rassta: Raster-Based Spatial Stratification Algorithms

by Bryan A. Fuentes, Minerva J. Dorantes, and John R. Tipton

Abstract Spatial stratification of landscapes allows for the development of efficient sampling surveys, the inclusion of domain knowledge in data-driven modeling frameworks, and the production of information relating the spatial variability of response phenomena to that of landscape processes. This work presents the **rassta** package as a collection of algorithms dedicated to the spatial stratification of landscapes, the calculation of landscape correspondence metrics across geographic space, and the application of these metrics for spatial sampling and modeling of environmental phenomena. The theoretical background of **rassta** is presented through references to several studies which have benefited from landscape stratification routines. The functionality of **rassta** is presented through code examples which are complemented with the geographic visualization of their outputs.

1 Introduction

The application of robust, quantitative approaches for the spatial modeling of environmental phenomena has increased in the past few decades mainly due to an increase in computational power, advances in statistical modeling, and the availability of geospatial layers of environmental information (Scull et al. 2003; Elith and Leathwick 2009). Most of these approaches aim at building explicit quantitative relationships between environmental controls and response phenomena through statistical learning. Examples of these approaches include digital soil mapping (DSM) (McBratney, Mendonça Santos, and Minasny 2003), species distribution modeling (SDM) (Guisan and Zimmermann 2000), land use/land cover classification (Ham et al. 2005), and forest fire modeling (Chuvieco et al. 2010). Despite the extensively documented success of these approaches, there are still some challenges that limit their application. For instance, poor statistical performance is often reported in studies where input data is too limited to accurately represent control-response relationships (Araújo and Guisan 2006). Moreover, model parsimony and interpretation of results can be compromised when using 'black-box' algorithms (Arrouays et al. 2020). Similarly, including *a priori* knowledge about natural processes in purely statistical approaches can be challenging to achieve (Heuvelink and Webster 2001).

Several studies have suggested embedding spatial stratification routines within approaches such as DSM, SDM, land use/cover mapping, forest fire modeling, and others to overcome the challenges limiting their application. In such studies, the spatial stratification of landscapes creates units with reduced spatial variability of environmental phenomena as compared to the overall variability across a landscape. The use of these units allows the researcher to (a) obtain balanced representations of control-response relationships (Guisan and Zimmermann 2000; West et al. 2016); (b) include expert knowledge of physical processes for improving modeling with limited data (Zhu et al. 2008); (c) improve the performance of parameterization of mechanistic models (Park and Van De Giesen 2004; Baldwin, Naithani, and Lin 2017); and, (d) facilitate the interpretation of environmental conditions and their influence on the spatiotemporal variability of processes of interest (Rodrigues et al. 2019).

In general, landscape stratification routines follow fundamental ecological concepts that explain the hierarchical and multi-scale nature of relationships between environmental phenomena across space (Allen and Starr 1982). Therefore, landscape stratification methods have been applied in many studies that use geospatial information for environmental modeling, such as those previously cited. However, few packages exist in the R environment with functions strictly aimed at landscape stratification routines using geospatial data. Although one could implement custom stratification algorithms using multiple all-purpose geospatial analysis packages such as **terra** (Hijmans 2021) and **sf** (Pebesma 2018), the ease of use, reproducibility, and replicability of analysis is often enhanced when algorithms are implemented as part of a dedicated package. The **motif** package (Nowosad 2021) is the only example the authors could find of a package that is fully dedicated to landscape stratification in R using geospatial data. Although the methods offered by **motif** are effective for large-scale studies (Jasiewicz, Netzel, and Stepinski 2015; Nowosad 2021), their application is currently limited to rasters of categorical data. Thus, **motif** is not practical for the modeling of spatially continuous environmental phenomena, which is often a goal of landscape stratification routines.

This work presents the **rassta** package as a collection of algorithms for the spatial stratification of landscapes, sampling, and modeling of environmental phenomena. The **rassta** package is not intended as a drop-in replacement for statistically-robust environmental modeling approaches. Rather, it is intended to serve as a generalized framework to derive geospatial information that can be used to improve inference with these statistical approaches.

2 Conceptual overview and functionality

The algorithms in the **rassta** package assist in the analysis of environmental information related to the spatial variability of natural phenomena across landscapes. These functions focus on integrating standard geospatial techniques and quantitative analysis in a generalized framework for landscape stratification, sampling, and modeling. All of the functions in the **rassta** package take geospatial data in raster format as input. In the context of geographic information systems (GIS), the raster format can be considered a graphical representation of a matrix that is organized in rows and columns, and which may be stacked in multiple layers (e.g., multi-band satellite imagery). Each cell (pixel) in the raster contains a value representing a spatially-varying phenomenon, such as elevation or precipitation. A few functions in **rassta** also produce geospatial data in vector format. Vector data represents geometric entities in the form of points, lines, and polygons. The **rassta** package uses the highly efficient **terra** package as the backbone for handling raster and vector data. Most of the geospatial data manipulation with **terra** is performed in C++ and is based on two main R data types (classes): `SpatRaster` and `SpatVector`. Note that **terra** imports the Rcpp package (Eddelbuettel and François 2011) since **terra** uses C++ (including external pointers) to manipulate these classes.

Most of the functions implemented in **rassta** are interrelated in the sense that the outputs from some functions can be used as the inputs for others. This functional interrelation allows for a generalized framework to conduct spatial stratification, sampling, and modeling in a single package following a project-oriented approach. In general, the functions of **rassta** can be grouped into five categories: (a) landscape stratification; (b) landscape correspondence metrics; (c) stratified sampling; (d) spatial modeling; and (e) miscellaneous (Figure 1). Each category and its corresponding functions (except for miscellaneous) are theoretically founded on several studies focused on understanding spatially-varying natural phenomena across landscapes. In the next sections, the rationale behind each category and its functions is described. This description is complemented with references to corresponding scientific literature and includes code examples showing the application of each function with extensive use of plotting functions (for visualization purposes only). Most of the plotting functions are derived from the **terra** package using the `SpatRaster` and `SpatVector` classes. [Note: To reduce the extension of code examples, all the map and graph plotting functions were consolidated in the function `figure()`].

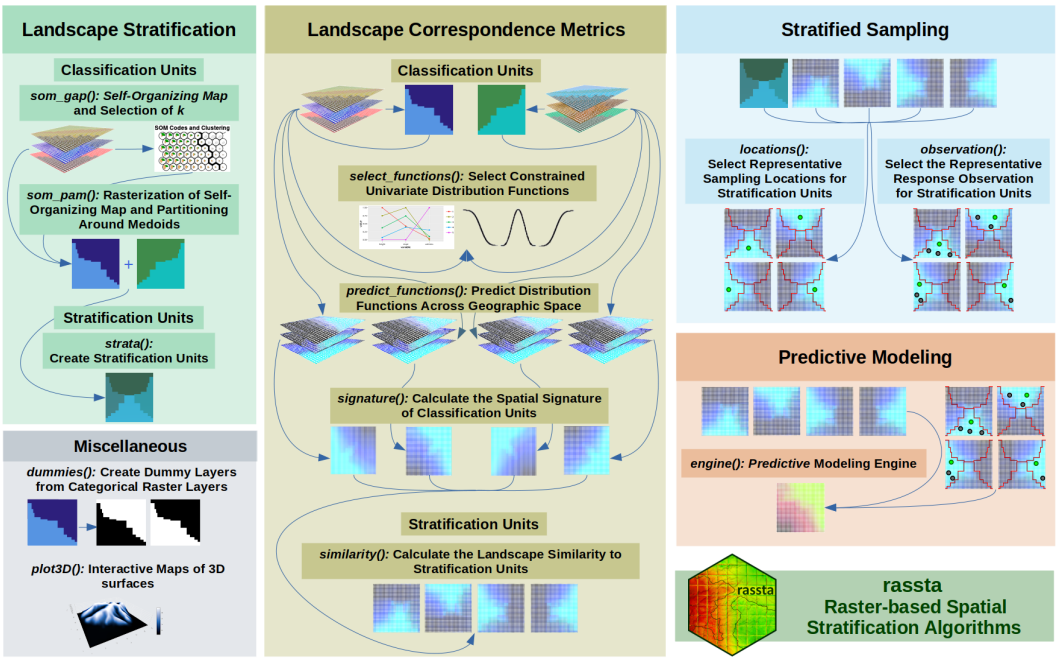

Figure 1: Functions of the rassta package. Connectors relate the inputs and outputs of the functions. The functions can be grouped in five categories: landscape stratification, landscape correspondence metrics, stratified sampling, predictive modeling, and miscellaneous.

Landscape stratification

Several studies have suggested the need to account for the hierarchical and multi-scale nature of landscape processes. Allen and Starr (1982) suggested that landscape processes can be explained through hierarchical multivariate structures given their multiple spatial and temporal scales. Based on

Dokuchaeiv's theory of soil formation (Glinka 1927) and the soil-landscape paradigm (Hudson 1992), McSweeney et al. (1994) proposed a nested model of soil-landscape processes at the physiographic, geomorphometric, and within-soil domains. Flügel (1995) suggested that the regionalization of hydrology-related processes should consider the multi-scale landscape heterogeneity in terms of soil, topography, geology, climate, and vegetation. These ideas have led these and other authors to formulate frameworks for the creation of spatial entities that stratify the landscape. The general purpose of these entities has been to define spatially-explicit domains that represent distinctive landscape processes and/or interactions (McSweeney et al. 1994). Accordingly, spatial stratification using **rassta** focuses on the creation of such domains (hereafter referred to as units).

The landscape stratification process with **rassta** follows a hierarchical approach similar to Austin and Heyligers (1989), who individually classified gradients of precipitation and elevation into intervals that were intersected with geologic classes for sampling purposes. Similarly, in **rassta**, a set of first-level units is created separately for each landscape factor under analysis. Then, multiple sets of first-level units are integrated into a single set of second-level units. The first-level units, called *classification units*, can be created outside of **rassta** via multicriteria analysis, statistical learning, or other methods. Moreover, the classification units can be formally defined through classification schemes, such as those based on taxonomic keys. The second-level units, called *stratification units*, result from the spatial intersection of multiple sets of classification units. Note that both classification and stratification units represent a spatial stratification for a given landscape. Figure 2 shows an example of a simple landscape stratification process based on two landscape factors, each with three raster layers representing continuous variables.

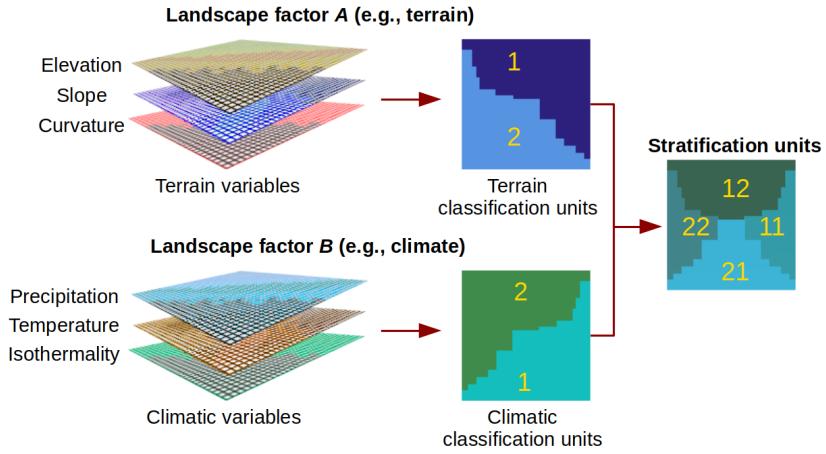

Figure 2: Schematic of a landscape stratification process. Raster layers of variables related to landscape factors are the inputs. The outputs are raster layers representing classification and stratification units.

There are three important aspects of the stratification approach used within **rassta** that must be considered. (a) One can simply create stratification units by incorporating variables from multiple landscape factors in a single classification process. However, the interpretation of results is often compromised when using a large number of variables in "all-in-the-bag" statistically driven classification schemes. (b) Multiple sets of classification units can belong to a single landscape factor, and each set can be created from variables at a distinct spatial scale. Presumably, this can account for the multi-scale nature of landscape factors in the stratification process. (c) A landscape factor can be represented by a single categorical variable, as in the case of geologic units or soil parent material. In this case, the landscape factor/variable is already in the form of classification units. Figure 3 shows a landscape stratification scenario like that addressed in (b) and (c).

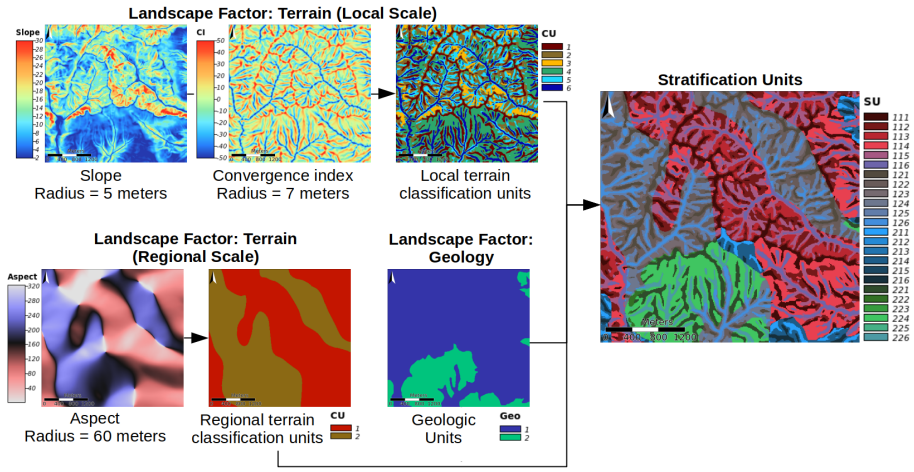

Figure 3: Schematic of a multi-scale landscape stratification process including a categorical variable. The stratification is based on three landscape factors: local scale terrain, regional scale terrain, and geology. Each terrain landscape factor is represented by raster layers of variables (local scale: slope and convergence index and regional scale: aspect and regional terrain). Geology is represented by a single categorical raster layer. Three sets of classification units (CU), one each for local terrain, regional terrain, and geology, are intersected to produce one set of stratification units (SU).

Classification units

A set of n classification units represents n distinct landscape configurations related to a single landscape factor. Note that the term *landscape configuration* is used here as a generic term for a particular pattern in the spatial variability of one or multiple variables belonging to a landscape factor. Currently, **rassta** allows the creation of classification units via unsupervised learning thanks to its functions `som_gap()` and `som_pam()`. The function `som_gap()` performs dimension reduction based on the self-organizing map (SOM) proposed by Kohonen (1990). The R package **kohonen** (Wehrens and Kruisselbrink 2018) is called internally by `som_gap()` to produce the SOM. The function `som_gap()` also performs cluster analysis on the SOM codes based on the partitioning around medoids (PAM) (Kaufman and Rousseeuw 1990), with estimation of the optimum number of clusters (k) through the gap statistic (Tibshirani, Walther, and Hastie 2001). It is important to mention that the output SOM object returned by `som_gap()` can be used as input for any other clustering algorithm (e.g., hierarchical, spectral, etc.) or statistical analysis outside of **rassta**.

The code below shows how `som_gap()` reduces the feature space and selects k clusters from four terrain variables. Note that the processing time of `som_gap()` is significant (around 162 seconds on a 4-cores Intel processor at 3.2 GHz for the following example). The processing time increases as the number of cells/layers in the argument `var.rast` increases, and/or as the argument `K.max` increases.

```
# Load the rassta and terra packages
library(rassta)
library(terra)
# Note that terra imports Rcpp, but if Rcpp is not automatically loaded then:
library(Rcpp)
# Get the data required to run the examples from rassta's installation folder
wasoil <- system.file("exdat/wasoil.zip", package = "rassta")
# Copy data to current working directory and extract files
file.copy(from = wasoil, to = getwd())
unzip("wasoil.zip")

# Set seed
set.seed(963)
# Multi-layer SpatRaster with 4 terrain variables
terr.var <- rast(c("height.tif", "midslope.tif", "slope.tif", "wetness.tif"))
# Scale variables to mean = 0 and standard deviation = 1
terr.varscale <- scale(terr.var)
# Dimensionality reduction and estimation of optimum k (max k to evaluate: 12)
terr.som <- som_gap(terr.varscale, xdim = 10, ydim = 10, K.max = 12)
# Plot results
```

```
figure(4, d = list(terr.var, terr.som))
```

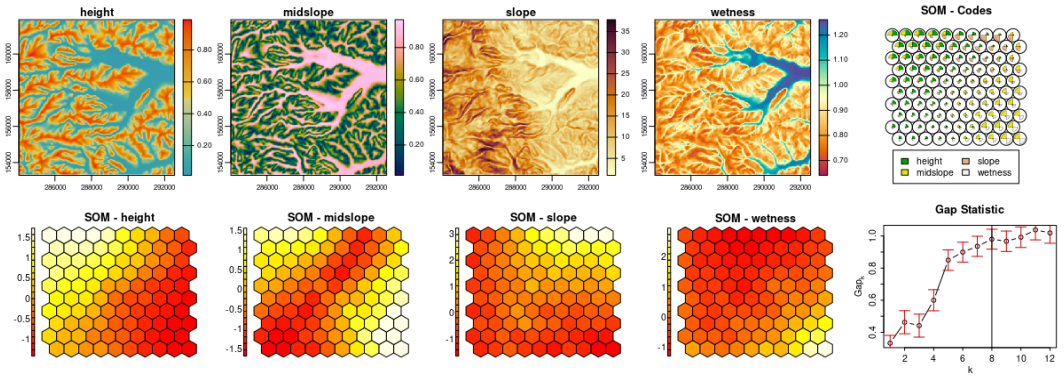

Figure 4: Dimension reduction and selection of number of clusters (k). The top row shows four terrain variables (height, midslope, slope, and wetness) that are used to generate the self-organizing map (SOM). The bottom row shows the reduced feature space of each variable and the Gap statistic that is used to select k for the construction of classification units.

The function `som_pam()` creates raster versions from the outputs of `som_gap()`. The code below shows how `som_pam()` creates raster versions of the SOM grid and PAM clustering computed in the previous example.

```
# Rasterization of terrain SOM grid and terrain PAM clustering
terr.sompam <- som_pam(ref.rast = terr.var[[1]], kohsom = terr.som$SOM,
                       k = terr.som$Kopt)
# Plot results
figure(5, d = list(terr.sompam, terr.var))
```

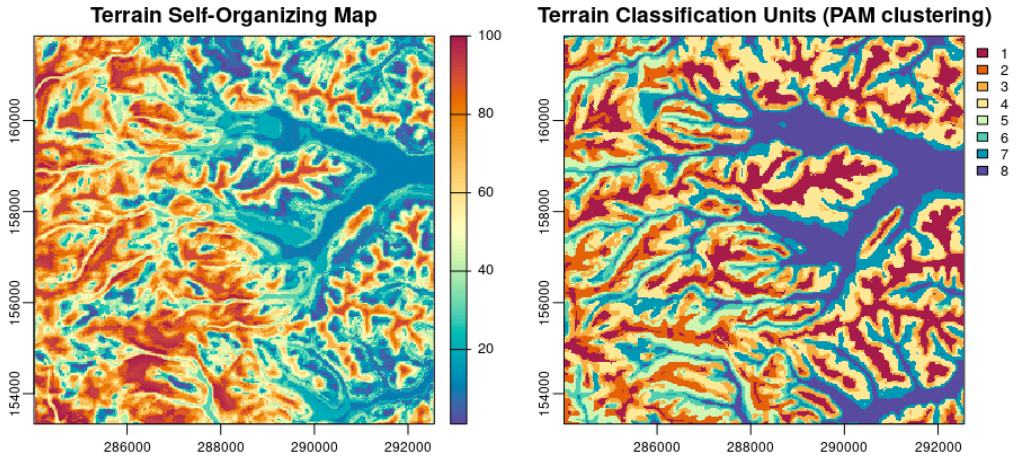

Figure 5: SOM grid and PAM clustering. Rasterized versions of the terrain SOM grid (left) and the terrain PAM clustering (right) are produced. The resulting clusters represent the classification units for the terrain landscape factor.

Note that the approach for creating classification units should not be limited to that offered by `som_gap()` and `som_pam()`. There are many other approaches outside of **rassta** that can be followed, such as supervised classification based on statistical learning, or GIS-based multicriteria analysis. The best approach may depend on the research question(s) being addressed. Therefore, the selection of the proper approach and the optional use of other R packages and/or GIS software is left to the user. Also, note that classification units created outside of **rassta** are completely compatible with **rassta** objects and methods if the units are represented through the `SpatRaster` class from **terra**.

Stratification units

A set of *n* stratification units represents *n* distinct landscape configurations related to multiple landscape factors. Note that the term *landscape configuration* is used here as a generic term for a particular pattern in the spatial variability of multiple variables belonging to multiple landscape factors, or to the same factor represented at multiple spatial scales. The function `strata()` allows the spatial intersection of multiple sets of classification units into a single set of stratification units. This function also assigns a unique numeric code to each stratification unit. The numeric code makes it possible to trace back each classification unit composing a given stratification unit. The code below shows the construction of stratification units with `strata()` using classification units from three landscape factors (climate, soil parent material, and terrain).

```
# Multi-layer SpatRaster with 3 sets of classification units
all.cu <- rast(c("climate.tif", "material.tif", "terrain.tif"))
# Stratification units
su <- strata(cu.rast = all.cu)
# Plot results
figure(6, d = list(su, all.cu))
```

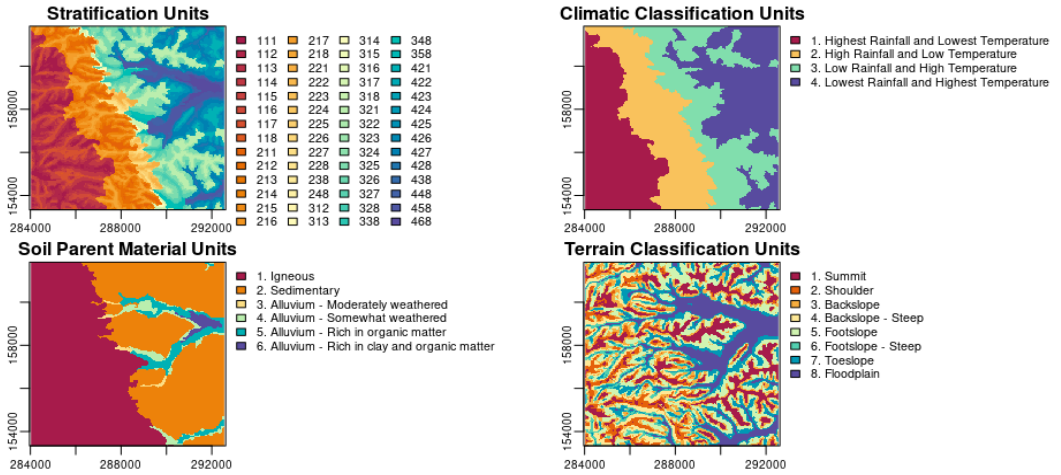

Figure 6: Creation of stratification units from sets of classification units. A set of classification units is produced for each of three landscape factors: climate, soil parent material, and terrain. The spatial intersection of these sets results in the stratification units for the landscape (upper left map).

Metrics of landscape correspondence

There are two metrics of landscape correspondence that can be calculated with **rassta**: (a) the spatial signature of classification units, and (b) the landscape similarity to stratification units. These metrics quantify the relative correspondence between any location across geographic space and landscape configurations represented by classification and stratification units. Several studies have applied similar concepts related to continuous correspondence between landscape configurations for the modeling of spatially-varying phenomena. Early examples include studies using multivariate distance metrics in the feature space for SDM (Carpenter, Gillison, and Winter 1993) and studies applying the fuzzy set theory (Zadeh 1965) for multicriteria evaluation (Burrough 1989), DSM (Zhu and Band 1994) and landform classification (MacMillan et al. 2000).

Spatial signature of classification units

The spatial patterns of the degree of correspondence between any landscape configuration and the configuration represented by a given classification unit are defined as the *spatial signature*. The spatial signature is represented by a raster layer of continuous values that results from the cell-wise aggregation of empirical distribution functions mapped over geographic space. Each distribution function corresponds to one variable and relates the classification unit to "typical" values of the variable within the classification unit. The concept of spatial signature is based on the work of Pike and Rozema (1975) and Pike (1988). These authors used the term *geometric signature* to describe a set of

sample statistics (e.g., mean, standard deviation) of terrain variables (e.g., slope, curvature) used to distinguish "geomorphically disparate landscapes" (Pike 1988).

The spatial signature in **rassta** replaces the geometric signature's measurements of central tendency and dispersion statistics with statistical distribution functions generated across geographic space. The statistical distribution functions used in **rassta** are: (a) the probability density function (PDF) based on the kernel density estimation, (b) the empirical cumulative distribution function (ECDF), and (c) an inverted version of the ECDF (iECDF). Note that the spatial signature concept is somewhat similar to the virtual ecological niche (Hirzel, Helfer, and Metral 2001) and the multivariate environmental similarity surface (Elith, Kearney, and Phillips 2010), which are implemented in R through the packages **virtualspecies** (Leroy et al. 2016) and **dismo** (Hijmans et al. 2020), respectively. Figure 7 and Figure 8 show an illustration and a pseudocode of the process to calculate the spatial signature of a classification unit, respectively. Note that the function `FUNSIG()` in the pseudocode is just a placeholder to encompass the three functions from **rassta** that are required to calculate spatial signatures. These functions are `select_functions()`, `predict_functions()`, and `signature()`, each will be further discussed next.

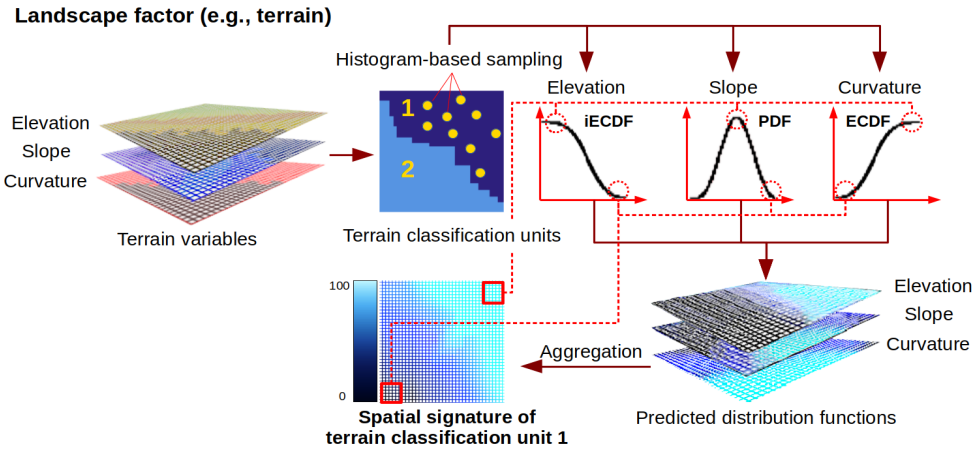

Figure 7: Schematic of the calculation process for spatial signatures. A set of classification units is produced using three variables. A distribution function is calculated for each variable within classification unit 1, and then predicted across geographic space. The predicted functions for unit 1 are aggregated, which results in the spatial signature of that unit.

Select and predict distribution functions, then calculate spatial signature

Require:
$V_{v1,\ldots,vn}$, multi-layer SpatRaster with n layers (i.e., variables)
$C_{c1,\ldots,cn}$, single-layer SpatRaster representing n classification units

function FUNSIG($V_{v1,\ldots,vn}, C_{c1,\ldots,cn}$)
  for each v in $V_{v1,\ldots,vn}$ **do**
    $S_{c1,\ldots,cn} \leftarrow$ SUMMARY($v, C_{c1,\ldots,cn}$) ▷ within-unit statistic (e.g., mean) of v
    for each c in $C_{c1,\ldots,cn}$ **do**
      $v_c \leftarrow$ EXTRACT(v, c) ▷ Extract observations of v within c
      if $S_c =$ MAX($S_{c1,\ldots,cn}$) **or** $S_c =$ MIN($S_{c1,\ldots,cn}$) **then** ▷ S_c: statistic of v within c
        $f \leftarrow ECDF$ ▷ select distribution function for v_c
        $v_{cs} \leftarrow$ QUANTILES(v_c) ▷ sample from v_c to reduce processing time
        $c_f \leftarrow f(v_{cs})$ ▷ calculate distribution function of v_{cs}
        if $S_c =$ MIN($S_{c1,\ldots,cn}$) **then**
          $c_f \leftarrow$ INVERT(c_f)
        end if
      else
        $f \leftarrow PDF$
        $v_{cs} \leftarrow$ CENTRALPOINTS(*histogram*(v_c))
        $c_f \leftarrow f(v_{cs})$
      end if
      $c_{loess} \leftarrow$ LOESS($y \sim x$); $y = c_f$, $x = v_{cs}$ ▷ fit *loess* regression
      $cv_{predfun} \leftarrow$ PREDICT(c_{loess}, v) ▷ predict the distribution function of v_{cs} across v
    end for
  end for
  for each c in C **do**
    $c_{predfun} \leftarrow cv1_{predfun}, ..., cvn_{predfun}$ ▷ Predicted distribution functions for c (one per each v)
    $c_{sig} \leftarrow$ AGGREGATE($c_{predfun}$) ▷ calculate Spatial signature of c
  end for
end function

Figure 8: Pseudocode of the calculation process for spatial signatures. The calculation process involves the selection, prediction, and aggregation of distribution functions. The spatial signature is calculated for each classification unit in a set.

An important assumption is made when using the PDF, ECDF, and iECDF to characterize the typical values of a given variable within a given classification unit. The position of a value within the distribution function is an indicator of how typical the value is in terms of the variable's distribution within the classification unit. For instance, values closer to, or at the peak of the PDF are assumed to be the most typical values of the variable within the classification unit. Contrarily, values at the tails of the PDF are the less typical. Although one could simply use the PDF as a generalized function to denote typical values, this function assigns the same weight to values at the tails of the distribution regardless of the tail's location (left or right). In some cases, *a priori* knowledge can dictate that typical values of a variable within a given classification unit are those approaching $+\infty$, or those approaching $-\infty$. The use of the ECDF and the iECDF is intended for those cases. More specifically, if a classification unit is known to be associated with a variable's extreme values toward $+\infty$, then the ECDF can be used to represent this association. Conversely, if the classification unit is associated with those variable's extreme values toward $-\infty$, then the iECDF can be used.

The function `select_functions()` allows the user to select the statistical distribution function used to represent the typical values for a given variable within a specific classification unit. Both automatic and interactive selection modes are supported, with the latter based on a **shiny** app (Chang et al. 2021). The automatic selection of distribution functions is based on within-unit statistics, also referred to as *zonal statistics* in the GIS literature, and it follows the criteria described next:

- PDF = when the mean (or median) of the variable's values within the classification unit is neither the maximum nor the minimum of all the mean (or median) values across all the units.
- ECDF = when the mean (or median) of the variable's values within the classification unit is the maximum of all the mean (or median) values across all the units.
- iECDF = when the mean (or median) of the variable's values within the classification unit is the minimum of all the mean (or median) values across all the units.

The code below shows the automatic selection of statistical distribution functions for four climatic classification units and two variables with `select_functions()`.

```
# Multi-layer SpatRaster with 2 climatic variables
clim.var <- rast(c("precipitation.tif", "temperature.tif"))
# Single-layer SpatRaster with 4 climatic classification units
```

```
clim.cu <- rast("climate.tif")
# Automatic selection of statistical distribution functions
clim.difun <- select_functions(cu.rast = clim.cu,
                               var.rast = clim.var,
                               mode = "auto")
# Plot results
figure(8, d = list(clim.difun, clim.cu, clim.var))
```

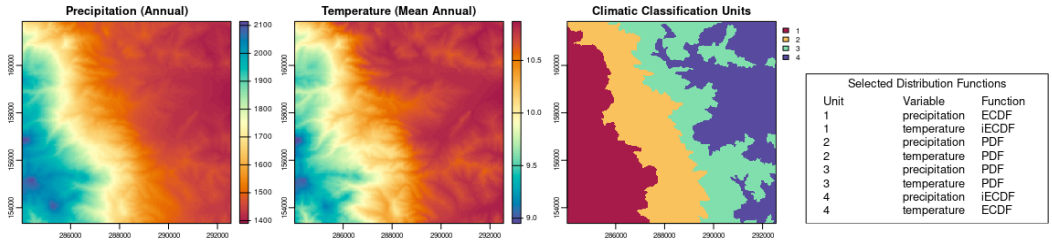

Figure 9: Selection of distribution functions. A set of four climatic classification units are produced using two variables: precipitation and temperature. A distribution function is selected for each variable within each classification unit.

The selected distribution functions can be used to generate predictions of distribution function values over geographic space with the function `predict_functions()` as shown in the code below. The predictions are generated by fitting a locally estimated scatterplot smoothing (LOESS) regression with the within-unit distribution function's values (y) and the within-unit variable's values (x). The fitted LOESS and the raster layer of the variable are then used to predict new distribution function values across geographic space.

```
# Multi-layer SpatRaster of climatic variables and classification units
clim.all <- c(clim.var, clim.cu)
# Ouput table from select_functions()
df <-  clim.difun$distfun
# Predicted distribution functions for climatic variables
clim.pdif <- predict_functions(cuvar.rast = clim.all,
                               cu.ind = 3,
                               cu = df$Class.Unit,
                               vars = df$Variable,
                               dif = df$Dist.Func)
# Plot results
figure(9, d = list(clim.pdif, clim.cu))
```

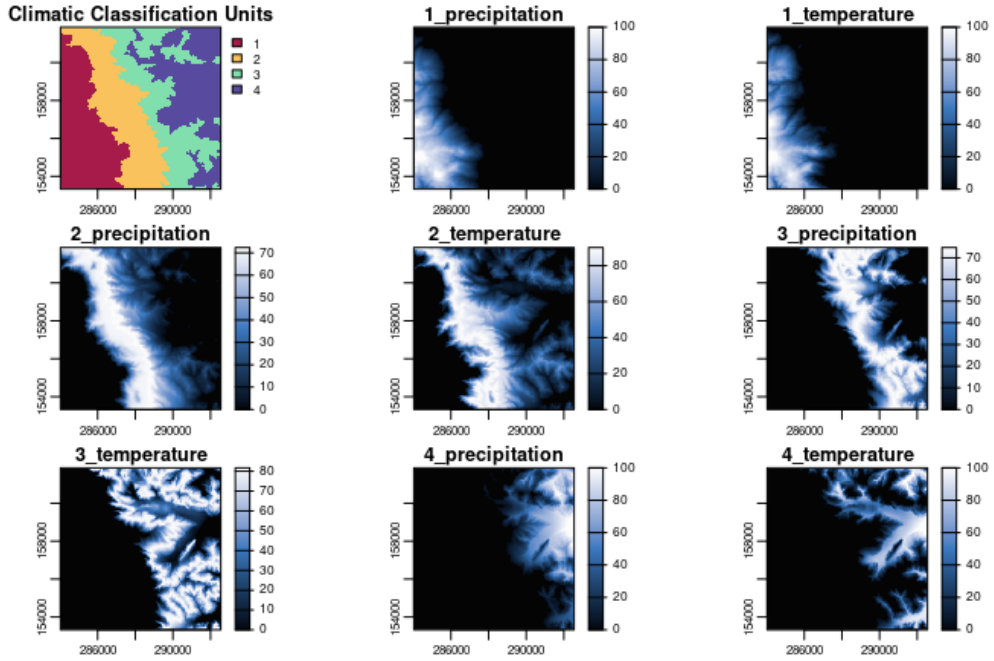

Figure 10: Prediction of distribution functions. A selected distribution function for each variable is predicted across geographic space. The predicted distribution function relates the landscape to a classification unit with regard to a variable.

The function `signature()` calculates the spatial signature of a given classification unit by aggregating all of the predicted distribution functions associated with the unit. The code below shows the calculation of spatial signatures with `signature()`. Note that the arguments `inprex` and `outname` allow the user to identify the raster layers representing the predicted distribution functions associated with each classification unit in a set, and to assign a unique name to each resulting raster layer of spatial signature, respectively.

```
# Spatial signatures from distribution functions predicted for climatic variables
clim.sig <- signature(pdif.rast = clim.pdif,
                      inprex = paste(seq(1, 4), "_", sep = ""),
                      outname = paste("climate_", seq(1, 4), sep = ""))
# Plot results
figure(10, d = list(clim.sig, clim.cu))
```

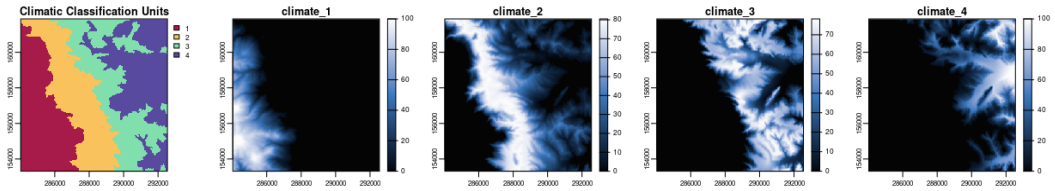

Figure 11: Calculation of spatial signatures. For each climatic classification unit (1 thru 4), the distribution functions (see Figure 10) are aggregated (e.g., mean pixel value) to produce the spatial signature of the unit. The spatial signature relates each position in the landscape to the landscape configuration represented by a classification unit.

Landscape similarity to stratification units

The spatial patterns of the degree of correspondence between any landscape configuration and the landscape configuration represented by a given stratification unit are defined as the *landscape similarity*. The landscape similarity is represented by a raster layer of continuous values, which results from the cell-wise aggregation of the spatial signatures of multiple classification units. This aggregation is possible because any given stratification unit is the result of the spatial intersection of multiple classification units, commonly one per landscape factor or factor scale (see Figure 2 and 3). Moreover,

each classification unit has one spatial signature associated with it. Therefore, any given stratification unit will be associated with multiple spatial signatures, which can be cell-wise aggregated to calculate the landscape similarity. Figure 12 shows an example of the calculation process for a layer of landscape similarity to stratification unit.

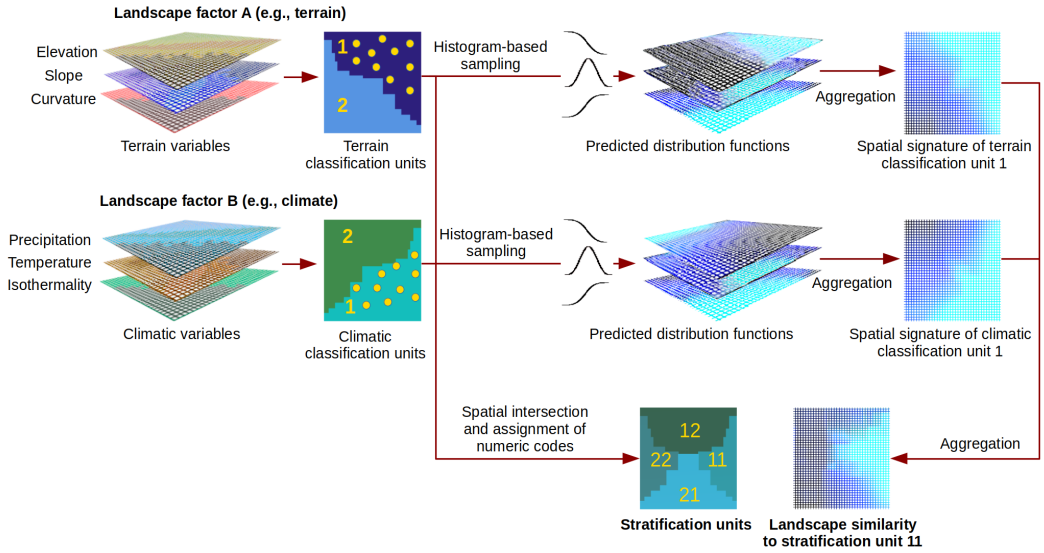

Figure 12: Schematic of the calculation process for landscape similarities. Sets of variables for each landscape factor (terrain and climate) are combined to produce sets of classification units (two each for terrain and climate), which are further combined to produce stratification units (12, 11, 21, and 22). Thus, each stratification unit has two classification units associated with it. Moreover, each classification unit has a spatial signature associated with it. Aggregating the spatial signatures of classification unit 1 for climate and unit 1 for terrain, both associated with stratification unit 11, results in the landscape similarity to that stratification unit.

The function `similarity()` calculates the landscape similarity layer for each stratification units in a given set (with the set being represented by a single-layer `SpatRaster` object), as shown in the following example. The argument `su.code` indicates the name of the landscape factors/factor scales used to create the stratification units, and the digit position (start, end) of the classification units' ID in the stratification unit's numeric code.

```
# Multi-layer SpatRaster with spatial signatures of classification units
clim.sig <- rast(list.files(pattern = "climate_")) # For climatic units
mat.sig <- rast(list.files(pattern = "material_")) # For soil parent material units
terr.sig <- rast(list.files(pattern = "terrain_")) # For terrain units
# Single-layer SpatRaster of stratification units
su <- rast("su.tif")
# Landscape similarity to stratification units
su.ls <- similarity(su.rast = su, sig.rast = c(clim.sig, mat.sig, terr.sig),
                    su.code = list(climate = c(1, 1), material = c(2, 2),
                                   terrain = c(3, 3)))
# Plot results
figure(12, d = list(su.ls, su, clim.sig, mat.sig, terr.sig))
```

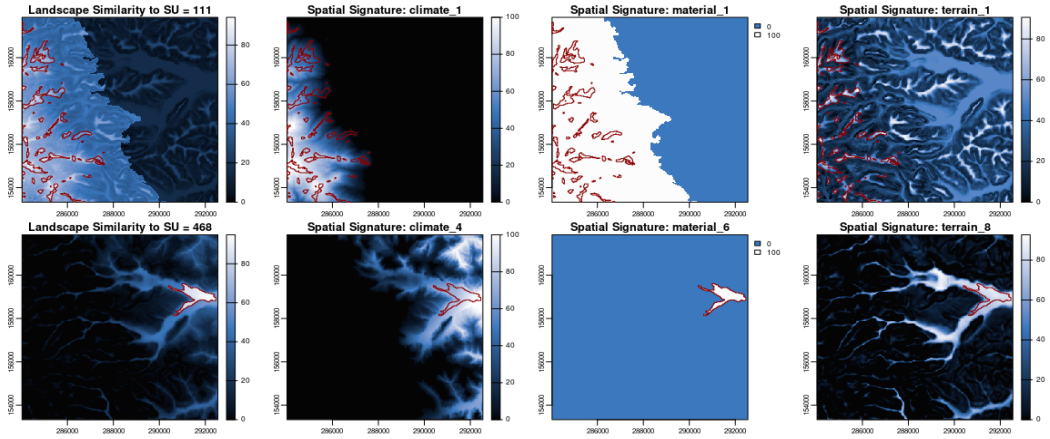

Figure 13: Metrics of landscape correspondence. Landscape similarity (extreme left) and spatial signatures for climate, parent material (material), and terrain associated with stratification units (SU) 111 (top row) and 468 (bottom row). The red polygons indicate the boundaries of the corresponding SU defined through the aggregation (i.e., mean pixel value) of the set of spatial signatures for that SU.

Stratified non-probability sampling

Stratified sampling is an efficient technique for achieving an adequate representation of environmental variability, reducing cost of field work, and improving modeling with limited observations (Austin and Heyligers 1989; Wessels et al. 1998; Guisan and Zimmermann 2000; Zhu et al. 2008; West et al. 2016). Accordingly, sampling with **rassta** to select observations/sampling locations is performed in a stratified fashion using stratification units. Additionally, the raster layers of landscape similarity to stratification units can be included in the sampling process. Including the landscape similarity layers results in a non-probability sample. For each stratification unit, the sampling process selects the observation(s)/sampling location(s) at the raster cell where the highest landscape similarity value occurs, resulting in a stratified, non-probability sample that is biased towards maximizing the representativeness of landscape configurations. This idea of biased, stratified sampling is based on the work of Gillison (1983); Gillison and Brewer (1985), Austin and Heyligers (1989), and Zhu et al. (2008). These authors have suggested that bias related to landscape configurations is relevant for the maximization of environmental representativeness, detection of maximum diversity, and representation of non-stochastic control-response relationships.

The function `observation()` performs the automatic selection of the representative response observation for each stratification unit in a given set. Given a stratification unit, the unit's representative response observation is that whose value best reflects the influence that the unit's landscape configuration exerts on the response. This function requires a set of observations/samples already collected for a set of stratification units. Currently, `observation()` selects observations based on the following methods: (a) *mls*: select the observation at the raster cell with the maximum landscape similarity value; (b) *mrv*: select the observation whose response value is the median of all the values; and (c) *random*: select an observation at random. Note that the latter represents a case of stratified random sampling.

The code below shows the selection of representative soil organic carbon (SOC) observations based on the maximum landscape similarity method. Note that the arguments `su.rast` and `ls.rast` require the stratification units and landscape similarity layers previously created with `strata()` and `similarity()`, respectively.

```
# SpatVector with SOC observations for stratification units
soc.obs <- vect("soc.shp")
# Representative SOC observation for each stratification unit
su.obs <- observation(su.rast = su, obs = soc.obs, col.id = 1, col.resp = 2,
                      method = "mls", ls.rast = su.ls$landsim)
# Plot results
figure(13, d = list(su.obs, soc.obs, su))
```

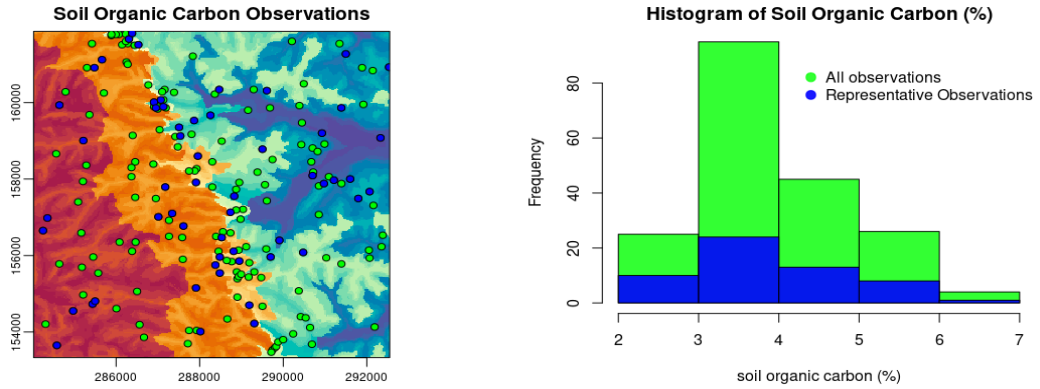

Figure 14: Selection of representative observations. Green points in the map represent the complete set of observations. Blue points represent the representative observation for each stratification unit.

The function `locations()` performs the automatic selection of the representative sampling location(s) for each stratification unit in a given set, where the representative sampling location is the raster cell where the highest landscape similarity value occurs. Currently, `locations()` implements two selection methods: (a) *buffer*: select sampling locations within areas with landscape similarity values above a certain threshold; and (b) *absolute*: select sampling locations with the highest landscape similarity values. The code below shows the use of `locations()` based on the buffer method.

```
# Representative sampling location and its buffer area for each stratification unit
su.samp <- locations(ls.rast = su.ls$landsim, su.rast = su, method = "buffer")
# Plot results
figure(14, d = list(su.samp, su))
```

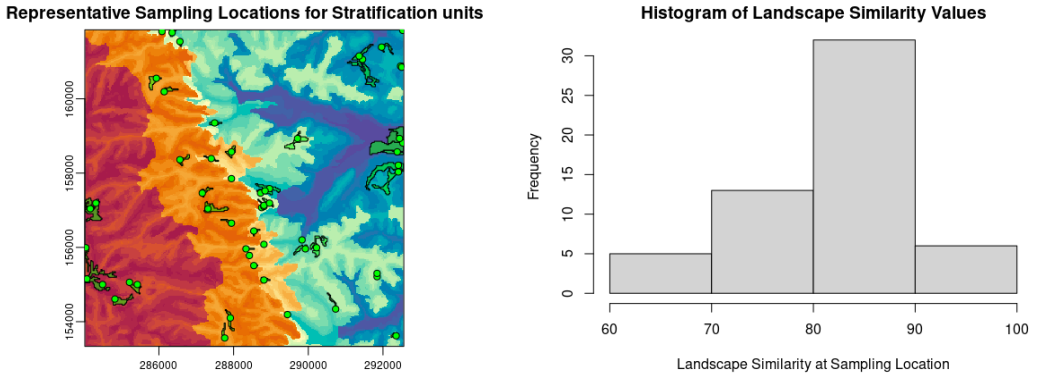

Figure 15: Selection of representative sampling locations. Green points in the map represent the sampling location for each stratification unit. Green polygons represent the buffer area for each sampling location.

Predictive modeling

Predictive modeling with **rassta** is based on the assumption that each stratification unit represents a distinct landscape configuration and that this configuration influences a natural phenomenon in a distinctive manner. It is assumed that the influence that a stratification unit's landscape configuration has on response phenomena at a specific location (i.e, raster cell) is proportional to the unit's landscape similarity value at that raster cell. Therefore, given a stratification unit x, the corresponding raster layer of landscape similarity ls, the response y, and a raster cell c, the greater the value of ls at c, the more similar y at c will be to the typical y for x. The typical (i.e., representative) value of a response phenomenon for a given stratification unit can be defined in several ways. For instance, if a response phenomenon was sampled/measured multiple times within a given stratification unit, the typical response value could be that from the sample/measurement at the raster cell with the highest landscape similarity value (see `observation()`).

Several studies have used landscape similarity layers to model the spatial variability of natural phenomena. These studies argue that the use of similarity layers is appropriate in cases when (a)

available observations for modeling are limited (Zhu et al. 2008); (b) initial spatial distribution patterns are needed for survey design (Carpenter, Gillison, and Winter 1993); (c) expert-driven selection of informative variables is possible (Knick and Dyer 1997); (d) *a priori* knowledge of response-control relationships in the form of conceptual models is available (Zhu et al. 2010; Schmidt, Tonkin, and Hewitt 2005); and (e) discriminating between (ecologically) positive and negative deviations from reference environments is required (Watrous et al. 2006). Accordingly, `engine()` allows the modeling of environmental phenomena with a number of training observations as few as the number of landscape similarity layers [cases (a) and (b)]; training observations and landscape similarity layers as outcomes of expert-driven landscape stratification [cases (c) and (d)]; and landscape similarity layers derived from spatial signatures that discriminate between the tails of distribution functions [case (e)].

Modeling with **rassta** is performed using the function `engine()`. For continuous responses, `engine()` performs a weighted average involving representative response values and landscape similarity layers. For a raster cell c, the modeled response value is equal to the weighted average of the representative values for those stratification units with the highest landscape similarity values at c. The stratification units with the highest landscape similarity values at c can be considered as the *nearest neighbors* (in feature space) of the landscape configuration at c. These nearest neighbors are called *winning stratification units*, and the weight of their corresponding representative value is proportional to the winning unit's landscape similarity value at c. For categorical responses, the modal response value of the winning stratification units replaces the weighted average. Figure 16 shows an example of the modeling process for continuous responses with **rassta**.

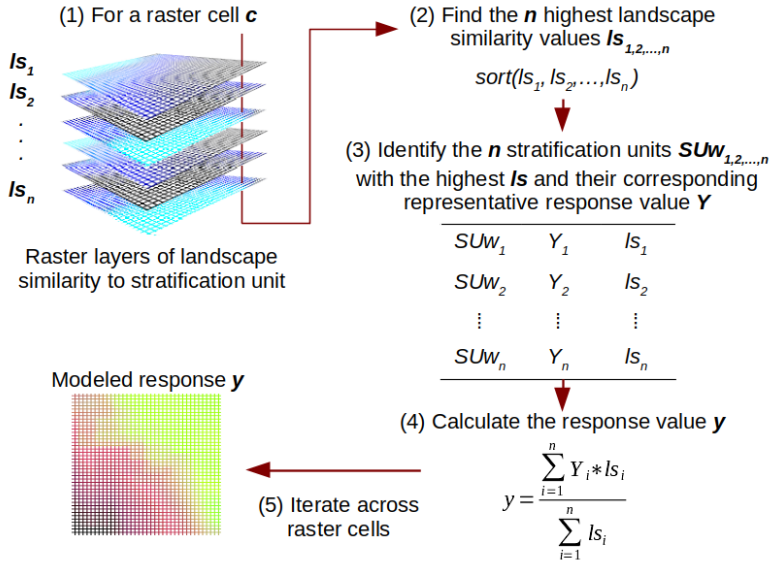

Figure 16: Schematic of the modeling process with rassta. The modeling process is performed in a cell-wise fashion. The inputs required are the raster layers of landscape similarity and the representative observations for each stratification unit.

Note that the weighted average for modeling phenomena across geographic space has been widely applied in GIS-based multicriteria decision analysis (GIS-MCDA). In GIS-MCDA, attributes (i.e., variables) in the form of raster layers are weighted according to expert criteria. The weighted variables are then combined through (cell-wise) overlay operators such as multiplication, addition and (ordered) averaging. The resulting value at each cell represents the relative suitability for a certain condition/decision (Malczewski 2006). The function `engine()` generalizes the weighted overlay process of GIS-MCDA by allowing the use of sampled/measured data of a response phenomenon in conjunction with the landscape similarity layers acting as weighted variables. This generalization allows the modeling of real-valued phenomena in continuous or categorical form. The modeling approach of `engine()` is almost the same as that proposed by Zhu (1997) to model landscape attributes across geographic space. The difference between `engine()` and the approach of Zhu (1997) is that `engine()` allows the selection of the number of landscape similarity layers for the weighted average calculation. Presumably, restricting the number of layers will reduce the shortening ('shrinking') effect that weighted averaging has on the range of modeled continuous response values (Nolan et al. 2019).

The code below demonstrates the use of `engine()` for the predictive modeling of soil organic carbon. Note that the representative response values (argument `su.repobs`) are those previously selected with `observation()`, and that the layers of landscape similarity (argument `ls.rast`) are those

previously created with `similarity()`.

```
# Table with the numeric code of stratification units and representative SOC values
su.soc <- su.obs$su_repobs[, c("SU", "soc")]
# engine() requires a (tiled) SpatVector with the boundaries of the area of interest
aoi <- vect("aoi.shp")
# engine() writes results directly on disk
if (dir.exists("soc") == FALSE) {dir.create("soc")}  # Create directory
# Spatial modeling of SOC across the landscape based on 3 winning stratification units
soc <- engine(ls.rast = su.ls$landsim, n.win = 3, su.repobs = su.soc,
              tiles = aoi, outdir = "soc", overwrite = TRUE)
figure(16, d = list(soc, "soc_valid.shp")) # Plot results
```

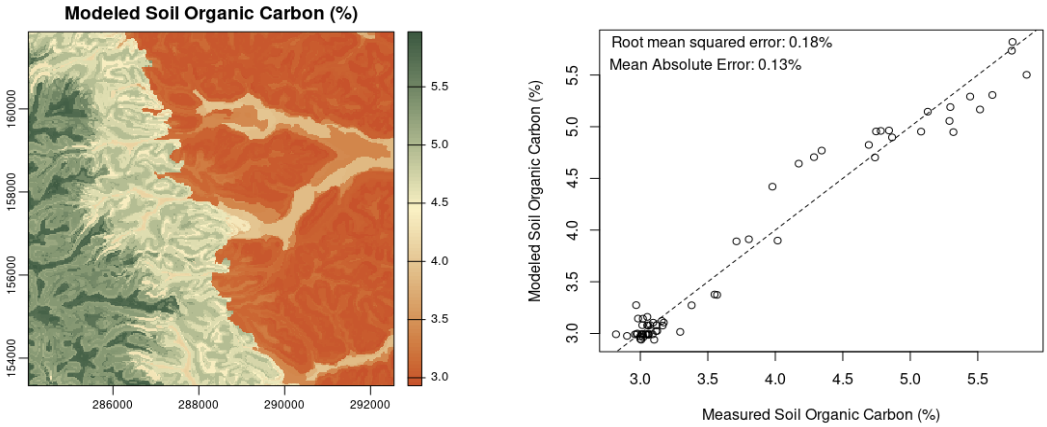

Figure 17: Modeled soil organic carbon (SOC) content (percent). The map shows the modeled SOC values across the landscape. The plot shows the the modeled (y) versus the measured (x) SOC values based on 62 independent observations.

Miscellaneous

The spatial signature only applies to classification units created from continuous variables. Thus, spatial signatures cannot be calculated for classification units that represent categorical variables, such as land use/land cover. In such cases, a one-hot encoding can be applied to produce binary layers for the units. These layers are considered the spatial signatures of the classification units. The code below shows the creation of binary layers for soil parent material units with `dummies()`.

```
# Multi-layer SpatRaster of soil parent material units
mat.cu <- rast("material.tif")
# Binary layers for each soil parent material unit and their maps
mat.sig <- dummies(mat.cu, preval = 100, absval = 0)
figure(17, d = mat.sig) # Plot results
```

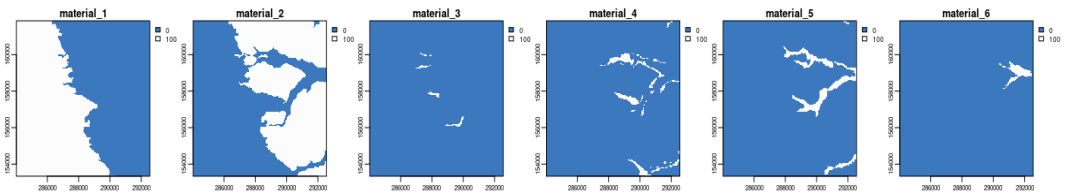

Figure 18: Construction of binary layers. Binary layers act as the spatial signatures for categorical variables. In this example, soil parent material acts as both landscape factor and classification units.

The function `plot3D()` produces interactive maps showing the 3-dimensional (*XYZ*) variability in raster layers representing continuous variables. The *XYZ* reference positions are obtained from a user-supplied elevation layer. For large raster layers (large spatial coverage and/or high spatial resolution), this function allows the option to decrease resolution and subset the data. The code below shows how `plot3D()` creates a 3D map for SOC, as modeled with `engine()`.

```
# Single-layer SpatRaster of terrain elevation and the 3D SOC map
elev <- rast("elevation.tif")
plot3D(c(elev, soc), z = 1, ex = 0.2, pals = "Fall", rev = TRUE) # 3D map
```

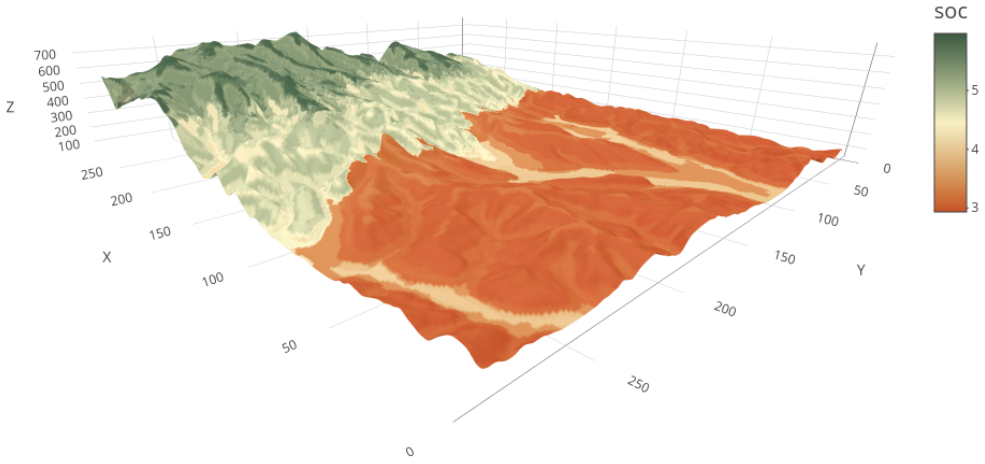

Figure 19: 3D map of SOC (percent). The Z dimension is obtained from a reference terrain model. Visit the online article to access the interactive version of the map.

3 Future versioning and summary

This work presented the **rassta** package for spatial stratification, sampling, and modeling of environmental phenomena within the R environment. Future versioning of the **rassta** package will focus on developing new approaches for spatial stratification. Stratification based on spatial intersection may not be feasible to implement in highly complex landscapes because these landscapes may require many (sets of) classification units to accurately represent the spatial variability of landscape factors, leading to over-stratification, and thus, greater demand for samples/observations to conduct predictive modeling based on landscape similarity. One plausible solution is the application of the stratification methods presented by Jasiewicz, Netzel, and Stepinski (2015), Jasiewicz, Stepinski, and Niesterowicz (2018), Nowosad (2021), and Nowosad and Stepinski (2021). However, these methods have been purposely designed for studies with continental/global applications. Therefore, these methods should be adapted for **rassta** to tailor their application at local scales to allow for more precise representations of natural phenomena and their spatial variability. Another focus of versioning can be new functions to visualize the variability of response phenomena relative to the hierarchical structure represented by the stratification units. Lastly, future versioning of **rassta** should also consider the user's experiences to ensure its general applicability.

The core ideas implemented in the **rassta** package include the multi-scale, hierarchical landscape stratification based on spatial intersection, the application of non-parametric distribution estimators to define the typical landscape configuration of stratification units, and the use of spatially explicit landscape correspondence metrics for non-probability sampling and predictive modeling. Some of these ideas have previously been implemented in R through a few packages dedicated to the analysis of geospatial data. Nevertheless, **rassta** offers a unified, generalized framework to conduct multiple landscape stratification routines through a dedicated set of algorithms. Moreover, spatially-explicit information created with **rassta**, like stratification units, landscape similarity layers, and representative observations, can be embedded into statistically robust modeling approaches to optimize the analysis of environmental phenomena.

References

Allen, T. F. H., and B. Starr. 1982. *Hierarchy: Perspectives for Ecological Complexity*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226489711.001.0001.

Araújo, Miguel B., and Antoine Guisan. 2006. "Five (or so) Challenges for Species Distribution Modelling." *Journal of Biogeography* 33 (10): 1677–88. https://doi.org/10.1111/j.1365-2699.2006.01584.x.

Arrouays, D., A. McBratney, J. Bouma, Z. Libohova, A. Richer-de-Forges, C. Morgan, P. Roudier, L. Poggio, and V. Mulder. 2020. "Impressions of Digital Soil Maps: The Good, the Not so Good, and

Making Them Ever Better." *Geoderma Regional* 20: e00255. https://doi.org/10.1016/j.geodrs.2020.e00255.

Austin, M. P., and P. C. Heyligers. 1989. "New Approach to Vegetation Survey Design: Gradsect Sampling." *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* 5: 31–36.

Baldwin, Doug, Kusum Naithani, and Henry Lin. 2017. "Combined Soil-Terrain Stratification for Characterizing Catchment-Scale Soil Moisture Variation." *Geoderma* 285: 260–69. https://doi.org/10.1016/j.geoderma.2016.09.031.

Burrough, Peter. 1989. "Fuzzy Mathematical Methods for Soil Survey and Land Evaluation." *Journal of Soil Science* 40 (3): 477–92. https://doi.org/10.1111/j.1365-2389.1989.tb01290.x.

Carpenter, Guy, A. N. Gillison, and J. Winter. 1993. "DOMAIN: A Flexible Modelling Procedure for Mapping Potential Distributions of Plants and Animals." *Biodiversity & Conservation* 2 (6): 667–80. https://doi.org/10.1007/bf00051966.

Chang, W., J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges. 2021. **shiny***: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

Chuvieco, E., I. Aguado, M. Yebra, H. Nieto, J. Salas, M. Pilar Martín, L. Vilar, et al. 2010. "Development of a Framework for Fire Risk Assessment Using Remote Sensing and Geographic Information System Technologies." *Ecological Modelling* 221 (1): 46–58. https://doi.org/10.1016/j.ecolmodel.2008.11.017.

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

Elith, Jane, Michael Kearney, and Steven Phillips. 2010. "The Art of Modelling Range-Shifting Species." *Methods in Ecology and Evolution* 1 (4): 330–42. https://doi.org/10.1111/j.2041-210x.2010.00036.x.

Elith, Jane, and John R. Leathwick. 2009. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time." *Annual Review of Ecology, Evolution, and Systematics* 40: 677–97. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Flügel, Wolfgang-Albert. 1995. "Delineating Hydrological Response Units by Geographical Information System Analyses for Regional Hydrological Modelling Using PRMS/MMS in the Drainage Basin of the River Bröl, Germany." *Hydrological Processes* 9 (3-4): 423–36. https://doi.org/10.1002/hyp.3360090313.

Gillison, Andrew. 1983. "Gradient Oriented Sampling for Resource Surveys - the Gradsect Method." *Survey Methods for Nature Conservation* 2: 349–74.

Gillison, Andrew, and Kenneth Brewer. 1985. "The Use of Gradient Directed Transects or Gradsects in Natural Resource Surveys." *Journal of Environmental Management* 20: 103–27.

Glinka, Konstantin. 1927. "Dokuchaev's Ideas in the Development of Pedology and Cognate Sciences." *Russian Pedological Investigations* 1.

Guisan, Antoine, and Niklaus Zimmermann. 2000. "Predictive Habitat Distribution Models in Ecology." *Ecological Modelling* 135 (2): 147–86. https://doi.org/10.1016/S0304-3800(00)00354-9.

Ham, Jisoo, Yangchi Chen, Melba Crawford, and Joydeep Ghosh. 2005. "Investigation of the Random Forest Framework for Classification of Hyperspectral Data." *IEEE Transactions on Geoscience and Remote Sensing* 43 (3): 492–501. https://doi.org/10.1109/tgrs.2004.842481.

Heuvelink, G., and R. Webster. 2001. "Modelling Soil Variation: Past, Present, and Future." *Geoderma* 100 (3): 269–301. https://doi.org/10.1016/S0016-7061(01)00025-8.

Hijmans, Robert J. 2021. **terra***: Spatial Data Analysis*. https://CRAN.R-project.org/package=terra.

Hijmans, Robert J., Steven Phillips, John Leathwick, and Jane Elith. 2020. **dismo***: Species Distribution Modeling*. https://CRAN.R-project.org/package=dismo.

Hirzel, Alexandre, Véronique Helfer, and F. Metral. 2001. "Assessing Habitat-Suitability Models with a Virtual Species." *Ecological Modelling* 145 (2-3): 111–21. https://doi.org/10.1016/S0304-3800(01)00396-9.

Hudson, Berman. 1992. "The Soil Survey as Paradigm-Based Science." *Soil Science Society of America Journal* 56 (3): 836–41. https://doi.org/10.2136/sssaj1992.03615995005600030027x.

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

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

Knick, Steven, and Deanna Dyer. 1997. "Distribution of Black-Tailed Jackrabbit Habitat Determined by GIS in Southwestern Idaho." *The Journal of Wildlife Management*, 75–85. https://doi.org/10.2307/3802416.

Kohonen, Teuvo. 1990. "The Self-Organizing Map." *Proceedings of the IEEE* 78 (9): 1464–80. https:

//doi.org/10.1016/s0925-2312(98)00030-7.

Leroy, Boris, Christine Meynard, Céline Bellard, and Franck Courchamp. 2016"**virtualspecies**, an R Package to Generate Virtual Species Distributions." *Ecography* 39 (6): 599–607. https://doi.org/10.1111/ecog.01388.

MacMillan, R., W. Pettapiece, S. Nolan, and T. Goddard. 2000. "A Generic Procedure for Automatically Segmenting Landforms into Landform Elements Using DEMs, Heuristic Rules and Fuzzy Logic." *Fuzzy Sets and Systems* 113 (1): 81–109. https://doi.org/10.1016/S0165-0114(99)00014-7.

Malczewski, Jacek. 2006. "GIS-Based Multicriteria Decision Analysis: A Survey of the Literature." *International Journal of Geographical Information Science* 20 (7): 703–26. https://doi.org/10.1080/13658810600661508.

McBratney, Alex, M. L. Mendonça Santos, and Budiman Minasny. 2003. "On Digital Soil Mapping." *Geoderma* 117 (1-2): 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

McSweeney, K., B. K. Slater, R. David Hammer, J. C. Bell, P. E. Gessler, and G. W. Petersen. 1994. "Towards a New Framework for Modeling the Soil-Landscape Continuum." *Factors of Soil Formation: A Fiftieth Anniversary Retrospective* 33: 127–45. https://doi.org/10.2136/sssaspecpub33.c8.

Nolan, Connor, John Tipton, Robert K Booth, Mevin B Hooten, and Stephen T Jackson. 2019. "Comparing and Improving Methods for Reconstructing Peatland Water-Table Depth from Testate Amoebae." *The Holocene* 29 (8): 1350–61. https://doi.org/10.1177/0959683619846969.

Nowosad, Jakub. 2021. "**motif**: An Open-Source R Tool for Pattern-Based Spatial Analysis." *Landscape Ecology* 36 (1): 29–43. https://doi.org/10.1007/s10980-020-01135-0.

Nowosad, Jakub, and Tomasz Stepinski. 2021. "Pattern-Based Identification and Mapping of Landscape Types Using Multi-Thematic Data." *International Journal of Geographical Information Science* 35 (8): 1634–49. https://doi.org/10.1080/13658816.2021.1893324.

Park, S. J., and N. Van De Giesen. 2004. "Soil-Landscape Delineation to Define Spatial Sampling Domains for Hillslope Hydrology." *Journal of Hydrology* 295 (1-4): 28–46. https://doi.org/10.1016/j.jhydrol.2004.02.022.

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.

Pike, Richard. 1988. "The Geometric Signature: Quantifying Landslide-Terrain Types from Digital Elevation Models." *Mathematical Geology* 20 (5): 491–511. https://doi.org/10.1007/BF00890333.

Pike, Richard, and Wesley Rozema. 1975. "Spectral Analysis of Landforms." *Annals of the Association of American Geographers* 65 (4): 499–516. https://doi.org/10.1111/j.1467-8306.1975.tb01058.x.

Rodrigues, Marcos, Sergi Costafreda-Aumedes, Carles Comas, and Cristina Vega-García. 2019. "Spatial Stratification of Wildfire Drivers Towards Enhanced Definition of Large-Fire Regime Zoning and Fire Seasons." *Science of the Total Environment* 689: 634–44. https://doi.org/10.1016/j.scitotenv.2019.06.467.

Schmidt, Jochen, Phil Tonkin, and Allan Hewitt. 2005. "Quantitative Soil - Landscape Models for the Haldon and Hurunui Soil Sets, New Zealand." *Soil Research* 43 (2): 127. https://doi.org/10.1071/sr04074.

Scull, Peter, Janet Franklin, Oliver Chadwick, and D. McArthur. 2003. "Predictive Soil Mapping: A Review." *Progress in Physical Geography* 27 (2): 171–97. https://doi.org/10.1191/0309133303pp366ra.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. "Estimating the Number of Clusters in a Data Set via the Gap Statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2): 411–23. https://doi.org/10.1111/1467-9868.00293.

Watrous, Kristen, Therese Donovan, Ruth Mickey, Scott Darling, Alan Hicks, and Susanna Von Oettingen. 2006. "Predicting Minimum Habitat Characteristics for the Indiana Bat in the Champlain Valley." *The Journal of Wildlife Management* 70 (5): 1228–37. https://doi.org/10.2193/0022-541X(2006)70%5B1228:PMHCFT%5D2.0.CO;2.

Wehrens, Ron, and Johannes Kruisselbrink. 2018. "Flexible Self-Organizing Maps in Kohonen 3.0." *Journal of Statistical Software* 87 (1): 1–18. https://doi.org/10.18637/jss.v087.i07.

Wessels, K. J., S. Van Jaarsveld, J. D. Grimbeek, and M. J. Van der Linde. 1998. "An Evaluation of the Gradsect Biological Survey Method." *Biodiversity & Conservation* 7 (8): 1093–1121. https://doi.org/10.1023/a:1008899802456.

West, Amanda, Sunil Kumar, Cynthia Brown, Thomas Stohlgren, and Jim Bromberg. 2016. "Field Validation of an Invasive Species Maxent Model." *Ecological Informatics* 36: 126–34. https://doi.org/10.1016/j.ecoinf.2016.11.001.

Zadeh, Lofti A. 1965. "Fuzzy Sets." *Information and Control* 8 (1): 338–53. https://doi.org/10.1142/9789814261302_0001.

Zhu, A. Xing. 1997. "A Similarity Model for Representing Soil Spatial Information." *Geoderma* 77 (2-4): 217–42. https://doi.org/10.1016/S0016-7061(97)00023-2.

Zhu, A. Xing, and Lawrence Band. 1994. "A Knowledge-Based Approach to Data Integration for Soil Mapping." *Canadian Journal of Remote Sensing* 20 (4): 408–18. https://doi.org/10.1080/07038992.1994.10874583.

Zhu, A. Xing, Feng Qi, Amanda Moore, and James Burt. 2010. "Prediction of Soil Properties Using

Fuzzy Membership Values." *Geoderma* 158 (3-4): 199–206. https://doi.org/10.1016/j.geoderma.2010.05.001.

Zhu, A. Xing, Lin Yang, Baolin Li, Chengzhi Qin, Edward English, James Burt, and Chenghu Zhou. 2008. "Purposive Sampling for Digital Soil Mapping for Areas with Limited Data." In *Digital Soil Mapping with Limited Data*, 233–45. Springer. https://doi.org/10.1007/978-1-4020-8592-5_20.

Bibliography

T. F. H. Allen and B. Starr. *Hierarchy: Perspectives for Ecological Complexity*. University of Chicago Press, 1982. URL https://doi.org/10.7208/chicago/9780226489711.001.0001. [p]

M. B. Araújo and A. Guisan. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10):1677–1688, 2006. URL https://doi.org/10.1111/j.1365-2699.2006.01584.x. [p]

D. Arrouays, A. McBratney, J. Bouma, Z. Libohova, A. Richer-de Forges, C. Morgan, P. Roudier, L. Poggio, and V. Mulder. Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional*, 20:e00255, 2020. URL https://doi.org/10.1016/j.geodrs.2020.e00255. [p]

M. P. Austin and P. C. Heyligers. New approach to vegetation survey design: gradsect sampling. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, 5:31–36, 1989. [p]

D. Baldwin, K. Naithani, and H. Lin. Combined soil-terrain stratification for characterizing catchment-scale soil moisture variation. *Geoderma*, 285:260–269, 2017. URL https://doi.org/10.1016/j.geoderma.2016.09.031. [p]

P. Burrough. Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science*, 40(3):477–492, 1989. URL https://doi.org/10.1111/j.1365-2389.1989.tb01290.x. [p]

G. Carpenter, A. N. Gillison, and J. Winter. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity & Conservation*, 2(6):667–680, 1993. URL https://doi.org/10.1007/bf00051966. [p]

W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges. **shiny***: Web Application Framework for R*, 2021. URL https://CRAN.R-project.org/package=shiny. R package version 1.6.0. [p]

E. Chuvieco, I. Aguado, M. Yebra, H. Nieto, J. Salas, M. Pilar Martín, L. Vilar, J. Martínez, S. Martín, P. Ibarra, J. de la Riva, J. Baeza, F. Rodríguez, J. Molina, M. Herrera, and R. Zamora. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. *Ecological Modelling*, 221(1):46–58, 2010. URL https://doi.org/10.1016/j.ecolmodel.2008.11.017. [p]

E. Eddelbuettel and R. François. **Rcpp**: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. URL https://10.18637/jss.v040.i08. [p]

J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009. URL https://doi.org/10.1146/annurev.ecolsys.110308.120159. [p]

J. Elith, M. Kearney, and S. Phillips. The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4):330–342, 2010. URL https://doi.org/10.1111/j.2041-210x.2010.00036.x. [p]

W.-A. Flügel. Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the river Bröl, Germany. *Hydrological Processes*, 9(3-4):423–436, 1995. URL https://doi.org/10.1002/hyp.3360090313. [p]

A. Gillison. Gradient oriented sampling for resource surveys - the gradsect method. *Survey Methods for Nature Conservation*, 2:349–374, 1983. [p]

A. Gillison and K. Brewer. The use of gradient directed transects or gradsects in natural resource surveys. *Journal of Environmental Management*, 20:103–127, 1985. [p]

K. Glinka. Dokuchaev's ideas in the development of pedology and cognate sciences. *Russian Pedological Investigations*, 1, 1927. [p]

A. Guisan and N. Zimmermann. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2):147–186, 2000. URL https://doi.org/10.1016/S0304-3800(00)00354-9. [p]

J. Ham, Y. Chen, M. Crawford, and J. Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3): 492–501, 2005. URL https://doi.org/10.1109/tgrs.2004.842481. [p]

G. Heuvelink and R. Webster. Modelling soil variation: past, present, and future. *Geoderma*, 100(3): 269–301, 2001. URL https://doi.org/10.1016/S0016-7061(01)00025-8. [p]

R. J. Hijmans. **terra***: Spatial Data Analysis*, 2021. URL https://CRAN.R-project.org/package=terra. R package version 1.3-4. [p]

R. J. Hijmans, S. Phillips, J. Leathwick, and J. Elith. **dismo***: Species Distribution Modeling*, 2020. URL https://CRAN.R-project.org/package=dismo. R package version 1.3-3. [p]

A. Hirzel, V. Helfer, and F. Metral. Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, 145(2-3):111–121, 2001. URL https://doi.org/10.1016/S0304-3800(01)00396-9. [p]

B. Hudson. The soil survey as paradigm-based science. *Soil Science Society of America Journal*, 56(3): 836–841, 1992. URL https://doi.org/10.2136/sssaj1992.03615995005600030027x. [p]

J. Jasiewicz, P. Netzel, and T. Stepinski. GeoPAT: A toolbox for pattern-based information retrieval from large geospatial databases. *Computers & Geosciences*, 80:62–73, 2015. URL https://doi.org/10.1016/j.cageo.2015.04.002. [p]

J. Jasiewicz, T. Stepinski, and J. Niesterowicz. Multi-scale segmentation algorithm for pattern-based partitioning of large categorical rasters. *Computers & Geosciences*, 118:122–130, 2018. URL https://doi.org/10.1016/j.cageo.2018.06.003. [p]

L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990. [p]

S. Knick and D. Dyer. Distribution of black-tailed jackrabbit habitat determined by GIS in Southwestern Idaho. *The Journal of Wildlife Management*, pages 75–85, 1997. URL https://doi.org/10.2307/3802416. [p]

T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990. URL https://doi.org/10.1016/s0925-2312(98)00030-7. [p]

B. Leroy, C. Meynard, C. Bellard, and F. Courchamp. **virtualspecies**, an R package to generate virtual species distributions. *Ecography*, 39(6):599–607, 2016. URL https://doi.org/10.1111/ecog.01388. [p]

R. MacMillan, W. Pettapiece, S. Nolan, and T. Goddard. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. *Fuzzy Sets and Systems*, 113(1):81–109, 2000. URL https://doi.org/10.1016/S0165-0114(99)00014-7. [p]

J. Malczewski. GIS-based multicriteria decision analysis: a survey of the literature. *International Journal of Geographical Information Science*, 20(7):703–726, 2006. URL https://doi.org/10.1080/13658810600661508. [p]

A. McBratney, M. L. Mendonça Santos, and B. Minasny. On digital soil mapping. *Geoderma*, 117(1-2): 3–52, 2003. URL https://doi.org/10.1016/S0016-7061(03)00223-4. [p]

K. McSweeney, B. K. Slater, R. David Hammer, J. C. Bell, P. E. Gessler, and G. W. Petersen. Towards a new framework for modeling the soil-landscape continuum. *Factors of Soil Formation: A Fiftieth Anniversary Retrospective*, 33:127–145, 1994. URL https://doi.org/10.2136/sssaspecpub33.c8. [p]

C. Nolan, J. Tipton, R. K. Booth, M. B. Hooten, and S. T. Jackson. Comparing and improving methods for reconstructing peatland water-table depth from testate amoebae. *The Holocene*, 29(8):1350–1361, 2019. URL https://doi.org/10.1177/0959683619846969. [p]

J. Nowosad. **motif**: an open-source R tool for pattern-based spatial analysis. *Landscape Ecology*, 36(1): 29–43, 2021. URL https://doi.org/10.1007/s10980-020-01135-0. [p]

J. Nowosad and T. Stepinski. Pattern-based identification and mapping of landscape types using multi-thematic data. *International Journal of Geographical Information Science*, 35(8):1634–1649, 2021. URL https://doi.org/10.1080/13658816.2021.1893324. [p]

S. J. Park and N. Van De Giesen. Soil-landscape delineation to define spatial sampling domains for hillslope hydrology. *Journal of Hydrology*, 295(1-4):28–46, 2004. URL https://doi.org/10.1016/j.jhydrol.2004.02.022. [p]

E. Pebesma. Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10(1): 439–446, 2018. URL https://doi.org/10.32614/RJ-2018-009. [p]

R. Pike. The geometric signature: quantifying landslide-terrain types from digital elevation models. *Mathematical Geology*, 20(5):491–511, 1988. URL https://doi.org/10.1007/BF00890333. [p]

R. Pike and W. Rozema. Spectral analysis of landforms. *Annals of the Association of American Geographers*, 65(4):499–516, 1975. URL https://doi.org/10.1111/j.1467-8306.1975.tb01058.x. [p]

M. Rodrigues, S. Costafreda-Aumedes, C. Comas, and C. Vega-García. Spatial stratification of wildfire drivers towards enhanced definition of large-fire regime zoning and fire seasons. *Science of the Total Environment*, 689:634–644, 2019. URL https://doi.org/10.1016/j.scitotenv.2019.06.467. [p]

J. Schmidt, P. Tonkin, and A. Hewitt. Quantitative soil - landscape models for the Haldon and Hurunui soil sets, New Zealand. *Soil Research*, 43(2):127, 2005. URL https://doi.org/10.1071/sr04074. [p]

P. Scull, J. Franklin, O. Chadwick, and D. McArthur. Predictive soil mapping: a review. *Progress in Physical Geography*, 27(2):171–197, 2003. URL https://doi.org/10.1191/0309133303pp366ra. [p]

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (statistical Methodology)*, 63(2):411–423, 2001. URL https://doi.org/10.1111/1467-9868.00293. [p]

K. Watrous, T. Donovan, R. Mickey, S. Darling, A. Hicks, and S. Von Oettingen. Predicting minimum habitat characteristics for the Indiana bat in the Champlain Valley. *The Journal of Wildlife Management*, 70(5):1228–1237, 2006. URL https://doi.org/10.2193/0022-541X(2006)70[1228:PMHCFT]2.0.CO;2. [p]

R. Wehrens and J. Kruisselbrink. Flexible self-organizing maps in kohonen 3.0. *Journal of Statistical Software*, 87(1):1–18, 2018. URL https://doi.org/10.18637/jss.v087.i07. [p]

K. J. Wessels, S. Van Jaarsveld, J. D. Grimbeek, and M. J. Van der Linde. An evaluation of the gradsect biological survey method. *Biodiversity & Conservation*, 7(8):1093–1121, 1998. URL https://doi.org/10.1023/a:1008899802456. [p]

A. West, S. Kumar, C. Brown, T. Stohlgren, and J. Bromberg. Field validation of an invasive species maxent model. *Ecological Informatics*, 36:126–134, 2016. URL https://doi.org/10.1016/j.ecoinf.2016.11.001. [p]

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(1):338–353, 1965. URL https://doi.org/10.1142/9789814261302_0001. [p]

A. X. Zhu. A similarity model for representing soil spatial information. *Geoderma*, 77(2-4):217–242, 1997. URL https://doi.org/10.1016/S0016-7061(97)00023-2. [p]

A. X. Zhu and L. Band. A knowledge-based approach to data integration for soil mapping. *Canadian Journal of Remote Sensing*, 20(4):408–418, 1994. URL https://doi.org/10.1080/07038992.1994.10874583. [p]

A. X. Zhu, L. Yang, B. Li, C. Qin, E. English, J. Burt, and C. Zhou. Purposive sampling for digital soil mapping for areas with limited data. In *Digital Soil Mapping with Limited Data*, pages 233–245. Springer, 2008. URL https://doi.org/10.1007/978-1-4020-8592-5_20. [p]

A. X. Zhu, F. Qi, A. Moore, and J. Burt. Prediction of soil properties using fuzzy membership values. *Geoderma*, 158(3-4):199–206, 2010. URL https://doi.org/10.1016/j.geoderma.2010.05.001. [p]

Bryan A. Fuentes
University of Arkansas
Department of Crop, Soil, and Environmental Sciences
Fayetteville, Arkansas
ORCiD: 0000-0003-3506-7101
bafuente@uark.edu

Minerva J. Dorantes
University of Arkansas
Department of Crop, Soil, and Environmental Sciences
Fayetteville, Arkansas
ORCiD: 0000-0002-2877-832X
mjdorant@uark.edu

John R. Tipton
University of Arkansas
Department of Mathematical Sciences
Fayetteville, Arkansas
ORCiD: 0000-0002-6135-8141
jrtipton@uark.edu