

9-19-2019

Quantitative spatial upscaling of categorical information: The multi-dimensional grid-point scaling algorithm

Daniel Gann

Follow this and additional works at: https://digitalcommons.fiu.edu/com_facpub



Part of the [Biology Commons](#)

This work is brought to you for free and open access by the Herbert Wertheim College of Medicine at FIU Digital Commons. It has been accepted for inclusion in HWCOC Faculty Publications by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

Quantitative spatial upscaling of categorical information: The multi-dimensional grid-point scaling algorithm

Daniel Gann^{1,2} 

¹Department of Biological Sciences, Florida International University, Miami, FL, USA

²Geographic Information Systems and Remote Sensing Center, Florida International University, Miami, FL, USA

Correspondence

Daniel Gann
Email: gannd@fiu.edu

Handling Editor: Robert Freckleton

Abstract

1. Categorical raster datasets often require upscaling to a lower spatial resolution to make them compatible with the scale of ecological analysis. When aggregating categorical data, two critical issues arise: (a) ignoring compositional information present in the high-resolution grid cells leads to high and uncontrolled loss of information in the scaled dataset; and (b) restricting classes to those present in the high-resolution dataset assumes validity of the classification scheme at the lower, aggregated resolution.
2. I introduce a new scaling algorithm that aggregates categorical data while simultaneously controlling for information loss by generating a non-hierarchical, representative, classification system for the aggregated scale. The Multi-Dimensional Grid-Point (MDGP) scaling algorithm acknowledges the statistical constraints of compositional count data. In a neutral-landscape simulation study implementing a full-factorial design for landscape characteristics, scale factors and algorithm parameters, I evaluated consistency and sensitivity of the scaling algorithm. Consistency and sensitivity were assessed for compositional information retention (IR_{cmp}) and class-label fidelity (CLF, the probability of recurring scaled class labels) for neutral random landscapes with the same properties.
3. The MDGP-scaling algorithm consistently preserved information at a significantly higher rate than other commonly used algorithms. Consistency of the algorithm was high for IR_{cmp} and CLF, but coefficients of variation of both metrics across landscapes varied most with class-abundance distribution. A diminishing return for IR_{cmp} was observed with increasing class-label precision. Mean class-label recurrence probability was consistently above 75% for all simulated landscape types, scale factors and class-label precisions.
4. The MDGP-scaling algorithm is the first algorithm that generates data-driven, scale-specific classification schemes while conducting spatial data aggregation. Consistent gain in IR_{cmp} and the associated reproducibility of classification systems strongly suggest that the increased precision of scaled maps will improve ecological models that rely on upscaling of high-resolution categorical raster data.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Author. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

KEYWORDS

categorical raster data, compositional count data, multi-dimensional grid-point, neutral landscape models, scaling algorithm, simplex, simulation, spatial scaling

1 | INTRODUCTION

Spatially explicit ecological models rely on spatially exhaustive data layers at appropriate scales for the ecological process of interest (Lam & Quattrochi, 1992; Mas, Gao, & Pacheco, 2010), which often requires scaling datasets to the scale of analysis. Upscaling of raster data to coarser spatial resolutions aggregates data of multiple high-resolution grid cells into lower resolution grid cells. As data are aggregated, generalization leads to information loss. Since the goal of data aggregation is to retain sufficient information relevant to a scientific question addressing a phenomenon at the aggregated scale, it is important to quantify the amount of information retained in the aggregated product and to control information loss.

Classification schemes of categorical data are valid for the range of spatial scales for which they were defined. However, commonly applied spatial-aggregation methods *majority rule*, *nearest-neighbor rule*, or *random rule*, or more complex methods such as spatial scan statistic (Coulston, Zaccarelli, Riitters, Koch, & Zurlini, 2014), only consider the original, high-resolution class scheme when assigning class labels to aggregated, larger landscape units. Complex spatial co-occurrence patterns are oversimplified, often resulting in vastly

increasing abundance of dominant classes and elimination of rare classes (Gann, Richards, & Biswas, 2012; He, Ventura, & Mladenoff, 2002; Ju, Gopal, & Kolaczyk, 2005). The presumption that the original class descriptors are valid at the aggregated lower resolution, regardless of scale factor, leads to uncontrolled loss of information content in each grid cell of the aggregated map, and potentially to fallacy in ecological models that use the oversimplified aggregated data.

For illustration, consider information loss for a landscape with two cover classes, 'grass' and 'tree' (Figure 1a). Aggregating the landscape subset of 49 grid cells with a class percentage distribution of 73% 'grass' and 27% 'tree' cover into a single coarser resolution cell requires a new class label assignment. The *majority rule*, a simple plurality decision rule that assigns the output category with the highest proportion of sub-samples (i.e. mode) applies the label 'grass' (Figure 1a). The *nearest-neighbour rule* assigns the 'tree' class, the category closest to the centre of the scaled grid cell, and the *random rule* assigns the output class at random (Figure 1a). Application of these three algorithms to the same input data results in completely different class assignments to the up-scaled grid cell, resulting in pure (100% cover) classes of either 'tree' or 'grass'. Assignment to a single class reduces the compositional information content of the aggregated grid

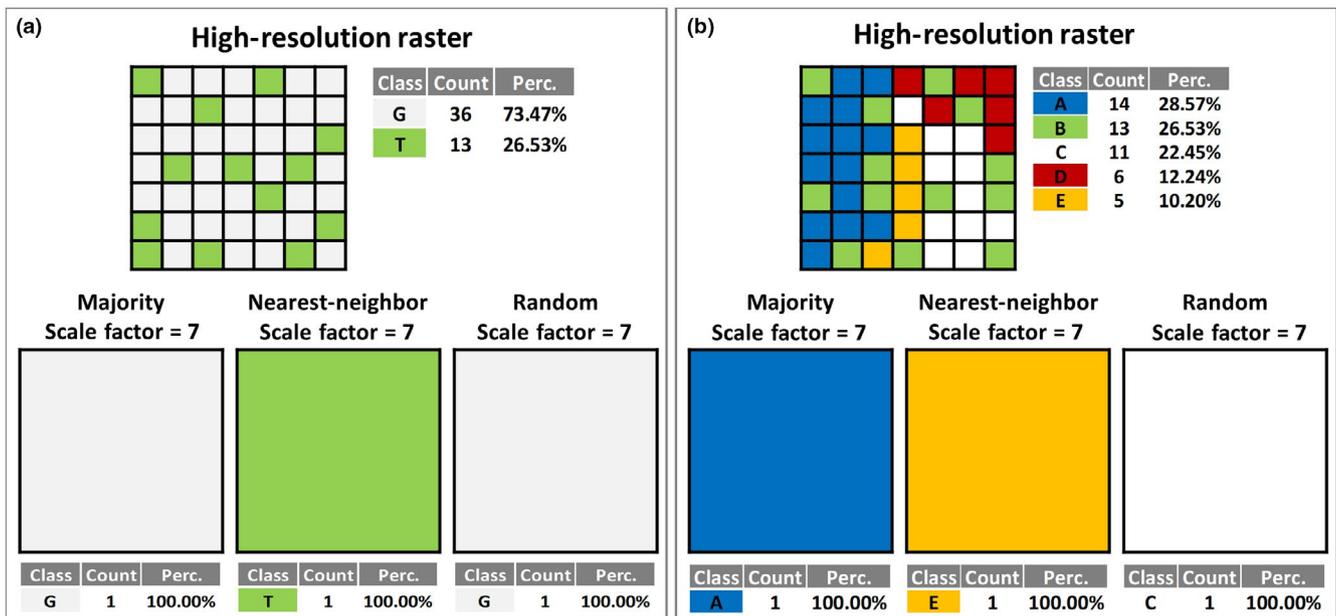


FIGURE 1 Information loss associated with *majority-*, *nearest-neighbor-* and *random-rule* scaling algorithms. (a) Applying the three algorithms to the two-class ('grass' (G) and 'tree' (T)) landscape example shows that none of the algorithms captures the more intuitive class of 'Woodland' at the aggregation scale. Depending on the algorithm, the amount of compositional information that is lost is either ~27%, if class 'G' is assigned, or ~73%, if class 'T' is assigned. (b) As richness increases, assigning a single input class label to the aggregated output cell represents only one of five original classes that were present, and the maximum compositional information that can be retained is less than 30%. A mixed class label that captures the heterogeneity is required at this scale to represent the landscape. In all three cases, four classes are omitted from the scaled class label. The single-class scaled class label over-represents its class with 100%, when in fact that class was present at only 28.57% for the outcome of the *majority rule*, 22.45% for the *random rule* and 10.2% for the *nearest-neighbor rule*

cells to 73% when assigning class 'grass', or to 27% when class 'tree' is assigned. With an increasing number of classes within a sample, compositional information decreases upon aggregation (Figure 1b).

Subsequently, if the aggregated product is used in combination with remotely sensed data to retrieve biophysical properties of the landscape, for instance biomass, then the biomass estimate will be coarsely over- or underestimated because it is much larger for trees than it is for grass. Magnitudes of error and uncertainty of estimated properties depend on the proportional misrepresentation of the generalized scaled landscape unit, and the difference in biophysical properties of the respective classes that are over- or underrepresented. In the example, to capture the relative abundance of both classes at the scale of lower resolution, a more appropriate class might have been 'Woodland', not a label option, since it did not exist in the original map. To date, no spatial aggregation algorithm generates scale-specific representative classes as landscape units are aggregated.

Several sub-disciplines of ecology have addressed defining representative classification schemes on the basis of quantitative measures of species co-occurrence data (Braun-Blanquet, 1932; De Cáceres et al., 2015; Mucina, 1997; Van Der Maarel, 1979). Species-association patterns, when randomly sampled on a 1-m² scale across a defined spatial extent, are expected to differ from the association patterns of the same species on a 50-m² scale (O'Neill et al., 1996; Schlup & Wagner, 2008). Consequently, plant communities vary along the continuum of spatial scales and definitions of communities or vegetation classes depend on the scale at which the landscape is sampled. Methods that generate scale-specific classification schemes from samples have to be consistent in delivering class descriptors (labels) that are reproducible and representative for the population, that is, the sampled landscape at the scale of interest (De Cáceres, Font, Vicente, & Oliva, 2009; De Cáceres & Wiser, 2012; Tichý, Chytrý, Hájek, Talbot, & Botta-Dukát, 2010; Tichý, Chytrý, & Šmarda, 2011; Wildi, 2010). The principles of reproducibility and representativeness also apply to scaling methods that aim to generate scale-specific classification schemes, recognizing class co-occurrence variability at different scales. An algorithm that implements scaling of classification schemes preferably also provides a control mechanism for information loss, considering the relative class abundance (composition) and the spatial arrangement (configuration) of sub-samples within samples. The algorithm presented here addresses the compositional information retention aspect.

2 | MATERIALS AND METHODS

2.1 | Defining the sample space

The sample space of spatially explicit, categorical data is finite and discrete, and samples of local neighborhoods result in count frequencies of classes. Sample space and relative class abundance distributions within samples depend on (a) diversity and spatial characteristics of the landscape and (b) the scale factor.

Richness (*rch*), the number of distinct classes, and evenness, which refers to the class abundance distribution (CAD) across the

landscape, define diversity. The spatial distribution patterns of the classes across the landscape range from systematic to random and from highly dispersed to completely aggregated.

Scale factor (*sf*) is the ratio of the spatial resolution of the scaled grid to the resolution of the original, high-resolution grid; when squared, *sf* provides the number of sub-units or grid cells within a sample (N_{smpl}). For instance, if the resolution of the original raster is 1 m and the scaled grid resolution is 7 m, the scale factor is 7 and the number of sample sub-units N_{smpl} is 49 (sf^2).

With *sf* and *rch* greater than 1, the number of possible distinct sample outcomes is the number of restricted or weak compositions with binomial coefficients:

$$\binom{N_{\text{smpl}} + rch - 1}{rch - 1}, \text{ or} \quad (1)$$

$$\frac{(N_{\text{smpl}} + rch - 1)!}{(rch - 1)! * (N_{\text{smpl}} + rch - 1 - (rch - 1))!}$$

Since percent-cover per sample is constrained to exactly 100%, the precision (*Prc*) of relative abundance of class (*c*) is:

$$Prc_c = \frac{100}{N_{\text{smpl}}} \quad (2)$$

For a given *sf*, as *rch* increases (Table 1, rows), or for a given *rch*, as *sf* increases (Table 1, columns), the number of unique compositions increases rapidly. The precision or granularity of measurement of relative class abundance of a sample is solely determined by *sf* (Table 1, Equation 2). For a given landscape with a specific *rch* and scaled with a specific *sf*, the frequency distribution of each possible composition then depends on CAD and the spatial dispersion or aggregation pattern of the classes across the landscape.

The constraint that the sum of all sample proportions = 1 makes the data compositional in nature (Aitchison, 1986). The sample space of compositional data is called the *simplex* or S^D (Aitchison, 1986).

$$S^D = \left\{ x = [x_1; \dots; x_D] \mid x_i \geq 0 \text{ and } \sum_{i=0}^D x_i = \kappa \right\} \quad (3)$$

The constraint of the *simplex* is that all $x_i \geq 0$, and that the sum of all $x_i = 1$. When dealing with count compositions (i.e. integers), quantitative grouping or classification methods that use distance metrics from the real space R and that assume multivariate normal distributions are inadequate and lead to spurious statistical results (Aitchison, 1986; van den Boogaart & Tolosana-Delgado, 2008; Jackson, 1997). Proposed solutions for statistical analysis of compositional data are log-ratio transformations of compositional data (Aitchison & Egozcue, 2005; Egozcue, Pawłowsky-Glahn, Mateu-Figueras, & Barceló-Vidal, 2003), which then allow application of analytical methods that are valid in R space. Log-ratio transformation, however, is not defined for count compositions with count zero, and the methods that have been proposed to deal with zero count

TABLE 1 Number of weak compositions for compositions with constraint of exactly 100% coverage. Precision is 100% divided by the number of subsamples which is the scale factor squared

Richness	Number of weak combinations					
	Scale factor					
	3	5	7	9	15	25
2	10	26	50	82	226	626
3	55	351	1,275	3,403	25,651	196,251
4	220	3,276	22,100	95,284	1,949,476	41,081,876
5	715	23,751	292,825	2,024,785	111,607,501	
6	2,002	142,506	3,162,510	34,826,302		
7	5,005	736,281	28,989,675			
8	11,440	3,365,856				
9	24,310	13,884,156				
10	48,620	52,451,256				
Precision	11.11%	4%	2.04%	1.23%	0.44%	0.16%

data add noise to the data (Martín-Fernández, Barceló-Vidal, & Pawłowsky-Glahn, 2003; Martín-Fernández, Hron, Templ, Filzmoser, & Palarea-Albaladejo, 2014). When scaling categorical raster data, samples will contain zeros when not every class is present in every sample. The number of zeros in a sample depends on landscape diversity and sf . Observations with zeroes are anticipated whenever sf is small relative to rch and are always present when sf produces a sub-sample unit count less than rch . The frequency of zeros in a sample increases with rch and $sptAgg$ and is present in every sample when rch exceeds the number of sub-samples (sf^2). Hence, a scaling algorithm that generates representative classification schemes has to be robust in dealing with compositional data samples that have a high frequency of zeros.

2.2 | The multi-dimensional grid-point scaling algorithm

The new multi-dimensional grid-points (MDGP) scaling algorithm presented here conducts spatial aggregation of categorical data while simultaneously generating a non-hierarchical, representative classification system for the aggregated scale. The algorithm allows for user-control of information retention while addressing the constraints of the sample space of compositional data with a high probability of zeros in the sample data.

The scaling algorithm performs two integrated tasks: (a) classification (grouping) of landscape objects (scaled grid cells) on the basis of relative abundance of classes within the samples, resulting in a scale-specific classification system that is representative at the scale of aggregation; and (b) assignment of all spatially aggregated units of the landscape to one of the scaled classes in the new classification system. The algorithm recognizes S^D as the multi-dimensional feature space spanned by compositional data, resulting in polytopes, where the number of features (i.e. richness) defines the number of vertices of the polytope and with equal unit distance of all vertices.

For a given landscape, richness (number of original classes) and scale factor determine the number and location of regularly spaced multi-dimensional grid-points (MDGP) in the solid space of the polytope. As richness and scale factor increase, the number of possible scaled class combinations (i.e. scaled richness) increases (Table 1). With increasing scale factor, however, precision of class proportion (the distance between points in the evenly-spaced grid) rapidly increases beyond ecological and, in many cases, statistical significance (Table 1). The MDGP-scaling algorithm limits scaled class-label precision by implementing a partitioning parameter, which reduces the number of possible grid points (Table in Figure 2). The 'parts' parameter partitions each dimension of the sample space (0%–100%) into equal parts. The result is a polytope with regularly spaced MDGPs, where the number of vertices is still equal to the number of original classes, but now the number of partitions in each dimension determines the number of MDGPs (Figure 2). The 1-part partition is equivalent to the *majority-rule*, where the scaled classes are identical to the input classes (Figure 2) and the output label precision is 100% (pure classes only). Increasing the scaled class-label precision to 50% requires 2-part partitioning of the sample space in each dimension, adding MDGPs at the 50% marks in each dimension. Each grid point then gets a class-label assigned that is composed of class and class proportions.

Assignment of each scaled grid cell to one of the MDGPs requires a decision rule. The decision criterion applied for this algorithm is the percentage similarity or Czekanowski index or coefficient (Czekanowski, 1909), here descriptively called compositional information retention (IR_{cmp}) (Equation 4)

$$IR_{cmp} = \sum_{i=1}^N \min(P_i Smp, P_i MDGP), \quad (4)$$

where P_i = proportion of class i and N = the number of classes in the sample data (Smp) of the scaled grid cell. A scaled grid cell

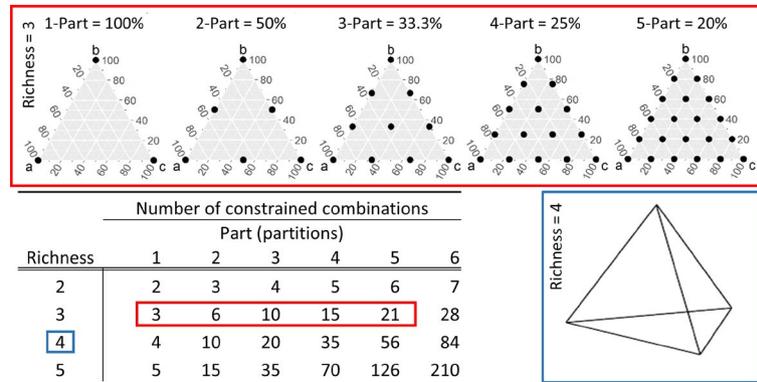


FIGURE 2 Red box: Ternary (2-simplex) plots of three classes a, b and c, for multi-dimensional grid-point solutions of 1–5 parts representing 100%, 50%, 33.3%, 25% and 20% class-label precisions, respectively (left to right). Numbers along the axes are percentages of class presence in each combination (dot). The outer points have one (the apices) or two classes; the inner points, when present, are composed of all three classes in different proportions. The total number of points and the distances between points is the class-label precision, which is determined by the number of partitions (Table). The 1-part precision solutions (first column) are identical to the majority-rule solution with the distance between the three points equal to 100%. Blue box: As richness increases the dimension of the polytope increases. For Richness = 4, grid points are evenly spaced within the 3-simplex (tetrahedron)

is assigned to the MDGP that maximizes IR_{cmp} (minimizes the number of flipped cells). Revisiting the example of a two-class landscape (Figure 1a) for each of the 1- to 5-part solutions, the number of

TABLE 2 Scaling solutions of example 1 (Figure 1) applying 1- to 5-part output class-label precision solutions. Maximized information retention for each class-label set in bold and maximized IR across all class-label precision solutions in bold red. Class-label precision of 25% (4-part solution) maximizes compositional information retention for this grid cell and the MDGP algorithm assigns a label nominally representing 75% 'grass' and 25% 'trees'

Parts	Label list	IR (%)
1 (majority)	G100	73.5
	T100	26.5
2	G100	73.5
	G5_T50	76.5
	T100	26.5
3	G100	73.5
	G67_T33	93.5
	G33_T67	59.5
	T100	26.5
4	G100	73.5
	G75_T25	98.5
	G50_T50	76.5
	G25-T75	51.5
	T100	26.5
5	G100	73.5
	G80_T20	93.5
	G60_T40	86.5
	G40_T60	66.5
	G20-T80	46.5
	T100	26.5

MDGPs increases from two to six classes (Table 2). A grid cell with a composition of 73% 'grass' and 27% 'tree' cover (Figure 1a) is assigned to the majority class grass for the 1-part solution (label precision = 100%), retaining 73.5% compositional information. For the 4-part (label precision = 25%) solution, it is assigned to the MDGP with 75% grass and 25% tree, retaining a maximum of 98.5% of compositional information (Table 2).

The effects of scaled class-label precision and sf on the number of potential and realized grid points and their frequency distributions are demonstrated in Figures 3–5. A landscape with three classes (Figure 3), when scaled with the MDGP-scaling algorithm for class-label precisions of 1-, 3- and 5-parts has a potential of 3, 10 and 21 grid points, respectively. The number of realized grid points for each class-label precision and sf depends on the CAD and $sptAgg$; in Figure 3, these are a geometric CAD with a low

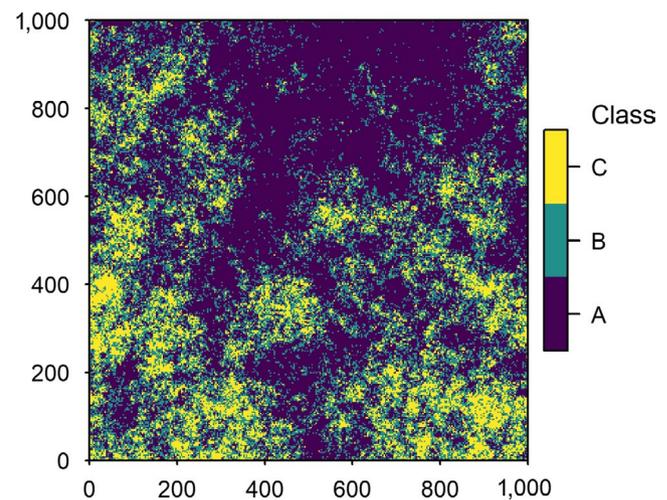
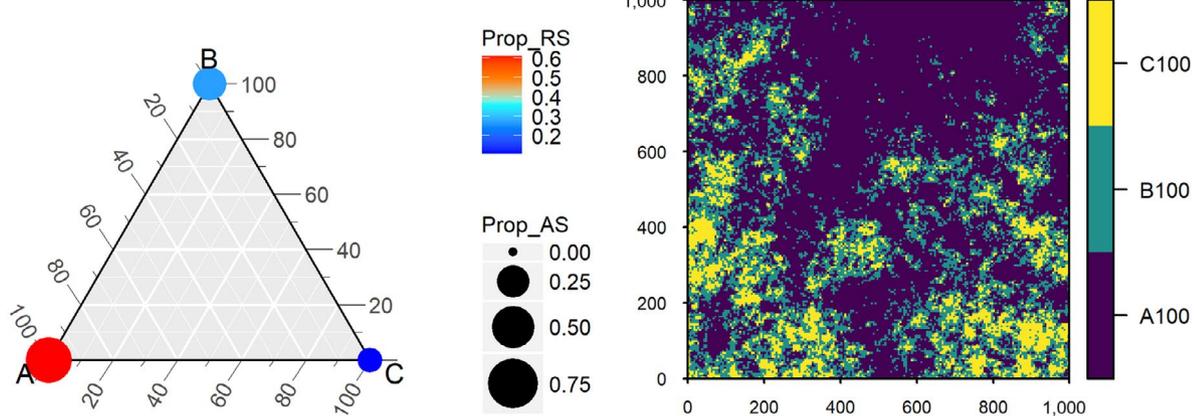
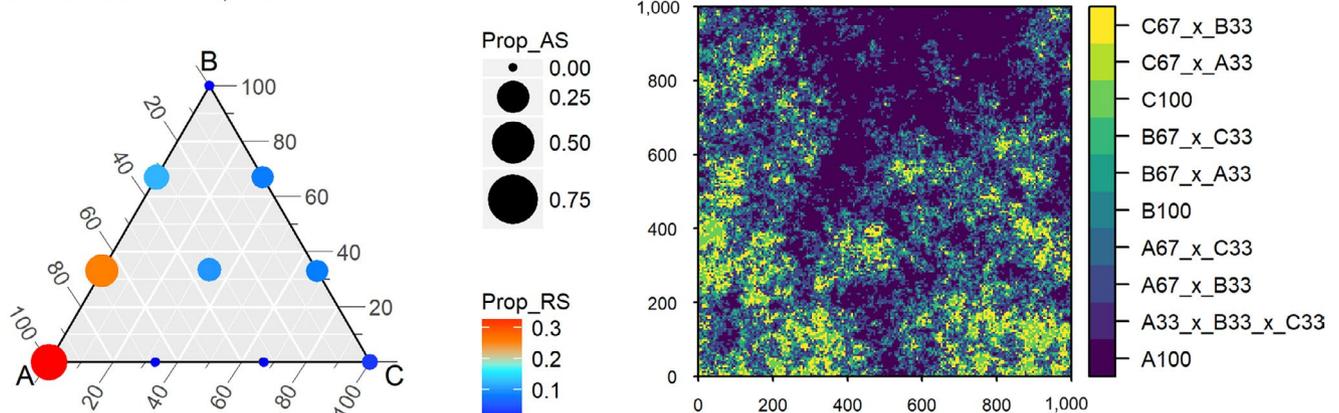


FIGURE 3 Neutral random landscape of three hypothetical classes (A, B, and C). Class abundance distribution is geometric and spatial aggregation low ($h = 0$)

(a) Scale factor = 5; Parts = 1



(b) Scale factor = 5; Parts = 3



(c) Scale factor = 5; Parts = 5

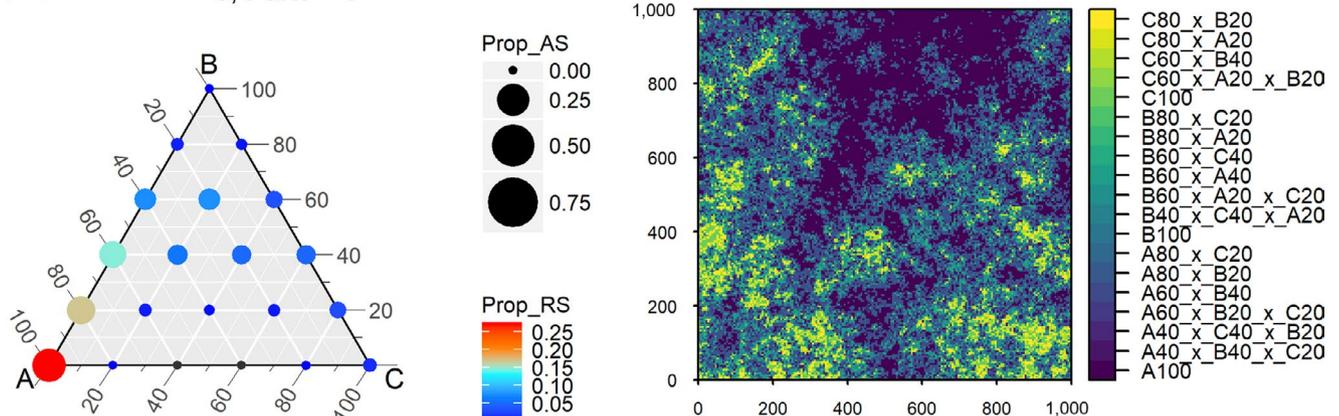
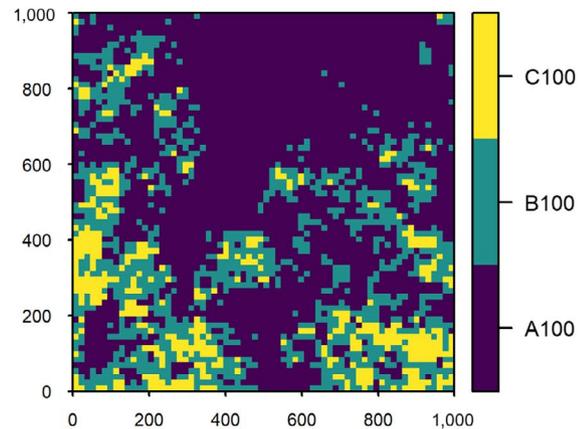
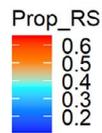
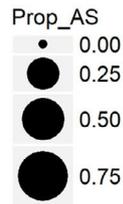
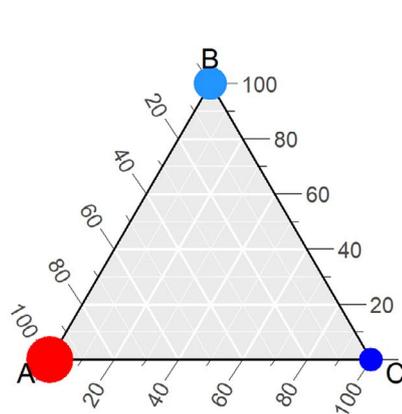


FIGURE 4 Landscape scaling results for landscape in Figure 3, scaled with a scale factor of five. Scaled landscape class frequencies (left) and associated maps (right) for label precisions of (a) 100% (1-part; *majority-rule*), (b) 33.3% (3-part), and (c) 20% (5-part). Circle size displays absolute scale of class proportions (Prop_AS) across all plots; colour rendered as relative scale of class proportions (Prop_RS) within each plot. Black dots indicate that the potential grid point was not realized at the aggregated scale. Class labels in the legend of the map as generated by the MDGP-scaling algorithm. Class labels are composed of class name and nominal percent representativeness

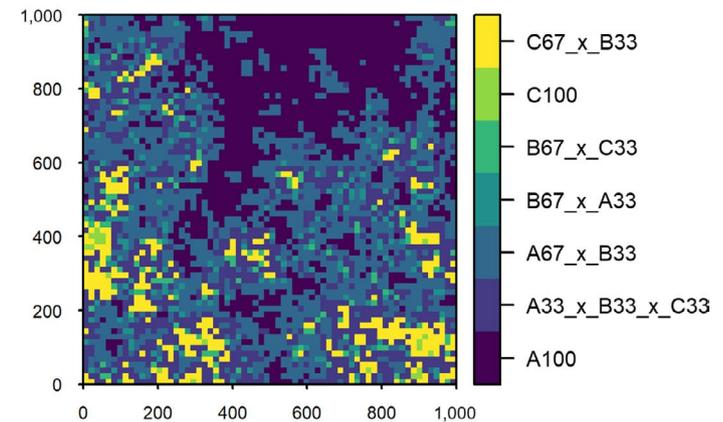
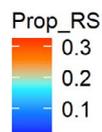
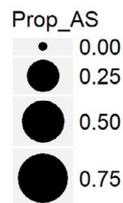
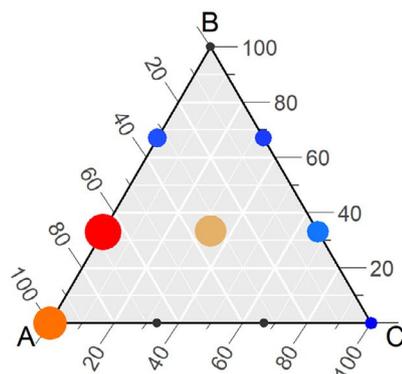
spatial aggregation. As output class-label precision increases to 33.3% and 20%, the number of realized output classes increases from 3 to 10 and 19 classes for a *sf* of 5 (Figure 4) and to 7 and 13 for a *sf* of 15 (Figure 5), respectively. The frequency distributions shift from 100% pure classes to a majority of grid cells assigned to mixed-label classes.

As richness increases, the number of possible scaled classes is still very high (Table 3), but many of the classes are expected at low frequencies across the landscape. Hence, to allow for removal of scaled classes with low proportions across the landscape, a threshold parameter for minimum representativeness was implemented. Output classes that are below the threshold are iteratively removed,

(a) Scale factor = 15; Parts = 1



(b) Scale factor = 15; Parts = 3



(c) Scale factor = 15; Parts = 5

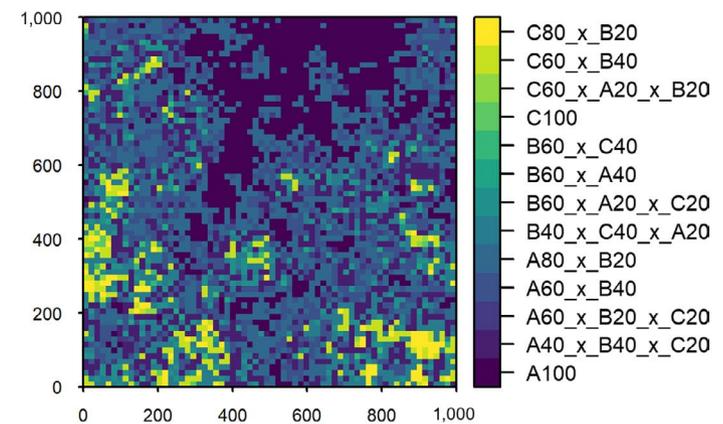
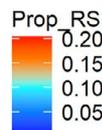
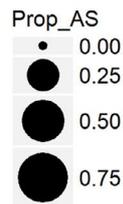
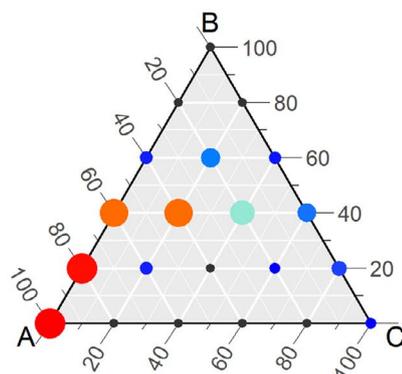


FIGURE 5 Landscape scaling results for landscape in Figure 3, scaled with a scale factor of 15. Scaled landscape class frequencies (left) and associated maps (right) for label precisions of (a) 100% (1-part; *majority rule*), (b) 33.3% (3-part), and (c) 20% (5-part). Circle size displays absolute scale of class proportions (Prop_AS) across all plots; colour rendered as relative scale of class proportions (Prop_RS) within each plot. Black dots indicate that the potential grid point was not realized at the aggregated scale. Class labels in the legend of the map as generated by the MDGP-scaling algorithm

and their assigned landscape units are reassigned to the remaining MDGPs that maximize their IR_{cmp} . The class removal process repeats until no class is below the representativeness threshold. Rare classes that occur in monotypic patches at the aggregation scale, however, might be of ecological significance. Maintaining rare classes even if they fall below the representativeness threshold is achieved with a

threshold parameter for homogeneity, which sets the minimum class percentage in a sample to declare it homogenous or monotypic. The pseudocode for the MDGP-scaling algorithm is presented in Figure 6.

The objective of the following simulation study was to conduct a consistency and sensitivity analysis for the MDGP-scaling algorithm

Richness	Number of constrained combinations					
	Parts (partitions)					
	1	2	3	4	5	6
2	2	3	4	5	6	7
3	3	6	10	15	21	28
4	4	10	20	35	56	84
5	5	15	35	70	126	210
6	6	21	56	126	252	462
7	7	28	84	210	462	924
8	8	36	120	330	792	1,716
9	9	45	165	495	1,287	3,003
10	10	55	220	715	2,002	5,005
Precision	100%	50%	33.33%	25%	20%	16.67%

TABLE 3 Number of constrained combinations and precision limits for equal part partitioning of n dimensions (richness). Precision is 100%/number of partitions

Pseudocode for MDGP-Scaling Algorithm

```

1) Input
   m:= map { raster map }
   o:= origin { random, xy }
   sf:= scale factor { odd integer }
   parts:= label precision parameter { integer }
   rprThr:= representativeness threshold in percent { 0 – 100 }
   hmgThr:= homogeneity threshold in percent { 0 – 100 }
2) Generate multi-dimensional grid point (MDGP) list for parts
3) Generate kernel (k) with dimensions sf x sf
4) Generate all enumerated scaled grid cells (cS) for sf and o
5) For all cS:
   Extract class values from m
   Tabulate relative class abundance in percent (cS_prc)
6) Split cS_prc into homogenetic (cS_hmg) and heterogenetic (cS_htr) samples:
   cS_hmg:= if any class percentage in cS_prc ≥ hmgThr
   cS_htr:= if all class percentages in cS_prc < hmgThr
7) For Each cS in cS_hmg:
   Calculate IR to each MDGP % (Eq. 4)
   Assign MDGP that maximizes IR
   Generate list of final monotypic MDGP classes (MDGP-mono)
8) For Each cS in cS_htr:
   Calculate IR to each MDGP % (Eq. 4)
   Assign MDGP that maximizes IR
9) Join cS_hmg and cS_htr (cS-MDGP)
10) Calculate relative abundance (ra) of all assigned MDGP in cS-MDGP
11) While any (ra(MDGP) in cS-MDGP = 0 OR ra(MDGP) in cS-MDGP < rprThr) AND MDGP
not in MDGP-mono:
   Remove MDGP from list and from all cS in cS-MDGP
   For Each cS with no MDGP in cS-MDGP:
     Calculate IR to each remaining MDGP % (Eq. 4)
     Assign MDGP that maximizes IR
   Calculate ra of all assigned MDGP in cS-MDGP
12) Generate class descriptors (label) for each remaining MDGP based on class
dominance and percent rounded to the nearest integer
13) Assign label to each cS in cS-MDGP

```

FIGURE 6 Pseudocode of the MDGP-scaling algorithm

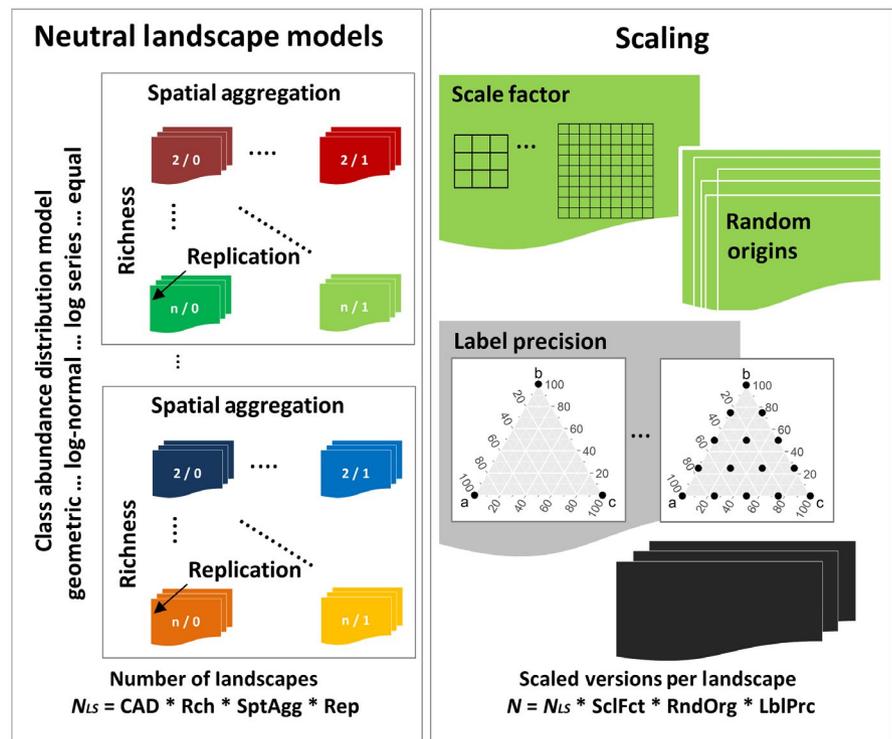
to evaluate the effects of landscape characteristics, scale factor, and scaled class-label precision on information retention and class representativeness for the larger landscape.

2.3 | Test framework

To evaluate consistency and sensitivity of MDGP-scaling, a test framework with a simulation component to generate neutral random

landscapes was implemented. The framework (Figure 7) integrates (a) the generation of neutral landscape models, (b) scaling of the landscapes, and (c) evaluation of the scaling results. Testing the performance of algorithms on replicates of simulated complex landscapes with known properties sets the statistical benchmark for applying them to real landscapes (Fahrig, 1991; With & King, 1997). A full factorial design for three levels of rch (3, 6, and 9 classes), two models of CAD (equal and geometric), and four levels of $sptAgg$ (0, 0.3, 0.6, 1)

FIGURE 7 Schema of framework to test the effects of landscape characteristics, scale factor and class-label precision on information retention, class-count consistency and class-label fidelity in a full factorial design



defined 24 landscape types. Equal CAD (Equation 5), although very unlikely in natural systems, provided the most neutral random landscape type, whereas the geometric CAD (Equation 6) was based on the ecological theory of resource limitation (Motomura, 1932), but any mathematical or statistical model (i.e. log-series, gamma, negative binomial, log-normal) that models the shape of relative abundance distributions (McGill et al., 2007) could be implemented. Class proportions P_c for each class c for equal and geometric CAD were calculated as:

$$P_c = \frac{1}{rch} \quad \text{for equal CAD} \quad (5)$$

$$P_c = \frac{2^{rch-1}}{(2 * 2^{rch-1} - 1) * 2^{c-1}} \quad \text{for geometric CAD} \quad (6)$$

Neutral landscape models were produced for landscapes that resemble distribution patterns that are driven by environmental gradients, one of many pattern types. An algorithm that produces such landscapes is the midpoint-displacement algorithm (Fournier, Fussell, & Carpenter, 1982; Palmer, 1992). The algorithm employed here was the implementation in the Python module 'NLMPY' (Etherington, Holland, & O'Sullivan, 2015). The parameters that determine the landscape pattern are the dimensions of the landscape (number of rows and columns), and a spatial aggregation parameter h that ranges from 0–1 and controls the level of spatial autocorrelation. The resulting array of continuous values was then converted to a raster with categorical data, using the `classifyArray` function in the 'NLMPY' module, where the `weights` parameter was

generated using rch (the number of classes) in combination with either the equal or geometric CAD (Equations 5 and 6). Spatial distribution patterns for rch of three and nine classes, equal and geometric CAD and $sptAgg$ factor h of 0, 0.3 and 1 are demonstrated in Figure 8.

For each of the 24 landscape types with unique characteristics, 10 replicates with $1,000 \times 1,000$ cells were generated, resulting in 240 neutral landscapes with known properties. Spatial-aggregation algorithms were evaluated for sf of 5, 9, 15 and 25. Origin of the scaled grid was randomized five times for each sf and landscape to account for effects of arbitrary origins of the scaled grid.

2.4 | MDGP-scaling algorithm consistency and sensitivity to scaling parameters and landscape characteristics

Efficacy of the MDGP-scaling algorithm to increase IR_{cmp} with the increase in class-label precision, was evaluated on the basis of mean IR_{cmp} of all scaled grid cells across the landscape using pairwise-paired Wilcoxon rank-sign tests (Wilcoxon, 1945). Test p -values were adjusted using the Bonferroni correction for multiple comparisons.

Consistency of the algorithm is crucial to build confidence in its application to real landscapes. Consistency was defined as reproducibility of scaling results across different simulated random landscapes that were congruent in the key characteristics of rch , CAD and $sptAgg$. It was expected that scaled landscapes originating at arbitrary grid origins of the same original landscape and across replicate landscapes with the same properties display low variability in scaling results.

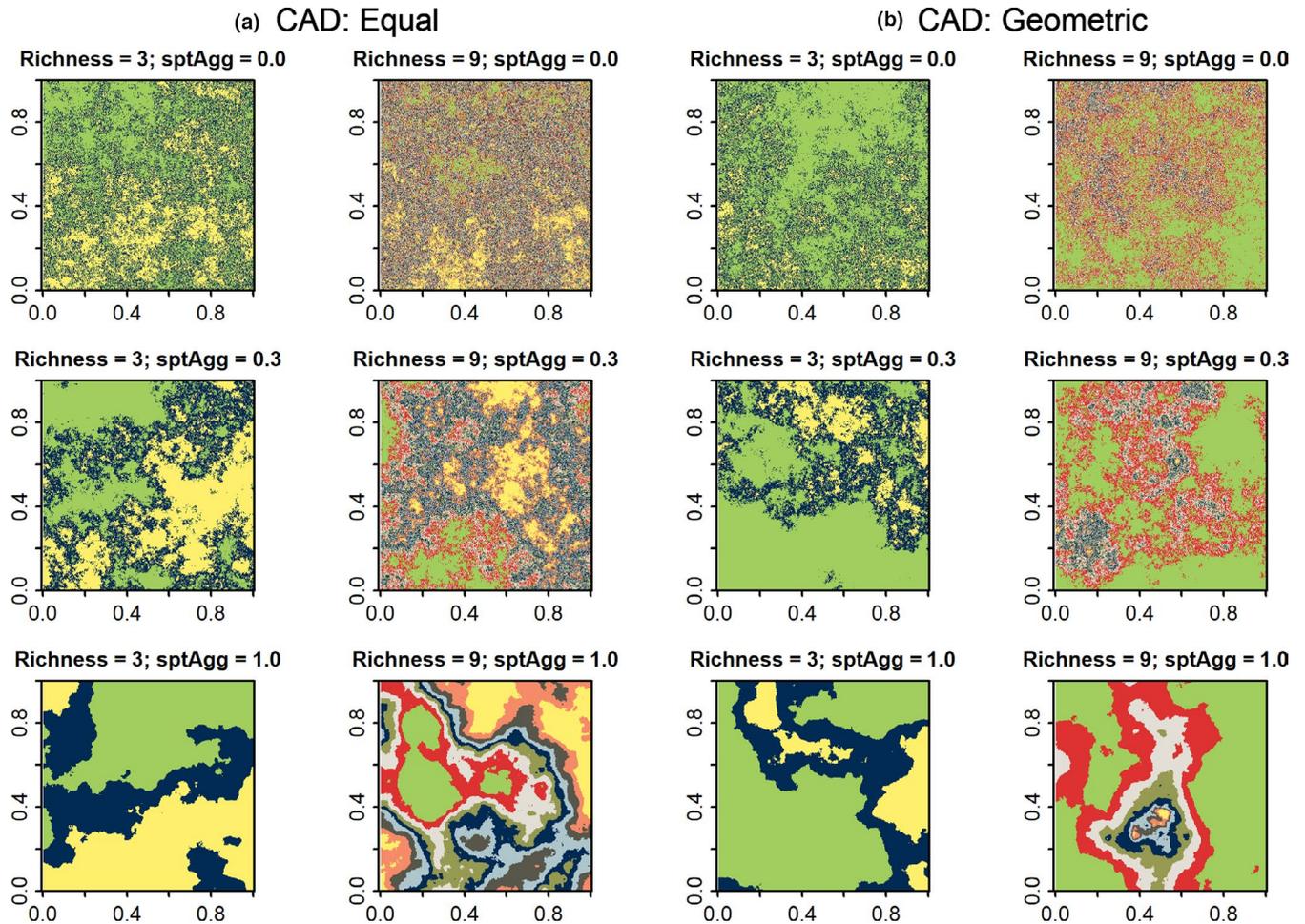


FIGURE 8 Neutral landscapes with (a) equal and (b) geometric CAD of three (left panels) and nine (right panels) classes for $sptAgg$ factor $h = 0.0$ (top), 0.3 (middle), and 1.0 (bottom)

Three indicators that were expected to display low variability are IR_{cmp} , class count (CC) of the scaled classification scheme, and class-label fidelity (CLF), which is the frequency of class-label recurrences across classification schemes generated at random grid origins.

To evaluate consistency of the algorithm when presented with random variation of landscapes with the same characteristics, consistency of mean IR_{cmp} , CC and CLF were evaluated at the landscape type level with coefficients of variability (CV). An algorithm, robust to random variations of the landscape, is expected to display low CV. For IR_{cmp} and CC, CV was calculated across the five random scaling results of each of the ten random landscapes per landscape type. For the five random origins of each landscape, (a) the mean probability of class-label recurrence across all class labels (CLF_{mnPrb}) and (b) the proportion of classes for which recurrence probability was one (CLF_{prp1}) were calculated. High CLF_{mnPrb} and CLF_{prp1} indicate consistent and reproducible classification schemes. Consistency for the CLF parameters was evaluated with the CV calculated across all landscapes of a landscape type and summarized by landscape type characteristics, scale factors and class-label precisions.

Sensitivity of the MDGP-scaling algorithm to scaling parameters and landscape characteristics was assessed with the magnitude of

change for IR_{cmp} and CLF when evaluated by landscape type and sf . With increasing class-label precision, regardless of landscape type and sf , IR_{cmp} was expected to significantly increase, while CLF was expected to decrease. Significance of differences in IR_{cmp} and CLF between class-label precisions was tested with pairwise-paired Wilcoxon rank-sign tests (Wilcoxon, 1945), and p -values were adjusted using the Bonferroni correction.

Consistency and sensitivity for the three indicators were assessed for four sf and four class-label precisions, ranging from 2- to 5-parts. For the simulation study, the representativeness threshold was maintained constant at 1% and class homogeneity at 90%.

The MDGP-scaling algorithm, simulation and test framework, data analysis and visualization were scripted in R (R Core Team, 2013), using packages 'RASTER' (Hijmans & van Etten, 2010), 'RGDAL' (Bivand, Keitt, & Rowlingson, 2013), 'COMPOSITIONS' (van den Boogaart & Tolosana-Delgado, 2008), 'FOREACH' and 'DOPARALLEL' (Revolution Analytics & Weston, 2013). Neutral landscape generation and scaled data aggregation for random landscape origins for the different scale factors were scripted in Python 2.7 (Python Software Foundation) utilizing the Python module 'NLMPY' (Etherington et al., 2015). All data were processed at the high-performance-computing cluster (HPC)

of the Instructional & Research Computing Center (IRCC) at Florida International University (FIU).

3 | RESULTS

Mean IR_{cmp} was significantly higher for class-label precisions of 50% and greater (pairwise-paired Wilcoxon Rank-Sign tests; Bonferroni adjusted $p < .001$; $N = 50$; Figure 9). Scale factor had a greater effect on IR_{cmp} for *majority-rule* aggregated landscapes than MDGP-scaled landscapes for all landscapes regardless of CAD and spatial aggregation factors greater than zero (Figure 9). Difference in IR_{cmp} between MDGP-scaled and *majority-rule* aggregated landscapes increased

with sf for all landscapes and was much greater for landscapes with low $sptAgg$.

3.1 | Algorithm consistency

Consistency of the algorithm was high for IR_{cmp} , CC and CLF. Landscape-specific coefficient of variation for IR_{cmp} ranged from 0.05% to 5.2%. Consistency of IR_{cmp} was high across all landscape types and scale factors and varied little with scale factor. However, evaluating consistency for individual landscape characteristics showed that CV was almost twice as high for landscapes with equal CAD, and variability increased with richness but decreased with spatial aggregation.

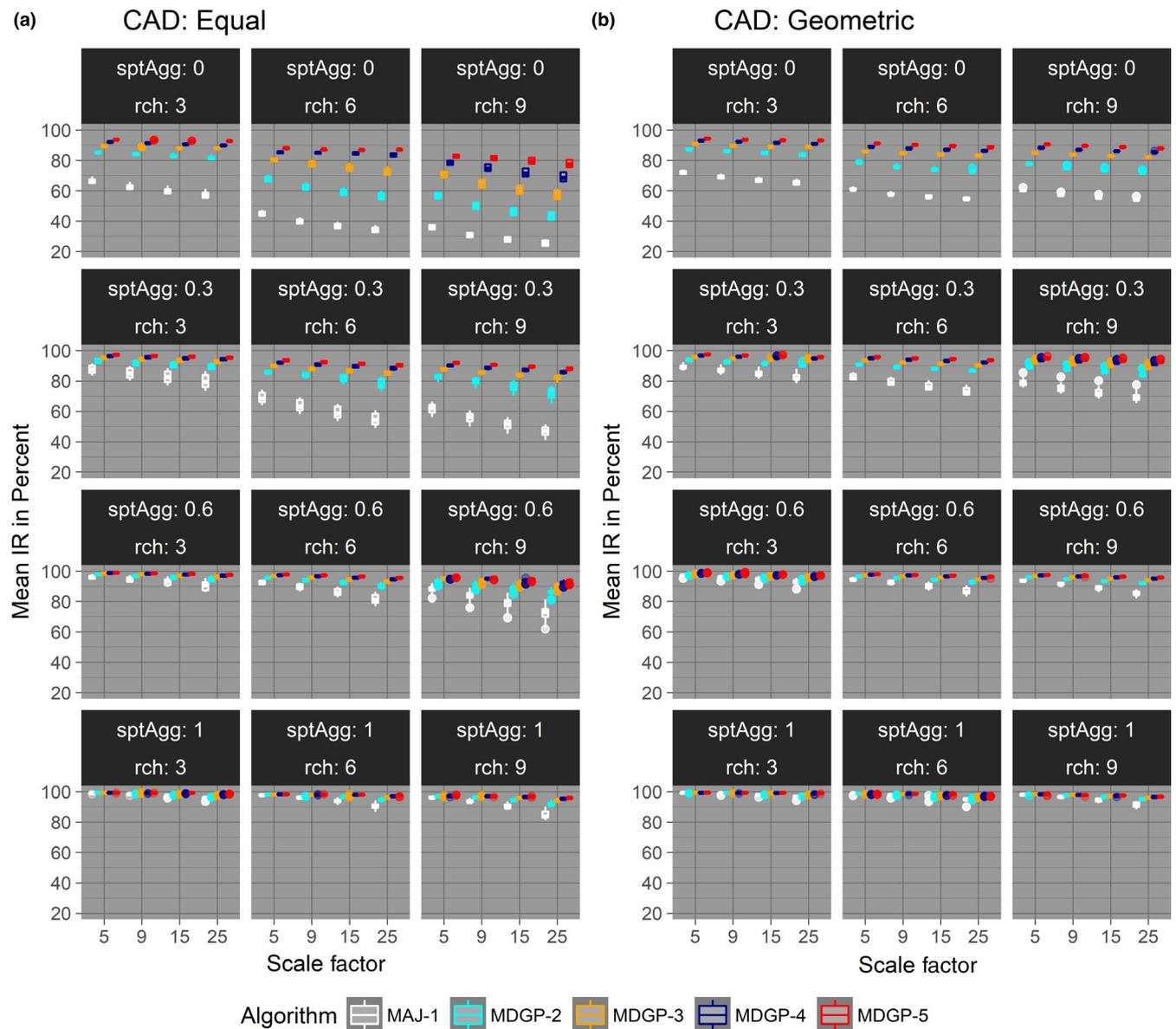


FIGURE 9 Sensitivity of information retention (IR) to scaling parameters for landscapes with (a) equal and (b) geometric class-abundance distribution (CAD). Richness (rch) increases across columns, while spatial aggregation ($sptAgg$) increases down rows. MAJ-1 = *majority-rule* algorithm with 100% class-label precision, MDGP = *multi-dimensional grid-point* scaling algorithm with 2 = 50%, 3 = 33%, 4 = 25% and 5 = 20% class-label precision

Variability in class count on average was below 10% when evaluated by landscape, class-label precision and scale factor (95th percentile = 19.7%). Lowest consistencies (CV > 20%) were observed only for a few landscapes with high spatial aggregation ($h = 1$), and in two instances for low aggregation ($h = 0$) when the scale factor was 25.

Consistency in CLF was high for both parameters, CLF_{mnPrb} and CLF_{prp1} . Mean CV for CLF_{mnPrb} was below 10% for all label precisions (95th percentile = 12.6%). The strongest effect on consistency of CLF_{mnPrb} was observed for spatial aggregation of landscapes. For landscapes with high spatial aggregation ($h = 1$) CV was on average 8.3% (95th percentile = 15.9%), at least twice as high as for all lower aggregation levels. Variability of CLF_{prp1} was much higher than for CLF_{mnPrb} . CLF_{prp1} varied most across spatial aggregation levels as well, but CV reached 18.8% (95th percentile = 34.6%). No effects on consistency of either CLF metric were observed for landscape characteristics richness and CAD, or for scale factor.

3.2 | Algorithm sensitivity

Mean IR_{cmp} increased with increasing class-label precision and decreasing sf for all landscapes (Figure 9). Mean IR_{cmp} was significantly lower (paired Wilcoxon rank-sign test; $p < .001$) for landscapes with equal CAD than for those with geometric CAD, and it increased with $sptAgg$ and decreased with rch (Figure 9). Considering the magnitude of effect, a diminishing return for IR_{cmp} with increasing class-label precision was observed across all landscapes and sf (Figure 9). Largest gains were consistently observed when increasing label precision from 1-part to 2-part solutions (*majority-rule* or 100% to MDGP-2 or 50% precision). Increase in IR_{cmp} with label precision was reduced as $sptAgg$ increased and was lowest for landscapes with aggregation of one. Largest gains in IR_{cmp} with increasing label precision were observed for landscapes with high rch and for high sf (Figure 9).

Class-label fidelity was high for all landscape types, scale factors and class-label precisions, but decreased with increasing class-label precisions (Figure 10). Mean probability of class recurrence for landscapes with equal CAD ranged from 0.99 ± 0.02 for a class-label precision of 50%, decreasing to 0.91 ± 0.1 for a precision of 20%. For landscapes with geometric CAD, a mean probability reduction of 0.02 ± 0.05 was observed when compared to the corresponding equal CAD landscapes (Figure 10). With increasing class-label precision, the greatest losses of CLF were observed for fully aggregated landscapes, regardless of rch and CAD. For landscapes with a geometric CAD and low $sptAgg$, CLF actually increased with class-label precision as rch increased to nine classes (Figure 10).

4 | DISCUSSION

The scale of analysis is crucial when developing ecological models, as results for environmental and ecological processes can vary

significantly when evaluated at different scales. Essential components for reliable interpretation of results are selecting the appropriate analytical scale for the ecological processes modelled and providing data with adequate precision to support the models. The MDGP-scaling algorithm is the first algorithm that generates data-driven, scale-representative classification schemes while conducting spatial data aggregation. The simulation study demonstrated that the algorithm consistently delivers similarly scaled class labels when generating scale-specific classification systems. Representativeness of generalized data is application-specific. When scaling categorical data, two thresholds are of interest: the minimum level of thematic class precision required to maintain enough information to answer the scientific question; and the threshold of minimum relative abundance of a class, below which it is of no ecological interest at the aggregated scale. The minimum level of class precision is the point beyond which location-specific generalization reduces information content to levels where the question of interest can no longer be addressed. To attain a desired precision in the thematic domain, the MDGP-scaling algorithm provides control parameters that allow for IR_{cmp} optimization in the thematic domain that can be tuned with respect to ecological significance for subsequent modeling. Spatially explicit and exhaustive layers of compositional information retention that are provided by the MDGP-scaling algorithm provides valuable input for ecological models that consider the spatially explicit propagation of uncertainty and error.

Gains in IR_{cmp} with increasing class-label precision followed the law of diminishing returns. Richness in scaled classification systems increased while CLF diminished, which complicates optimization of the precision parameter. Decreasing class-label precision in several instances reduced IR_{cmp} marginally while significantly enhancing CLF and reducing class count, producing a more general classification scheme. An increase in class-label precision did not always increase CC or reduce CLF, which indicates that the class-label precision parameter needs to be optimized for individual landscapes and scale factors.

4.1 | Ecological applications

Spatially explicit models of landscape dynamics have their advantages over spatially implicit models (DeAngelis & Yurek, 2017), but they require the detection of spatially explicit change at adequate spatial and temporal resolutions. Detection of changes in land cover and ecosystem properties are common remote sensing applications. Interpretation of changes in spectral-reflectance patterns, as they relate to biophysical parameters of the land surface or as changes in land cover depends on the accurate identification of land cover at the spatial, temporal and thematic precision at which changes are modeled. Landscapes that display high spatial heterogeneity complicate retrieval of biophysical parameters using remotely sensed data (Jacob & Weiss, 2014; Liu, Hiyama, Kimura, & Yamaguchi, 2006; Lu, 2006). For instance, Leaf Area Index (LAI), and Fraction of Photosynthetically Active Radiation (FPAR) are two

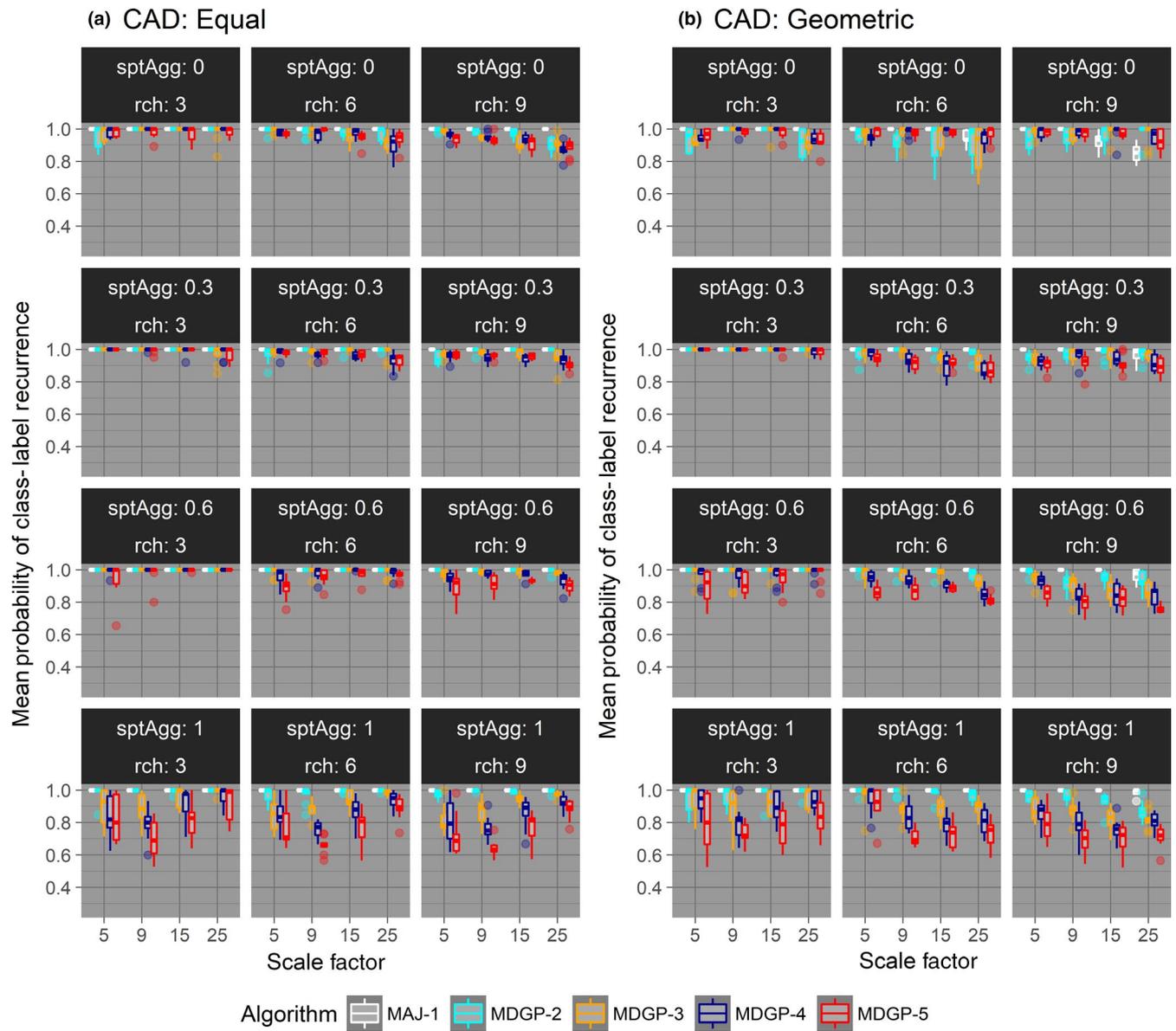


FIGURE 10 Sensitivity of class-label fidelity evaluated across all landscape iterations with the same characteristics for five random origins ($N = 50$) for landscapes with (a) equal and (b) geometric class-abundance distribution (CAD). Richness (rch) increases across columns, while spatial aggregation ($sptAgg$) increases down rows. MAJ-1 = majority-rule algorithm with 100% class-label precision, MDGP = multi-dimensional grid-point scaling algorithm with 2 = 50%, 3 = 33%, 4 = 25% and 5 = 20% class-label precision

important biophysical variables in ecosystem productivity models that rely on prior knowledge of land-cover information (Ganguly et al., 2008; Knyazikhin et al., 1999; le Maire, Marsden, Nouvellon, Stape, & Ponzoni, 2012; Steltzer & Welker, 2006; Zhao et al., 2016). Sensitivity of LAI and FPAR to land-cover and high heterogeneity of vegetation types within a pixel affects LAI estimates in a nonlinear fashion (Garrigues, Allard, Baret, & Weiss, 2006; Lotsch, Tian, Friedl, & Myneni, 2003), and LAI estimate errors at coarse resolution are inversely related to the proportion of the dominant land cover in a pixel (Tian et al., 2002). Consequently, scaling of land cover maps that maintains more precise plant community information reduces error and uncertainty of biophysical parameter estimates from moderate-resolution remotely sensed data.

Another application that requires scaling of land-cover information is modeling land-cover change across long temporal extents. Since the early 2000s, availability of multi-spectral datasets with high spatial resolution has increased. Modeling spatially explicit and exhaustive changes in the past, however, requires resorting to data with lower spatial resolution. Combining categorical land-cover maps derived at different scales requires high spatial resolution products to be scaled to the lower resolution reconciling scaled differences of classification systems. A high priority in this case is to determine class-label precisions at which the low-resolution sensor can spectrally differentiate the most common co-occurrence patterns of mixed classes. Applying the MDGP-scaling algorithm can assist in the optimal class-label precision selection for a variety of

sensors for which larger spatial extents or longer time-series data are available. In a test study, the MDGP-scaling method was applied to scale 2 m resolution plant community maps derived from WorldView-2 data to the 30 m Landsat resolution (Gann, 2018a). Scaling the original map with a class-label precision of 33% increased IR_{cmp} by 15%, while also increasing class-detection accuracy by 5.2% when compared to *majority-rule* aggregation. Classification accuracy increased because the mixed classes had more refined class definitions (labels) that translated into more specific multi-spectral reflectance patterns. How scaled maps with higher class-label precision increase accuracy and precision of ecological modeling still needs to be evaluated.

ACKNOWLEDGEMENTS

I am grateful to Jennifer Richards for providing valuable feedback as the research developed. I also thank Cassian D'Cunha for providing technical support with the high-performance data processing.

DATA AVAILABILITY STATEMENT

Python and R scripts with documentation are available at <https://doi.org/10.5281/zenodo.3406889> (Gann, 2019). The datasets of neutral landscape models and the scaling results for random origins that were generated and used in the simulation study are available at <https://doi.org/10.5281/zenodo.1400061> (Gann, 2018b).

ORCID

Daniel Gann  <https://orcid.org/0000-0001-6131-3391>

REFERENCES

- Aitchison, J. (1986). *The statistical analysis of compositional data. Monographs on statistics and applied probability*. London: Chapman and Hall. <https://doi.org/10.1007/978-94-009-4109-0>
- Aitchison, J., & Egozcue, J. J. (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, 37(7), 829–850. <https://doi.org/10.1007/s11004-005-7383-7>
- Bivand, R., Keitt, T., & Rowlingson, B. (2013). *RGDAL: Bindings for the geo-spatial data abstraction library*. Retrieved from <http://cran.r-project.org/package=rgdal>.
- Braun-Blanquet, J. (1932). *Plant sociology—The study of plant communities* (1st ed.). London, UK: McGraw-Hill Book Company.
- Coulston, J. W., Zaccarelli, N., Riitters, K. H., Koch, F. H., & Zurlini, G. (2014). The spatial scan statistic: A new method for spatial aggregation of categorical raster maps. *Quantitative Methods*. ArXiv E-Prints, 1–14. Retrieved from <http://arxiv.org/abs/1408.0164>.
- Czekanowski, J. (1909). Zur differential Diagnose der Neandertalgruppe. *Korrespondenzblatt der deutschen Gesellschaft für Anthropologie. Ethnologie Und Urgeschichte*, 40, 44–47.
- De Cáceres, M., Chytrý, M., Agrillo, E., Attorre, F., Botta-Dukát, Z., Capelo, J., ... Wiser, S. K. (2015). A comparative framework for broad-scale plot-based vegetation classification. *Applied Vegetation Science*, 18(4), 543–560. <https://doi.org/10.1111/avsc.12179>
- De Cáceres, M., Font, X., Vicente, P., & Oliva, F. (2009). Numerical reproduction of traditional classifications and automatic vegetation identification. *Journal of Vegetation Science*, 20(4), 620–628. <https://doi.org/10.1111/j.1654-1103.2009.01081.x>
- De Cáceres, M., & Wiser, S. K. (2012). Towards consistency in vegetation classification. *Journal of Vegetation Science*, 23(2), 387–393. <https://doi.org/10.1111/j.1654-1103.2011.01354.x>
- DeAngelis, D. L., & Yurek, S. (2017). Spatially explicit modeling in ecology: A review. *Ecosystems*, 20(2), 284–300. <https://doi.org/10.1007/s10021-016-0066-z>
- Egozcue, J. J., Pawłowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300. <https://doi.org/10.1023/A:1023818214614>
- Etherington, T. R., Holland, E. P., & O'Sullivan, D. (2015). *NLMPLY: A python software package for the creation of neutral landscape models within a general numerical framework. Methods in Ecology and Evolution*, 6(2), 164–168. <https://doi.org/10.1111/2041-210X.12308>
- Fahrig, L. (1991). Simulating methods for developing general landscape-level hypotheses of single-species dynamics. In M. G. Turner, & R. H. Gardner (Eds.), *Quantitative methods in landscape ecology: The analysis and interpretation of landscape heterogeneity* (vol. 82, pp. 417–442). New York: Springer-Verlag.
- Fournier, A., Fussell, D., & Carpenter, L. (1982). Computer rendering of stochastic models. *Communications of the ACM*, 25(6), 371–384. <https://doi.org/10.1145/358523.358553>
- Ganguly, S., Schull, M., Samanta, A., Shabanov, N., Milesi, C., Nemani, R., ... Myneni, R. (2008). Generating vegetation leaf area index earth system data record from multiple sensors. Part 1: Theory. *Remote Sensing of Environment*, 112(12), 4333–4343. <https://doi.org/10.1016/j.rse.2008.07.014>
- Gann, D. (2018a). Quantitative spatial upscaling of categorical data in the context of landscape ecology: A new scaling algorithm. Florida International University. Retrieved from <https://digitalcommons.fiu.edu/etd/3641/>.
- Gann, D. (2018b). Simulated Neutral Landscape Models and Scaling Results for Testing of Multi-Dimensional Grid-Point Scaling Algorithm (Version 1.0.0) [Data set]. *Zenodo*. <http://doi.org/10.5281/zenodo.1400061>
- Gann, D. (2019). Multi-Dimensional Grid Point Scaling Algorithm - Framework and Algorithm Evaluation (Version v1.0.0). *Zenodo*. <http://doi.org/10.5281/zenodo.3406889>
- Gann, D., Richards, J. H., & Biswas, H. (2012). *Determine the effectiveness of vegetation classification using WorldView 2 satellite data for the Greater Everglades*. West Palm Beach, FL: South Florida Water Management District. Retrieved from <http://digitalcommons.fiu.edu/gis/25>.
- Garrigues, S., Allard, D., Baret, F., & Weiss, M. (2006). Influence of landscape spatial heterogeneity on the non-linear estimation of leaf area index from moderate spatial resolution remote sensing data. *Remote Sensing of Environment*, 105(4), 286–298. <https://doi.org/10.1016/j.rse.2006.07.013>
- He, H. S., Ventura, S. J., & Mladenoff, D. J. (2002). Effects of spatial aggregation approaches on classified satellite imagery. *International Journal of Geographical Information Science*, 16(1), 93–109. <https://doi.org/10.1080/13658810110075978>
- Hijmans, R. J., & van Etten, J. (2010). *RASTER: Geographic analysis and modeling with raster data. R package version*.
- Jackson, D. A. (1997). Compositional data in community ecology: The paradigm or peril of proportions? *Ecology*, 78(3), 929–940. <https://doi.org/10.2307/2266070>
- Jacob, F., & Weiss, M. (2014). Mapping biophysical variables from solar and thermal infrared remote sensing: Focus on agricultural landscapes with spatial heterogeneity. *IEEE Geoscience and Remote Sensing Letters*, 11(10), 1844–1848. <https://doi.org/10.1109/LGRS.2014.2313592>

- Ju, J., Gopal, S., & Kolaczyk, E. D. (2005). On the choice of spatial and categorical scale in remote sensing land cover classification. *Remote Sensing of Environment*, 96(1), 62–77. <https://doi.org/10.1016/j.rse.2005.01.016>
- Knyazikhin, Y., Glassy, J., Privette, J. L., Tian, Y., Lotsch, A., Zhang, Y., ...Running, S. W. (1999). MODIS Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation Absorbed by Vegetation (FPAR) product (MOD15) algorithm theoretical basis document.
- Lam, N.-S.-N., & Quattrochi, D. A. (1992). On the issues of scale, resolution, and fractal analysis in the mapping sciences. *The Professional Geographer*, 44(1), 88–98. <https://doi.org/10.1111/j.0033-0124.1992.00088.x>
- le Maire, G., Marsden, C., Nouvellon, Y., Stape, J. L., & Ponzone, F. J. (2012). Calibration of a species-specific spectral vegetation index for leaf area index (LAI) monitoring: Example with MODIS reflectance time-series on eucalyptus Plantations. *Remote Sensing*, 4(12), 3766–3780. <https://doi.org/10.3390/rs4123766>
- Liu, Y., Hiyama, T., Kimura, R., & Yamaguchi, Y. (2006). Temporal influences on Landsat-5 Thematic Mapper image in visible band. *International Journal of Remote Sensing*, 27(15), 3183–3201. <https://doi.org/10.1080/01431160600647258>
- Lotsch, A., Tian, Y., Friedl, M. A., & Myneni, R. B. (2003). Land cover mapping in support of LAI and FPAR retrievals from EOS-MODIS and MISR: Classification methods and sensitivities to errors. *International Journal of Remote Sensing*, 24(10), 1997–2016. <https://doi.org/10.1080/01431160210154858>
- Lu, D. (2006). The potential and challenge of remote sensing-based biomass estimation. *International Journal of Remote Sensing*, 27(7), 1297–1328. <https://doi.org/10.1080/01431160500486732>
- Martín-Fernández, J.-A., Barceló-Vidal, C., & Pawłowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3), 253–278. <https://doi.org/10.1023/A:1023866030544>
- Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2014). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2), 134–158. <https://doi.org/10.1177/1471082X14535524>
- Mas, J.-F., Gao, Y., & Pacheco, J. A. N. (2010). Sensitivity of landscape pattern metrics to classification approaches. *Forest Ecology and Management*, 259(7), 1215–1224. <https://doi.org/10.1016/j.foreco.2009.12.016>
- McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., ... White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10(10), 995–1015. <https://doi.org/10.1111/j.1461-0248.2007.01094.x>
- Motomura, I. (1932). On the statistical treatment of communities. *Zoological Magazine*, 44, 379–383.
- Mucina, L. (1997). Classification of vegetation: Past, present and future. *Journal of Vegetation Science*, 8(6), 751–760. <https://doi.org/10.2307/3237019>
- O'Neill, R. V., Hunsaker, C. T., Timmins, S. P., Jackson, B. L., Jones, K. B., Riitters, K. H., & Wickham, J. D. (1996). Scale problems in reporting landscape pattern at the regional scale. *Landscape Ecology*, 11(3), 169–180. <https://doi.org/10.1007/BF02447515>
- Palmer, M. W. (1992). The coexistence of species in fractal landscapes. *The American Naturalist*, 139(2), 375–397. <https://doi.org/10.1086/285332>. Retrieved from <http://www.jstor.org/stable/2462417>.
- R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Revolution Analytics, & Weston, S. (2013). FOREACH: Foreach looping construct for R. Retrieved from <http://cran.r-project.org/package=foreach>.
- Schlup, B. M., & Wagner, H. H. (2008). Effects of study design and analysis on the spatial community structure detected by multiscale ordination. *Journal of Vegetation Science*, 19(5), 621–632. <https://doi.org/10.3170/2008-8-18421>
- Steltzer, H., & Welker, J. M. (2006). Modeling the effect of photosynthetic vegetation properties on the NDVI-LAI relationship. *Ecology*, 87(11), 2765–2772. [https://doi.org/10.1890/0012-9658\(2006\)87\[2765:MTEOPV\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2006)87[2765:MTEOPV]2.0.CO;2)
- Tian, Y., Wang, Y., Zhang, Y., Knyazikhin, Y., Bogaert, J., & Myneni, R. B. (2002). Radiative transfer based scaling of LAI retrieval from reflectance data of different resolutions. *Remote Sensing of the Environment*, 84, 143–159. [https://doi.org/10.1016/S0034-4257\(02\)00102-5](https://doi.org/10.1016/S0034-4257(02)00102-5)
- Tichý, L., Chytrý, M., Hájek, M., Talbot, S. S., & Botta-Dukát, Z. (2010). OptimClass: Using species-to-cluster fidelity to determine the optimal partition in classification of ecological communities. *Journal of Vegetation Science*, 21(2), 287–299. <https://doi.org/10.1111/j.1654-1103.2009.01143.x>
- Tichý, L., Chytrý, M., & Šmarda, P. (2011). Evaluating the stability of the classification of community data. *Ecography*, 34(5), 807–813. <https://doi.org/10.1111/j.1600-0587.2010.06599.x>
- van den Boogaart, K. G., & Tolosana-Delgado, R. (2008). COMPOSITIONS: A unified R package to analyze compositional data. *Computers and Geosciences*, 34(4), 320–338. <https://doi.org/10.1016/j.cageo.2006.11.017>
- Van Der Maarel, E. (1979). Transformation of cover-abundance values in phytosociology and its effects on community similarity. *Plant Ecology*, 39(2), 97–114. <https://doi.org/10.1007/bf00052021>
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>
- Wildi, O. (2010). *Data analysis in vegetation ecology*. West Sussex, UK: Wiley-Blackwell.
- With, K. A., & King, A. W. (1997). The use and misuse of neutral landscape models in ecology. *Oikos*, 79(2), 219–229. <https://doi.org/10.2307/3546007>
- Zhao, J., Wang, Y., Zhang, H., Zhang, Z., Guo, X., Yu, S., & Du, W. (2016). Spatially and temporally continuous LAI datasets based on the mixed pixel decomposition method. *SpringerPlus*, 5(1), 516. <https://doi.org/10.1186/s40064-016-2166-9>

How to cite this article: Gann D. Quantitative spatial upscaling of categorical information: The multi-dimensional grid-point scaling algorithm. *Methods Ecol Evol*. 2019;00:1–15. <https://doi.org/10.1111/2041-210X.13301>