.

A "local landscape" is a more appropriate unit of analysis for broad-scale studies. For categorical raster data is represented by a block of cells containing a local pattern of a cell-based variable. Pattern-based spatial analysis is a set of ideas and methods allowing for the description of spatial patterns and calculation of similarity between patterns. The core idea is to transform the data from a large raster consisting of cells having simple content (a single value) into a spatial signature - a statistical description of a pattern. A large number of landscape metrics were developed to quantify spatial patterns (O'Neill et al. 1988; Turner and Gardner 1991; Li and Reynolds 1993; He, DeZonia, and Mladenoff 2000; Jaeger 2000; McGarigal 2014; Nowosad and Stepinski 2019), and were implemented in the existing software (McGarigal et al. 2002; Hesselbarth et al. 2019). However, most of landscape metrics are a single number depicting specific characteristics of a local landscape. Spatial signatures, on the other hand, are multi-values representations of landscape composition and configuration, and therefore can be compared using a large number of existing distance or dissimilarity measures (Lin 1991; Cha 2007). This enables spatial analysis such as search, change detection, clustering, and segmentation.

Ideas related to the pattern-based spatial analysis were first tested in a collection of GRASS GIS modules called GeoPAT (Jasiewicz, Netzel, and Stepinski 2015). GeoPAT allowed for pattern-based search and change detection for a single layer raster data using a limited number of spatial signatures. It was also limited to a Linux operating system, as GRASS needs to be compiled together with GeoPAT. Further improvements led to the release of GeoPAT 2 (Netzel et al. 2018), which is a standalone command-line software, with versions for Linux and Windows. There were also a number of changes compared to the first version, including a new segmentation module, and experimental support for analysis of patterns in time-series spatial data.

Both GeoPAT and GeoPAT 2 support analysis on single-layer categorical rasters (e.g. a map of land cover or a map of geomorphons). Recently, new ideas for working on multi-layer categorical rasters were developed. It includes new spatial signatures, the weighted co-occurrence matrix (*wecoma*) that allows applying numerical weights for a categorical raster (Dmowska, Stepinski, and Nowosad 2020) and the integrated co-occurrence matrix (*incoma*) that encapsulates not only spatial patterns of many input layers but also spatial relationships between the layers (Vadivel, Sural, and Majumdar (2007)). This signature can be also converted into the normalized integrated co-occurrence histogram, allowing for content-based searches, comparisons, or clusterings.

The aim of this paper is to present a new open-source software called **motif**. Contrary

to GeoPAT and GeoPAT 2, it is a fully cross-platform, working on Windows, Linux, and macOS machines. This R package provides tools to derive several types of implemented spatial signatures based on regular and irregular regions. It includes single layer signatures like composition or a co-occurrence histogram, as well as, multi-layer signatures, e.g. an integrated co-occurrence matrix. The **motif** package also allows users to create and use their own spatial signatures. Importantly, it has functions allowing for pattern-based search, comparison, and clustering. Additionally, it is integrated with robust R packages for spatial data representation, namely **stars** and **sf** (Pebesma 2020, 2018), and therefore operations in this package can be added easily to existing workflows or be a basis for new workflows. Finally, the computationally demanding parts of the software were written in C++ (Eddelbuettel and François 2011), which together with a larger-than-RAM support, allows for working on high-resolution rasters or data on continental or global scales.

2 Pattern-based analysis

The main idea of the pattern-based analysis is that any categorical raster dataset can be described using some spatial signatures. A large number of possible signatures were developed (Table 1). The most basic one is a composition, which describes a number of cells of each category in a local landscape. Gustafson (1998) highlighted that alongside composition there is also a second fundamental element of spatial patterns - spatial configuration. Riitters (2019) stated that amount (composition) is a more fundamental metric than adjacency (configuration), however, he also underlines that they both are crucial for complete pattern description. Similarly, Remmel (2009) showed that a given composition could result in a large number of possible configurations. Therefore, the use of more complex signatures is desirable in many cases. One of the signatures encapsulating both composition and spatial configuration is a co-occurrence matrix. It is a `k` by `k` matrix, where `k` is a number of classes in the categorical raster, constructed by counting all of the pairs of the adjacent cells (Haralick, Shanmugam, and Dinstein 1973; Jasiewicz, Netzel, and Stepinski 2015). In it, diagonal values are related to the composition and non-diagonal ones are related to the configuration. Recently, extended versions of the co-occurrence matrix were developed for spatial data analysis. It includes a weighted co-occurrence matrix (*wecoma*) (Dmowska, Stepinski, and Nowosad 2020) and an integrated co-occurrence matrix (*incoma*) (Vadivel, Sural, and Majumdar (2007)). The weighted co-occurrence matrix is created based on two raster datasets. The first one is used to find adjacencies between classes, while the second one provides weights for each adjacency. It allows to put more importance to some locations in the data than others. Integrated co-occurrence matrix, on the other hand, allows incorporating spatial patterns of many rasters at the same time as it is created based on two or more rasters. This signature consists of independent co-occurrence matrices for each input dataset, but also co-located co-occurrence matrices for each combination of input datasets. The co-located co-occurrence matrix uses two matrices and counts adjacencies between the cells in the first dataset and related cells in the second dataset. Therefore, this signature incorporates not only the spatial patterns of all of the input layers but also the spatial relationship between layers.

Table 1: Spatial signatures available in the motif package

Signature	Abbreviation	Input data	Description
Composition	composition	One categorical raster	A vector representing the shares of cells for each category in a local landscape
Co-occurrence matrix	*coma*	One categorical raster	A matrix representing composition and spatial configuration by counting all of the pairs of the adjacent cells for each category in a local landscape
Co-occurrence histogram (vector)	cove	One categorical raster	A vector containing a normalized form of the co-occurrence matrix used for spatial search, comparison, and clustering
Weighted co-occurrence matrix	*wecoma*	One categorical and one numerical raster (weight matrix)	A modification of a co-occurrence matrix, in which each adjacency contributes to the output based on the values from the weight matrix
Weighted co-occurrence histogram (vector)	wecove	One categorical and one numerical raster (weight matrix)	A vector containing a normalized form of the weighted co-occurrence matrix used for spatial search, comparison, and clustering
Integrated co-occurrence matrix	*incoma*	Two or more categorical rasters	A matrix representing composition and spatial configuration of all of the input rasters, but also spatial interactions between them
Integrated co-occurrence histogram (vector)	incove	Two or more categorical rasters	A vector containing a normalized form of the integrated co-occurrence matrix used for spatial search, comparison, and clustering
User-defined signature	-	Depends on the user-defined function	Any user-defined function that can summarize stars objects

Spatial signatures can be calculated for either regular areas or irregular ones. A basic example of using regular areas is when a large region is divided into a number of smaller, square-sized regions. The size of the smaller regions can be provided as a number of cells in each direction. This number, however, should be large enough to be able to encapsulate meaningful local spatial patterns (Boots 2003). The pattern-based approach can be also applied to regions having several different sizes, and thus testing how the spatial patterns are scale-dependent. Another approach is to use irregular regions, which can be results of the previous studies (e.g. ecoregions, climate regionalizations) or represents some existing borders, such as counties, states, or countries.

The signatures store compressed information about local spatial patterns. They can be used in several possible workflows. First, several information theory metrics can be calculated for all types of co-occurrence matrices (Nowosad and Stepinski 2019). It includes marginal entropy, representing the diversity of spatial categories or relative mutual information quantifying the clumpiness of spatial categories. Second, all of the matrix signatures can be also transformed into a one-dimensional form - a normalized histogram (vector) representation (Table 1). A normalized histogram representation is created by restructuring a 2-dimensional matrix into a 1-dimensional vector, and next, by normalizing the vector to sum to one. This representation can be used as an input for a large number of existing distance or dissimilarity measures that allow comparing histograms of values (Lin 1991; Cha 2007). They make it possible to determine how similar patterns of different local landscapes (or precisely their signatures) are to each other.

It opens a several groups of possible applications, including spatial pattern search, comparison, and clustering. The spatial search takes a query area, computes a selected spatial signature based on the provided categorical data for this area, and calculates the spatial signatures for another dataset, usually divided into regular or irregular areas. Next, the distance/dissimilarity is calculated between the spatial signature for the area of interest and each of the areas in the second dataset. The results shows which areas have the most and the least similar spatial patterns to our area of interest based on the provided parameters (e.g., signature type, distance metrics). Spatial comparison, often called a change detection, also accepts two sets of data, where both should have the same resolution and extent. These two sets of data can be divided into many spatially consistent regular or irregular regions. Selected signature is calculated for each subregion in both sets of data, and a pair of signatures in each region is compared using the given distance/dissimilarity measures. This allows locating regions without any change in spatial patterns between datasets or regions that have very different spatial patterns in two sets of data. The third possible application involves clustering, in which regular or irregular regions are joined into several (possibly multi-parts) groups of similar spatial patterns. This process starts by calculating a distance matrix that is a result of measuring distance/dissimilarity between signatures of all provided regions. Next, one of a large number of possible clustering techniques can be applied, including hierarchical or fuzzy clustering. The result has the same extent as input data but divides the whole area into several clusters.

3 Software

3.1 Software design

The open-source R (R Core Team 2020) package **motif** provides tools for pattern-based spatial analysis. The package builds upon two robust R packages: **stars** representing spatial raster data (Pebesma 2020) and **sf** for vector data representation (Pebesma 2018). Most functions in this package use computationally fast and memory-efficient C++ code (Eddelbuettel and François 2011; Eddelbuettel and Sanderson 2014; Nowosad 2020). Moreover, the **motif** package also supports using larger-than-RAM raster datasets. All **motif** functions are consistently named using the `lsp_` (*l*ocal *s*patial *p*attern) prefix, which allows to quickly find any relevant function (Table 2).

The **motif** package describes spatial patterns of one or more categorical raster data for any defined regular and irregular regions. Patterns are represented quantitatively by built-in signatures based on co-occurrence matrices but any user-defined functions are also allowed. The main structure in this package is an extended data frame of a special class `lsp` (Müller and Wickham 2020). This data frame has several columns: `id` - unique identifier of each defined region, `na_prop` - a share (0-1) of NA cells in each defined region, and `signature` - a list-column containing calculated signatures, where each row relates to one of the defined regions. Two examples of the `lsp` objects are presented in section 4.1.

The package enables various types of pattern-based spatial analyses, such as search (section 4.2), comparison (change detection) (section 4.3), and clustering (section 4.4). In them, the similarity of spatial patterns between given regions is represented by a distance or dissimilarity measure between their spatial signatures. **motif** allows using 46 optimized distance and similarity measures implemented in the **philentropy** package (Drost 2018).

Each function in the package has an extensive help file containing a list of possible arguments and a set of examples. The source code of this package is thoroughly tested, with about 96% of the code lines executed using the automated tests. The package also has a website at https://nowosad.github.io/motif/, which contains installation instructions, documentation, examples on how to create user-defined functions, and several vignettes.

3.2 Software capabilities

Table 2 contains a list of the functions from the **motif** package. The functions are divided into several groups of application: (1) description, (2) search, (3) comparison and change detection, (4) clustering, and (5) various.

The `lsp_signature()` function accepts input data in form of an object of class `stars` with one or more attributes. It creates spatial signatures for either entire of the provided dataset (default), any defined regular or irregular regions. Regular regions are defined by a number expressing a length of the side of a square-shaped block of cells in the number of cells. For example, in case of providing a value of 100, each regular region will consist of 100 by 100 non-overlapping cells. Spatial vector objects of class `sf` can be used to define

Table 2: Overview of the function in the motif package

Function	Application	Description
lsp_signature()	Description	Creates spatial signatures
lsp_search()	Search	Searches for similar spatial pattern
lsp_compare()	Comparison or change detection	Compares spatial patterns
lsp_to_dist()	Clustering	Calulates distance matrix between spatial patterns
lsp_add_clusters()	Clustering	Adds custers' ids to a lsp object
lsp_add_quality()	Clustering	Calculates quality metrics for all of the clusters
lsp_extract()	Various	Extracts a local landscape based on a provided id
lsp_add_sf()	Various	Creates vector object (of the sf class) based on the input object or a set of parameters
lsp_add_stars()	Various	Creates raster object (of the stars class) based on the input object or a set of parameters

irregular regions. The `lsp_signature()` function has several built-in signatures, including `"composition"` that counts proportions of categories in a local landscape, `"coma"` - a co-occurrence matrix and `"cove"` - a co-occurrence histogram, `"wecoma"` and `"wecove"` - a weighted co-occurrence matrix/histogram, and `"incoma"` and `"incove"` - an integrated co-occurrence matrix/histogram. Additionally, this function accepts user-defined functions, which should allow only one argument in a form of a list containing one or more matrices (Table 1). The `lsp_signature()` function is also internally used in the functions for spatial search, comparison, and clustering.

The `lsp_search()` function performs query by example, searching for areas with similar spatial patterns in categorical data. It accepts a categorical raster dataset with one or more attributes (layers) and compares it to the second (usually larger) dataset with the same attributes. The first dataset can be either compared to the second dataset as a whole, divided into regular regions, or divided into irregular regions. A selected signature is calculated for the first dataset and all of the regions in the second dataset. Next, a distance between the signature of the first dataset and each of the signatures for the second dataset is calculated using selected measure available in the `philentropy::distance()` function. Additional parameters, such as neighborhood or normalization types, are also available.

The `lsp_compare()` function compares two spatial datasets with one or more layers containing categorical raster data for the same area. Similarly to the previous functions, both datasets can be either compared to as whole areas, areas divided into regular regions, or areas divided into irregular regions. Spatial signatures are calculated for all regions in both datasets and a distance between them is calculated using a selected measure from the `philentropy::distance()` function. Additional parameters, such as neighborhood or

normalization types, are also available.

The **motif** package makes it possible to find clusters of areas with similar spatial patterns. This process has a few steps. First, a spatial signature must be derived using the `lsp_signature()` function for the defined regular and irregular areas. Next, `lsp_to_dist()` should be used to calculate a distance between each of the areas. Its output is a distance matrix accepted by many existing R functions for clustering. It includes different approaches of hierarchical clustering (`hclust()`, `cluster::agnes()`, `cluster::diana()`) or fuzzy clustering (`cluster::fanny()`) (Kaufman and Rousseeuw 1990; Maechler et al. 2019). Based on the obtained vector with group memberships (clusters), the `lsp_add_clusters()` function adds clusters' ids back to a `lsp` object, creating a new spatial object. The output can be of `stars` or `sf` class. The **motif** package also allows us to calculate three quality metrics to evaluate spatial patterns' clustering: (1) inhomogeneity, (2) isolation, and (3) overall quality. Inhomogeneity measures a degree of mutual dissimilarity between all objects in a cluster. This value is between 0 and 1, where the small value indicates that all objects in the cluster represent consistent patterns so the cluster is pattern-homogeneous. Isolation is an average distance between the focus cluster and all of its neighbors. This value is between 0 and 1, where the large value indicates that the cluster stands out from its surroundings. Overall quality is calculated as $1 - (inhomogeneity/isolation)$. This value is also between 0 and 1, where higher values indicate better quality (Haralick and Shapiro 1985; Jasiewicz, Stepinski, and Niesterowicz 2018).

The above applications can be applied to many areas (local landscapes). The `lsp_extract()` function makes it easier to extract, visualize, and analyze a single area based on the provided id. For example, it allows us to select landscapes the most similar to the query one based on the result of `lsp_search()`, areas with the largest change obtained using `lsp_compare()`, or examples of members of a selected cluster. The `lsp_extract()` function expects three arguments: `x` - an input `stars` object, window - a numeric value or an `sf` object, and `id` an id number of selected area.

Additionally, the `lsp_add_sf()` and `lsp_add_stars()` functions create spatial objects based on a given set of parameters or provided `lsp` objects. It allows, for example, to visualize created regular or irregular grids.

4 Case studies

Four case studies are presented in the next sections to describe different capabilities of the **motif** package, including calculation of spatial signatures (section 4.1), spatial pattern search (section 4.2), spatial pattern comparison (section 4.3), and spatial pattern clustering (section 4.4). Complete code and data to recreate all of the case studies is available at https://github.com/Nowosad/motif-examples.

The case studies were based on four raster datasets. It includes the CCI land cover map for the year 1994, the C3S land cover map for the year 2018, European Digital Elevation Model (EU-DEM), and the Hammond's landform regions (European Space Agency 2017;

European Centre for Medium-Range Weather Forecasts 2019; European Environment Agency 2016; Karagulle et al. 2017). These datasets were derived from the following resources: http://maps.elie.ucl.ac.be/CCI/viewer/, https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover, https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1, and https://rmgsc.cr.usgs.gov/outgoing/ecosystems/Global/.

The CCI land cover data for the year 1994, C3S land cover data for the year 2018 were reclassified into nine broader IPCC (Intergovernmental Panel on Climate Change) categories. The land cover and landform regions data were also reprojected into the interrupted Goode homolosine projection and resampled into the resolution of 300 meters.

European Digital Elevation Model was cropped into the area of Poland and used as a basis for the calculations of geomorphons (Jasiewicz and Stepinski 2013). Geomorphons, used in section 4.2, classify a digital elevation model into one of ten most common terrain forms: flat, peak, ridge, shoulder, spur, slope, hollow, footslope, valley, and pit.

Besides the **motif** package, the case studies require the **stars** and **sf** packages to read spatial data (Pebesma 2020, 2018).

```
library(motif)
library(sf)
library(stars)
```

All of the following visualizations were created using the **tmap** package (Tennekes 2018).

4.1 Basic signatures

The basic functionality of the **motif** package is to derive spatial signatures based on the input data. Derived spatial signatures are also often used as an intermediate step in the other types of spatial pattern-based analysis, as well as a source of values to calculate some metrics related to spatial composition and configuration. For example, the co-occurrence matrix derived with the **motif** package can be an input for the calculation of the information theory metrics, such as marginal entropy (diversity) or relative mutual information (clumpiness) (Nowosad and Stepinski 2019). That being said, the most basic use of spatial signatures is to summarize the data. In this case study, we show how the spatial signatures are calculated and stored in the **motif** package. It is presented using two possible approaches - one with regular rectangular regions (Figures 1) and one with irregular regions (Figures 2).

We used the land cover classification for New Guinea for this purpose. The island is covered mostly by forest, but also has areas with six additional land cover classes, including agriculture or grasslands (Figures 1, 2). This dataset was read into R using the **stars** package.

```
landcover = read_stars("landcover.tif")
```

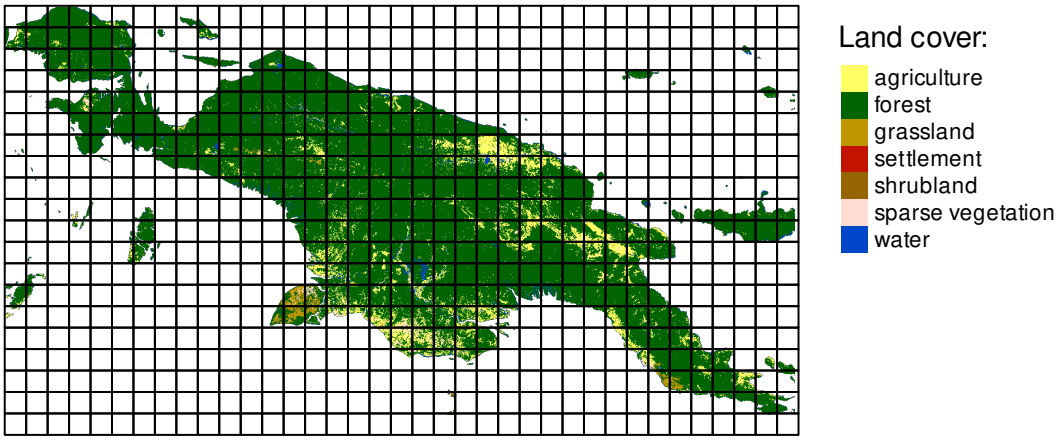

Figure 1: A land cover of New Guinea divided into a number of regular rectangular regions of 200 by 200 cells (60 by 60 km)

First, land cover composition (proportion of each of the land cover classes) was derived based on a set of regular non-overlapping regions using the `lsp_signature()` function (Figure 1). This function requires just a few arguments, including the input data, type of the calculated signature, and optionally, a value in the window argument. The default, `window = NULL`, calculates the spatial signature for the entire input dataset. In this example, we set `window` to 200, which means that each regular region will consist of 200 by 200 cells.

```
comp_output = lsp_signature(landcover,
                            type = "composition",
                            window = 200)
comp_output
```

```
## # A tibble: 300 x 3
##      id na_prop signature
## * <int>   <dbl> <list>
## 1     2  0.644  <dbl[,7] [1 x 7]>
## 2     3  0.0992 <dbl[,7] [1 x 7]>
## 3     4  0.145  <dbl[,7] [1 x 7]>
## 4     5  0.602  <dbl[,7] [1 x 7]>
## 5     6  0.775  <dbl[,7] [1 x 7]>
## # ... with 295 more rows
```

The output of the `lsp_signature()` function is an extended data frame of class `lsp`. It has three columns, `id` of each region, the proportion of cells with missing values in each region,

and a list-column containing calculated signatures. The list-column allows storing each type of spatial signatures from a single value signatures (or metrics) to multidimensional arrays.

Each signature can be accessed as a regular list object in R. For example, the spatial signature of the first region in this object is:

```
comp_output$signature[[1]]
```

```
##                  1         2 3 5 6 7          9
## [1,] 0.0009141411 0.9611138 0 0 0 0 0.03797201
```

It is a one-row matrix, where each column relates to the subsequent land cover category, and each value is a proportion (0-1) of each category. The `lsp_signature()` function makes it also possible to return the actual number of cells with each category with the `normalization` argument set to `"none"`.

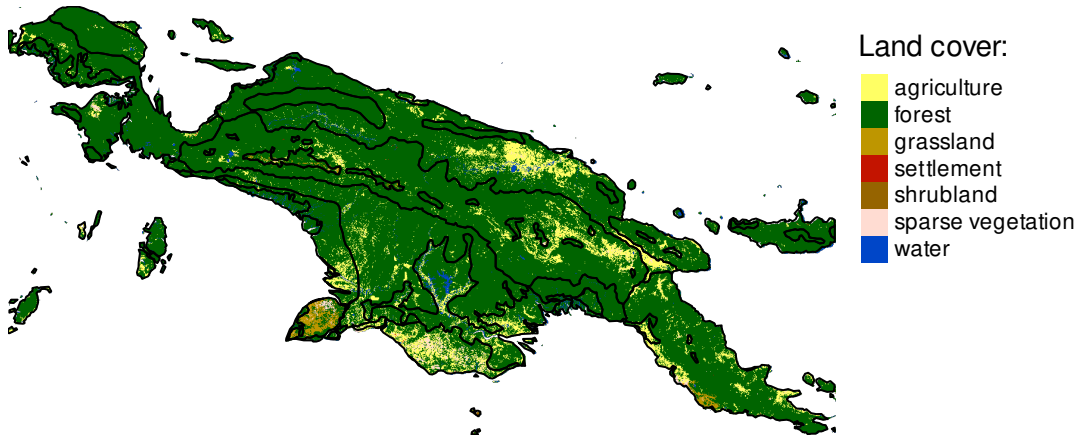

Figure 2: A land cover of New Guinea divided into 22 ecoregions

A second possible approach is to use irregular regions that can be provided as a spatial vector object of class `sf`.

```
ecoregions = read_sf("ecoregions.gpkg")
```

The `ecoregion` object contains 22 ecoregions for the New Guinea area derived from Dinerstein et al. (2017) (http://ecoregions2017.appspot.com/). Calculation of spatial signatures for irregular regions requires just using this object as a value in the `window` argument.

```
landcover_comp_e = lsp_signature(landcover, type = "composition",
                                 window = ecoregions["id"])
landcover_comp_e
```

The structure of the output is identical to the first example.

```
## # A tibble: 22 x 3
##      id na_prop signature
## * <int>   <dbl> <list>
## 1     1  0.114  <dbl[,7] [1 x 7]>
## 2     2  0.0851 <dbl[,7] [1 x 7]>
## 3     3  0.0377 <dbl[,7] [1 x 7]>
## 4     4  0.0914 <dbl[,7] [1 x 7]>
## 5     5  0      <dbl[,7] [1 x 7]>
## # ... with 17 more rows
```

4.2 Spatial pattern search

Spatial pattern search allows for quantifying similarity between the query region and the search space and finally finding regions that are the most similar to the query one. In this case study, we were interested in finding areas of similar topography to the area of Suwalski Landscape Park.

Suwalski Landscape Park is a protected area in north-eastern Poland with a post-glacial landscape consisting of young morainic hills. One possible approach to the raised question is to use a geomorphons map of this region. Geomorphons categorize cells in this area into one of ten landform types. Left panel at Figure 3 shows that this area has irregular spatial patterns with a significant part represented by slopes and only a limited number of flat areas. Spatial search requires two groups of input data, one representing the query region, and second that we want to search.

```
gm_suw = read_stars("gm_suw.tif")
gm_pol = read_stars("gm_pol.tif")
```

The `lsp_search()` function allows us to calculate distances between the query region and the search area. It accepts the query region and the search region and the type of signatures we want to compare. Next, an applicable distance measure needs to be specified. Here, we used the Jensen Shannon distance. Similarly to the first use case, it is also possible to search for either regular rectangular regions or irregular regions.

```
gm_search = lsp_search(gm_suw, gm_pol,
                       type = "cove", dist = "jensen-shannon",
                       window = 100)
```

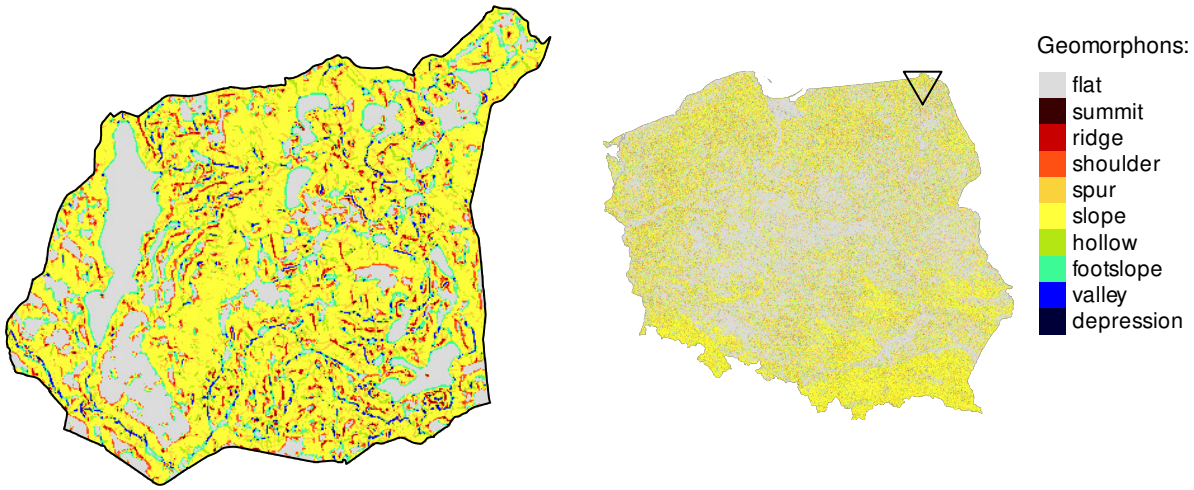

Figure 3: Geomorphons of: (left) Suwalski Landscape Park, (right) Poland with a triangle pointing to the location of the Suwalski Landscape Park

The output gives a distance value between the query region and every search region (Figure 4). The results show that majority of similar areas are located in northern Poland and forms a belt with homogeneous topography. The `lsp_extract()` function allows to pull out selected regions. Eight examples of the areas with the most similar patterns of geomorphons comparing to the Suwalski Landscape Park are presented in Figure 4. They are also similar to each other, suggesting a high-quality result.

4.3 Spatial pattern comparison

Spatial pattern comparison allows detecting changes in spatial patterns between two sets of data. Here, our goal is to compare how the land cover changed for the Amazon region between the years 1992 and 2018. For this purpose, two categorical raster datasets with the same resolution and extent need to be read (Figure 5).

```
lc92 = read_stars("lc_am_1992.tif")
lc18 = read_stars("lc_am_2018.tif")
```

Comparison of spatial patterns is a role for the `lsp_compare()` function. Its syntax is very similar to `lsp_search()`, however, it expects two datasets. Here, we used the co-occurrence vector as a signature, the Jensen-Shannon distance, and regular regions of 300 by 300 cells.

```
lc_am_compare = lsp_compare(lc92, lc18,
                      type = "cove", dist_fun = "jensen-shannon",
                      window = 300)
```

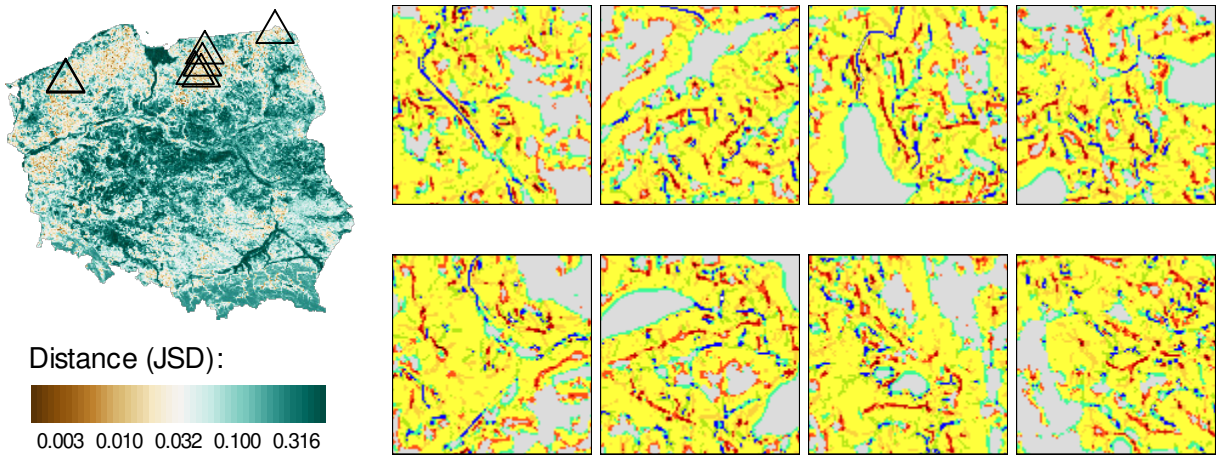

Figure 4: (Left) A result of the spatial pattern search between the Suwalski Landscape Park and the Poland area. Triangles represent eight areas with the most similar patterns of geomorphons comparing to the Suwalski Landscape Park (Right) Examples of the eight areas with the most similar patterns of geomorphons comparing to the Suwalski Landscape Park

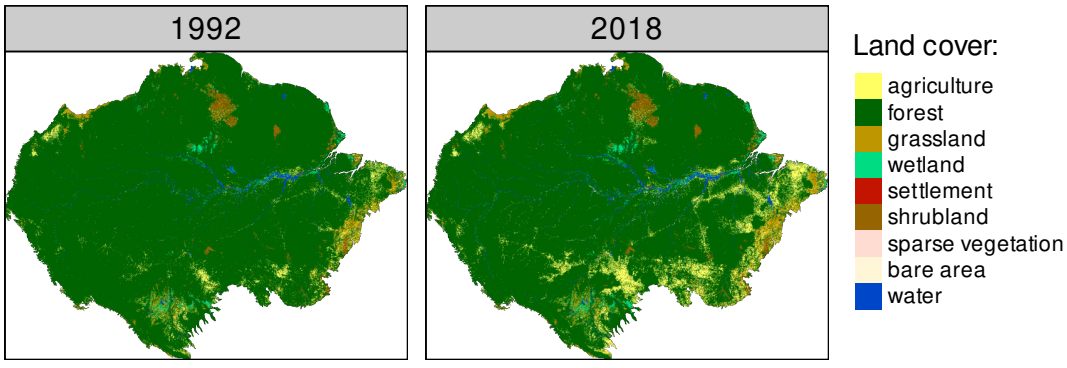

Figure 5: Land cover for the Amazon region for the years 1992 and 2018

The resulting object provides an overview of the land cover pattern changes for the Amazon region (Figure 6). Areas with the most intensive changes are located in the southern and eastern parts of the region. Figure 6 also shows seven areas with the largest land cover changes extracted with the `lsp_extract()` function. In all of the cases, the spatial pattern-change is similar. Regions covered mostly by forest in 1992 are now vastly replaced with extensive areas of agriculture.

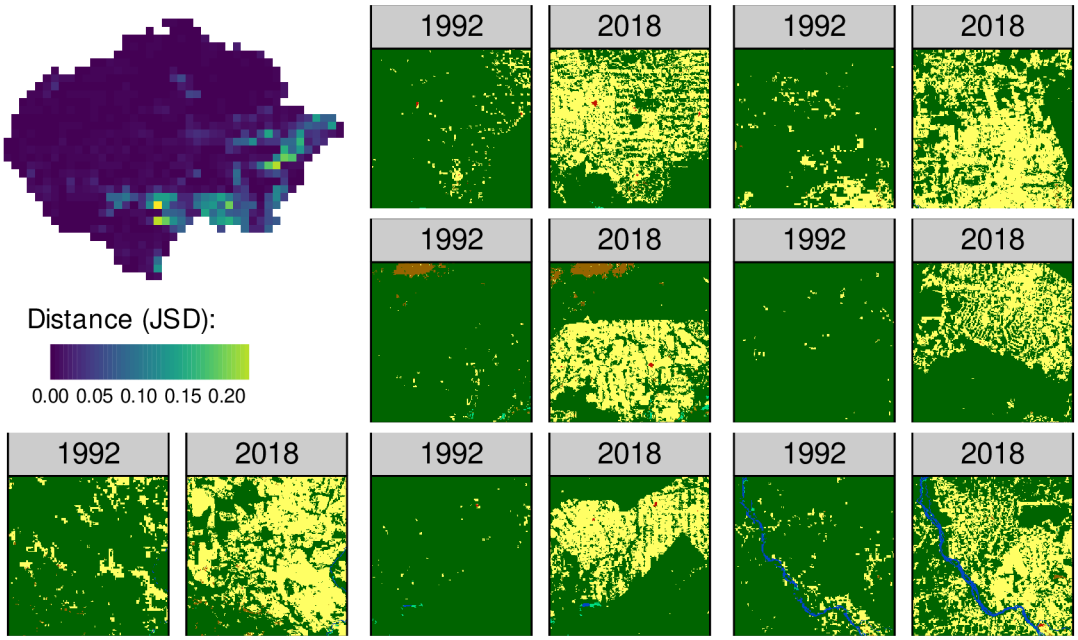

Figure 6: A visualization of the Amazon region presenting the change magnitude between the land cover patterns between 1992 and 2018 (top left), and the seven examples of areas with the largest changes of spatial land cover patterns between 1992 and 2018

4.4 Spatial pattern clustering

Spatial pattern clustering allows us to identify regions with similar spatial patterns and group them together. The previous case studies used the single layer datasets, either representing land cover categories or geomorphons. However, the pattern-based spatial analysis is not limited to just single layers. It also works on many layers, but in these cases, it requires using an appropriate spatial signature. Recently, a new signature, an integrated co-occurrence matrix, was developed to represent multilayer spatial patterns. This signature is implemented in the **motif** package in the form of a matrix (`incoma`) and vector (`incove`).

In this case study, we were interested in in finding clusters of regions with similar patterns of land covers and landforms in Africa. Two datasets representing land cover and landform regions of Africa were read into R, and combined into a single **stars** object (Figure 7).

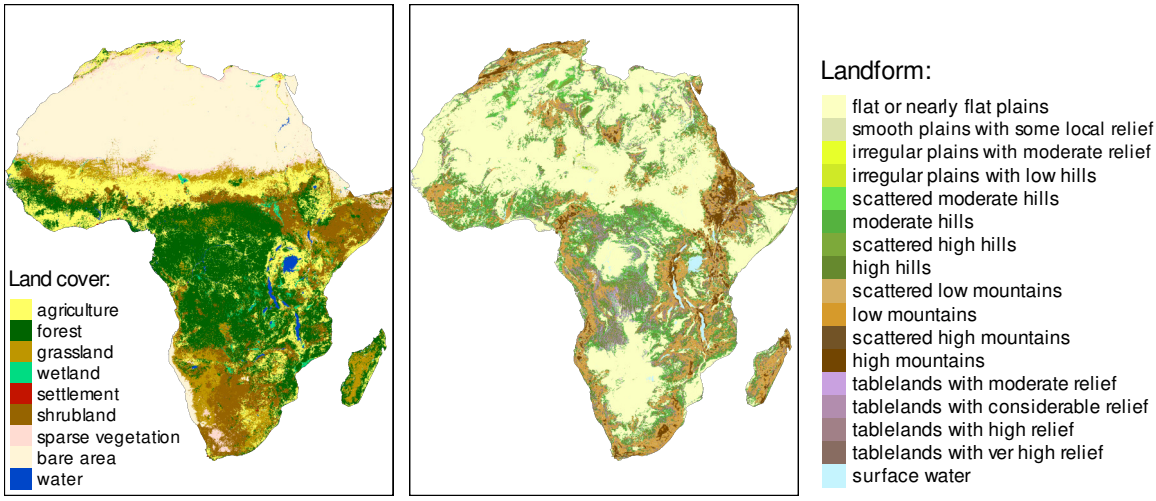

Figure 7: Land cover map (left) and landform regions map (right) of Africa

```
lc = read_stars("data/land_cover.tif")
lf = read_stars("data/landform.tif")
eco = c(lc, lf)
```

Next, spatial pattern clustering requires a few steps. The first one is to derive a spatial signature for all of the analyzed regions using `lsp_signature()`. Here, we are interested in the multi-layer spatial patterns of both land cover and landform regions, therefore we can use the `"incove"` signature. The second step uses `lsp_to_dist()` to calculate the distance between spatial signatures for each region. The result of the calculation is an object of class `dist` that can be used in any clustering technique requiring a distance matrix.

```
eco_signature = lsp_signature(eco_data,
                              type = "incove",
                              window = 300,
                              normalization = "pdf")
eco_dist = lsp_to_dist(eco_signature, dist_fun = "jensen-shannon")
```

In this example, we used the hierarchical clustering with the Ward agglomeration method, and based on the dendrogram plot, we decided to distinguish eight clusters. Cluster were added to the `lsp` object using the `lsp_add_clusters()` function and are presented on the left in Figure 8.

Figure 8 also contains an example for each of the clusters derived with the `lsp_extract()` function. The first cluster represents areas with diverse land cover (with dominating agriculture) on complex terrain. Contrary to the first one, the second cluster is mostly located in one region (Sahel), with agriculture, grassland, sparse vegetation, and bare areas on plains. Most areas from the third cluster form a belt south from the second cluster. It consists of areas with dominating agriculture on plains or hills. The fourth and fifth

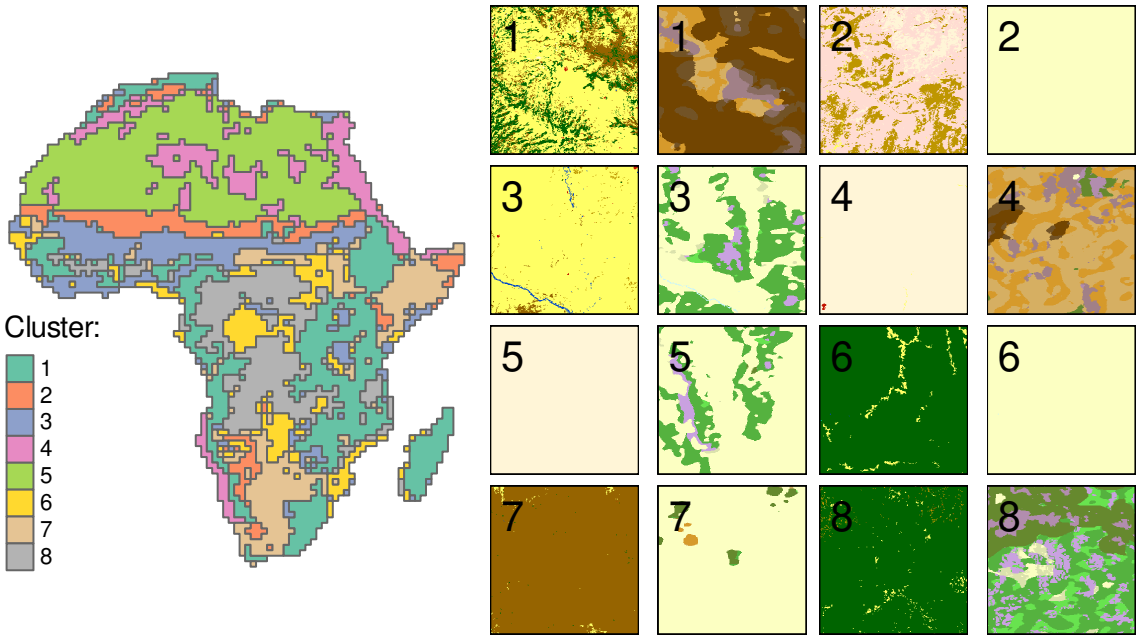

Figure 8: Africa divided into eight clusters of similar spatial patterns of land cover and landform regions (left) and examples of each cluster (right). Legends for the colors in the examples are presented in Figure 7

clusters are closely related to the parts of the Sahara desert, with the fourth cluster located on mountains and the fifth one on plains or hills. The sixth and eights clusters also have similar land cover, both covered mostly by forests with smaller areas of agriculture and grasslands. They differ in terms of their landforms, with the sixth cluster located mostly on plains, and eighth on complex terrain. Finally, the seventh cluster relates to areas with the domination of shrublands.

The **motif** package also makes possible to evaluate properties of each cluster with the `lsp_add_quality()` function. It returned three values to each cluster indicating their inhomogeneity, isolation, and quality (Table 3). All clusters showed similar levels of isolation, but differed in terms of inhomogeneity, and therefore overall quality. The best overall quality was found for the fifth cluster, which had the most homogeneous spatial patterns (lowest inhomogeneity). On the other hand, the first cluster showed the lowest values of overall quality having the most inhomogeneous spatial patterns of land cover and landform regions.

5 Discussion

In this paper, we introduced the **motif** R package for pattern-based spatial analysis. It allows for the extraction of single- and multi-layer spatial patterns, pattern-based search, comparison, and clustering. The software is fully operational, and its functionality has been tested on global datasets with several billions of cells. This software was designed to work

Table 3: Quality metrics for eight clusters divided based on similar spatial patterns of land cover and landform regions for Africa

Cluster	Inhomogeneity	Isolation	Quality
1	0.48	0.75	0.37
2	0.40	0.70	0.42
3	0.34	0.69	0.51
4	0.29	0.73	0.60
5	0.12	0.70	0.83
6	0.28	0.61	0.54
7	0.37	0.69	0.46
8	0.29	0.68	0.57

on any modern laptop and does not require any external servers to run its calculations.

The **motif** package is based on the R language due to several reasons. R proved to be a main or one of the most important programs in many fields of science that rely on spatial data and spatial data patterns. It includes different subfields of ecology (Sciaini et al. 2018; Lai et al. 2019), spatial statistics, and GIScience (Lovelace, Nowosad, and Muenchow 2019). A vast number of existing R packages allow using the **motif** package not only as an individual tool but also as a part of many possible workflows. It is built upon popular R spatial frameworks, such as **stars** and **sf**, and therefore each input to this package can be beforehand processed in R, and also all of the calculation outputs could be further used and modified using existing tools. Keeping most or all calculations inside of one toolbox increases replicability and reproducibility, and makes it a key element to validate scientific studies (Lovelace, Nowosad, and Muenchow 2019).

The existing software GeoPAT and GeoPAT 2 (Jasiewicz, Netzel, and Stepinski 2015; Netzel et al. 2018) already proved to be useful for the content-based search of Earth observation data archives (Peng et al. 2019), assessment of land cover change (Netzel and Stepinski 2015), or mapping geomorphological landscapes (Józsa and Fábián 2016). Their functionality, however, is limited to working on just one layer (a type of spatial data) at the time. The **motif** package provides most of the functionality of the existing software and opens new potential applications for pattern-based analyses of many layers.

Future improvements of the software will be aimed in several directions. The R language lacks a robust implementation of spatial segmentation. It would be worth not only to implement some spatial segmentation techniques using for the R language but also to investigate if the existing algorithms for spatial segmentation can be applied for multilayer data. The methods should be written efficiently to work on large spatial datasets, and thoughtfully tested and evaluated before adding them to the **motif** package. Second, new signatures can be developed to provide new, useful measures of spatial patterns. It especially includes signatures aimed at describing many spatial scales at the same time. Third, there is a need to develop a comprehensive approach to analyze how different selected spatial scales influence obtained results and to provide an evidence-based way to decide

on a selected spatial scale. Fourth, possible performance improvements, especially when working on irregular regions and obtaining quality of clusterings should be investigated. Fifth, a large number of different spatial signatures and distance measures exist, however, we are not aware of any study comparing their effectiveness, advantages, and disadvantages when applying to different types of spatial patterns. The design of this package allows us to quickly tests many combinations of possible parameters, which then can be used as a basis of the comparison study. Finally, future users' inputs and experiences will be appreciated, as they can shed light on missing features or further areas of development.

References

Boots, Barry. 2003. "Developing Local Measures of Spatial Association for Categorical Data." *Journal of Geographical Systems* 5 (2): 139–60. https://doi.org/10/fjm8dk.

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.

Dinerstein, Eric, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, et al. 2017. "An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm." *BioScience* 67 (6): 534–45. https://doi.org/10.1093/biosci/bix014.

Dmowska, Anna, Tomasz F Stepinski, and Jakub Nowosad. 2020. "Racial Landscapes - a Pattern-Based, Zoneless Method for Analysis and Visualization of Racial Topography." *Applied Geography*.

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.

Eddelbuettel, Dirk, and Romain François. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40 (8): 1–18. https://doi.org/10.18637/jss.v040.i08.

Eddelbuettel, Dirk, and Conrad Sanderson. 2014. "RcppArmadillo: Accelerating R with High-Performance C++ Linear Algebra." *Computational Statistics and Data Analysis* 71: 1054–63. http://dx.doi.org/10.1016/j.csda.2013.02.005.

European Centre for Medium-Range Weather Forecasts. 2019. "ICDR Land Cover Product User Guide and Specification."

European Environment Agency. 2016. "European Digital Elevation Model (EU-DEM), Version 1.1."

European Space Agency. 2017. "Land Cover CCI Product User Guide Version 2." ESA Libin, Belgium.

Gustafson, Eric J. 1998. "Quantifying Landscape Spatial Pattern: What Is the State of the Art?" *Ecosystems* 1: 143–56.

Haralick, Robert M, Karthikeyan Shanmugam, and Its' Hak Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on Systems, Man, and Cybernetics*, no. 6: 610–21. https://doi.org/10/bdqvtn.

Haralick, Robert M, and Linda G Shapiro. 1985. "Image Segmentation Techniques." *Computer Vision, Graphics, and Image Processing* 29 (1): 100–132.

He, Hong S, Barry E DeZonia, and David J Mladenoff. 2000. "An Aggregation Index (AI) to Quantify Spatial Patterns of Landscapes." *Landscape Ecology* 15: 591–601.

Hesselbarth, Maximilian H. K., Marco Sciaini, Kimberly A. With, Kerstin Wiegand, and Jakub Nowosad. 2019. "*Landscapemetrics* : An Open-Source *R* Tool to Calculate Landscape Metrics." *Ecography* 42 (10): 1648–57. https://doi.org/10.1111/ecog.04617.

Jaeger, Jochen A G. 2000. "Landscape Division, Splitting Index, and Effective Mesh Size: New Measures of Landscape Fragmentation." *Landscape Ecology* 15: 115–30.

Jasiewicz, Jarosław, Paweł Netzel, and Tomasz Stepinski. 2015. "GeoPAT: A Toolbox for Pattern-Based Information Retrieval from Large Geospatial Databases." *Computers & Geosciences* 80: 62–73. https://doi.org/10.1016/j.cageo.2015.04.002.

Jasiewicz, Jarosław, and Tomasz F. Stepinski. 2013. "Geomorphons a Pattern Recognition Approach to Classification and Mapping of Landforms." *Geomorphology* 182 (January): 147–56. https://doi.org/10.1016/j.geomorph.2012.11.005.

Jasiewicz, Jarosław, Tomasz Stepinski, and Jacek Niesterowicz. 2018. "Multi-Scale Segmentation Algorithm for Pattern-Based Partitioning of Large Categorical Rasters." *Computers & Geosciences* 118 (September): 122–30. https://doi.org/10.1016/j.cageo.2018.06.003.

Józsa, Edina, and Szabolcs Ákos Fábián. 2016. "Mapping Landforms and Geomorphological Landscapes of Hungary Using GIS Techniques," 13.

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): 1040–60. https://doi.org/10.1111/tgis.12265.

Kaufman, Leonard, and Peter J Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Lai, Jiangshan, Christopher J. Lortie, Robert A. Muenchen, Jian Yang, and Keping Ma. 2019. "Evaluating the Popularity of R in Ecology." *Ecosphere* 10 (1). https://doi.org/10.1002/ecs2.2567.

Li, Habin, and James F. Reynolds. 1993. "A New Contagion Index to Quantify Spatial Patterns of Landscapes." *Landscape Ecology* 8 (3): 155–62. https://doi.org/10.1007/BF00125347.

Lin, Jianhua. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37 (1): 145–51.

Lovelace, R, J Nowosad, and J Muenchow. 2019. *Geocomputation with R*. Chapman and Hall/CRC Press.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2019. *Cluster: Cluster Analysis Basics and Extensions*.

McGarigal, Kevin. 2014. "Landscape Pattern Metrics." *Wiley StatsRef: Statistics Reference Online*, 13.

McGarigal, Kevin, Sam A Cushman, Maile C Neel, and Eduard Ene. 2002. "FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps."

Müller, Kirill, and Hadley Wickham. 2020. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Netzel, Pawel, Jakub Nowosad, Jaroslaw Jasiewicz, Jacek Niesterowicz, and Tomasz F Stepinski. 2018. "GeoPAT 2: User's Manual."

Netzel, Pawel, and Tomasz F. Stepinski. 2015. "Pattern-Based Assessment of Land Cover Change on Continental Scale with Application to NLCD 2001." *IEEE Transactions on Geoscience and Remote Sensing* 53 (4): 1773–81. https://doi.org/10.1109/TGRS.2014.2348715.

Nowosad, Jakub. 2020. *Comat: Co-Occurrence Matrices of Spatial Data*. https://nowosad.github.io/comat/.

Nowosad, Jakub, and Tomasz F. Stepinski. 2019. "Information Theory as a Consistent Framework for Quantification and Classification of Landscape Patterns." *Landscape Ecology* 34 (9): 2091–2101. https://doi.org/10.1007/s10980-019-00830-x.

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): 153–62. https://doi.org/10.1007/BF00162741.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

———. 2020. *Stars: Spatiotemporal Arrays, Raster and Vector Data Cubes*.

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 Observation and Geoinformation* 81 (September): 13–26. https://doi.org/10.1016/j.jag.2019.05.006.

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

Remmel, T. K. 2009. "Investigating Global and Local Categorical Map Configuration Comparisons Based on Coincidence Matrices: Investigating Global and Local Categorical Map Configuration." *Geographical Analysis* 41 (2): 144–57. https://doi.org/10/bvbjrv.

Riitters, Kurt. 2019. "Pattern Metrics for a Transdisciplinary Landscape Ecology." *Landscape Ecology* 34 (9): 2057–63. https://doi.org/10.1007/s10980-018-0755-4.

Sciaini, Marco, Matthias Fritsch, Cédric Scherer, and Craig Eric Simpkins. 2018. "NLMR and Landscapetools: An Integrated Environment for Simulating and Modifying Neutral

Landscape Models in R." Edited by Nick Golding. *Methods in Ecology and Evolution* 9 (11): 2240–8. https://doi.org/10.1111/2041-210X.13076.

Tennekes, Martijn. 2018. "tmap: Thematic Maps in R." *Journal of Statistical Software* 84 (6): 1–39. https://doi.org/10.18637/jss.v084.i06.

Turner, Monica G, and Robert H Gardner. 1991. *Quantitative Methods in Landscape Ecologythe Analysis and Interpretation of Landscape Heterogeneity*. 574.5 T8.

Vadivel, A., Shamik Sural, and A. K. Majumdar. 2007. "An Integrated Color and Intensity Co-Occurrence Matrix." *Pattern Recognition Letters* 28 (8): 974–83. https://doi.org/10.1016/j.patrec.2007.01.004.