Landscape Ecol
https://doi.org/10.1007/s10980-019-00869-w

RESEARCH ARTICLE

Feasibility of pattern type classification for landscape patterns using the AG-curve

Bjorn-Gustaf J. Brooks · Danny C. Lee

Received: 15 May 2019 / Accepted: 2 July 2019

Abstract
Context Ecological data often contain spatial structures that are latent indicators of ecological processes of interest. The emergence of spatial pattern analysis has advanced ecological studies by identifying spatial autocorrelation and testing its relationship to underlying processes. Spatial point pattern tests such as Ripley's K function were designed for identifying spatial patterns, however they are not without their limitations.
Objectives Recently another graphical technique, AG-curve, was proposed. This paper examines its suitability for classifying disturbance patterns in remote sensing scenery containing tens of thousands of pixels.
Methods To answer the question, *Is there a significant pattern of disturbance or decline present?*, landscapes that were subject to disturbance from mining, wildfire and logging activities were analyzed and compared using the AG-curve technique, which classifies spatial patterns in a window as either random, aggregated, or regular (dispersed). 40 × 40 km windows of NDVI data covering the three prototypical disturbance landscapes and one undisturbed landscape were analyzed for the presence of patterns.
Results From a raster representing the net change in NDVI spanning 18 years, the AG-curve correctly classified the spatial pattern of disturbance in the three disturbance landscapes as a pattern of aggregation among the net-loss in NDVI pixels. In contrast, the undisturbed landscape was classified as random.
Conclusion The AG-curve is a descriptive classification technique useful for identifying spatial patterns in remote sensing imagery and discerning clustered from dispersed patterns. Results highlight that information about the spatial scale of the pattern is also apparent when interpreting the AG-curve graph.

Keywords Pattern analysis · Landscapes · Disturbance · Remote sensing · AG-curve · Hierarchical clustering

Introduction

Ecological data, when examined geographically, often contain structures that exhibit some degree of spatial autocorrelation. These structures are often latent indicators of ecological processes of interest such as epidemic tree death. Consequently, spatial pattern analysis has become an integral part of ecology and a focus of landscape studies. In general, pattern analysis

B.-G. J. Brooks (✉) · D. C. Lee
Eastern Forest Environmental Threat Assessment Center, USDA Forest Service, 200 W.T. Weaver Blvd, Asheville, NC 28804, USA
e-mail: bjorn-gustaf.brooks@usda.gov

Published online: 29 July 2019

tends to either focus on fitting an observed pattern to a theoretical model for significance testing and prediction, or focuses on testing if a pattern differs from a null model produced by random chance. Although many metrics for pattern analysis have been found (see Gustafson 2018; Frazier, this issue), the search for new methods continues. This paper examines the suitability of a new graphical pattern analysis technique, the AG-curve (Takai et al. 2017), as a possible method for identifying the presence of disturbance patterns (or any pattern) in remote sensing scenes containing on the order of 10^5–10^6 pixels.

Traditional statistical tests do not always work well for spatial pattern analyses, because they're often based on assumed independence among the observations. That is, processes at point p_i are assumed not influence processes at its neighboring point p_j. In recent decades spatial statistics have evolved from a contrasting premise, acknowledging that landscape features often give rise to spatial autocorrelation, necessitating new approaches. Spatial tests, such as Ripley's *K* and *L* (Ripley 1977) are used not only to examine the difference between the observed pattern and Complete Spatial Randomness (CSR), but also to compare against a non-random theoretical expectation. Both kinds of alternative hypotheses are represented graphically on the *K*-function graph by an envelope that indicates the range of possible realizations.

Since the introduction of the original *K*-function test (Ripley 1977), various improvements have expanded its capacity (e.g., Besag and Diggle 1977; Diggle 2003) for dealing with more challenging data sets. However, the *K*-function has some disadvantages that are related to its cumulative nature. While very useful for testing statistical significance, *K(t)* at distance t_i contains information at all scales less than and equal to t_i, which complicates interpretation. There are alternative techniques such as the *G*-function (Diggle 2003). However, this also has its own set of limitations. For example, certain kinds of patterns cannot be discriminated using the *G*-function graph because they produce identical curves (See attraction and repulsion patterns in Figs. 4, 5 in Takai et al. 2017).

Recently another graphical technique for classifying spatial patterns, called the AG-curve (agglomerative), was proposed as an approach for bridging this gap. As the name implies, it addresses issues in pattern identification using agglomerative hierarchical clustering. Takai et al. (2017) introduced and tested AG-curve performance using micro- and macroscopic data sets. Given the abundance of modern satellite imagery there are strong incentives to test the applicability of the AG-curve on remote sensing data, not only for analytical efficiency, but also for skill in simultaneously searching across spatial scales for patterns.

The AG-curve methodology

The AG-curve is used to graphically classify a given pattern into one of three categories:

Random: Point patterns that tend to reveal neither attractive nor repulsive processes;
Aggregated: Patterns characterized by clustered points due to attractive processes;
Regular: Patterns that tend to show regularly spaced dispersion among points due to repulsive processes.

The AG-curve combines two analytical steps (the initial test for CSR and subsequent determination of regular or aggregated pattern) into a single graphical test for classifying point patterns.

The steps in developing an AG-curve from a set of coordinate data are as follows (A link to our computer code for replicating these steps is provided at the end of this section.):

(1) From a set of *n* spatial points calculate an *n x n* distance matrix
(2) Hierarchically cluster the distance matrix
(3) From the cluster output, extract dendrogram height *h* at each cluster merge *k*, from $k = n - 1$ to $k = 1$.
(4) Graph the results as *h(k)* by *k*

It is apparent in the steps above that the AG-curve is actually an alternate representation of the cluster dendrogram that extracts the rate of change information embedded in the dendrogram branching. In its most basic form the curve can be expressed as:

$$h(k) \text{ for } k = 1, 2, \ldots, n - 1$$

where n is the total number of spatial points. (Note that the agglomerative clustering process actually proceeds in the opposite direction, $k = n-1, n-2, \ldots, 1$, beginning from those with the smallest distance until the most distant point/cluster is merged.

The graph of the AG-curve includes a second feature, the null model envelope. The envelope is also calculated through steps 1, 2, and 3 above, but is done repeatedly each time with a new set of random points. Each set of points is derived from a simple random sample, also with n points, from the null model domain. After all iterations, the minimum and maximum values of $h(k)$ are used to plot the upper and lower bounds of the envelope at each increment of k. As is discussed below in the examples, the intent of this paper is to explore patterns indicating loss of vegetation, such as disturbance, within the larger landscape. Therefore the AG-curve is based on the subset of n points showing decline, whereas the envelope is constructed from random sets of n points from the complete set of points in the landscape.

Computation of the AG-curve is a memory intensive process. To calculate the AG-curve for a landscape window containing n points, an $n \times n$ distance matrix must be constructed. For example, a 40 km × 40 km window consisting of 250 m × 250 m MODIS pixels contains 25,600 pixels, from which a 655 million ($25{,}600^2$) element distance matrix is constructed for hierarchical clustering. Assuming the distance matrix is single precision, 2.4 gigabytes ($25{,}600^2 \times 4$ bytes) of memory are required just to store the matrix. This can pose a problem for smaller computers, which may have as little as 4 GB of memory. In terms of computational time, the AG-curve of one full 40 km × 40 km window as described above with 655 million pixels took less than 30 s to compute on the standard desktop computer used here. However, computation time for the four prototypical AG-curves took 2 h per case, because each CSR envelope was based on 500 iterations of calculating the AG-curve of n pixels (where n was determined by the number of net-negative NDVI pixels in the window). It is noteworthy that both the number of iterations and the sample size affect the width of the CSR envelope. For example, the CSR envelope based on iterations over the complete set would simply produce a single line, because all iterations would result in the same min and max values. Alternatively iterations over a miniscule sample of say 5% would tend to produce a broad envelope, because sampling variance would likely change dramatically between iterations.

Computer code for calculating the AG-curve is available for the R programming language (R Core Team 2017), and can be downloaded from our repository at https://github.com/LandscapeDynamics/AGcurve/. Alternatively, within R execute "devtools::install_github('LandscapeDynamics/AGcurve')", then "library (AGcurve)", and finally "?AGcurve" to see the code documentation.

AG-curve Interpretation

The AG-curve is the product of clustered spatial coordinates. $h(k)$ represents the distance between the two merged-clusters at that point along the monotonic curve. For example, assuming single-linkage clustering is used, $h(k)$ at any given clustering stage (e.g., $k = n - 1, n - 2, \ldots$) represents the distance between the two nearest points from the two clusters being merged (Fig. 1). One way to view the curve is as a scale that indicates the capacity for resolving spatial patterns as a function of distance. That is, at any scale (distance), $h(k)$, the complete set of points are only resolvable from one another as though there were just k points. As a hypothetical example, the complete set of 100 points, viewed from a distance of 20 meters might only be resolvable as 40 distinct clusters of points.

Figure 1 shows a three-dimensional representation of an example hierarchical clustering of four points. On the left, the locations of the four points are marked by the diamonds. The first cluster merge of the nearest two points, at $k = 3$, occurs over the shortest distance (see cyan colored line). The next nearest points are merged at $k = 2$ over a slightly longer distance. Finally both groups are merged at $k = 1$, by the nearest point in each cluster, across the longest distance. The resulting topology of the AG-curve on the right reveals the rate of change in distance during spatial clustering, and it is this rate that provides novel information about the type of pattern present.

Since the AG-curve is an indication of the rate of change in the clustering of spatial points, a few principles follow. A backwards-L topology, where initially $h(k)$ changes little then rapidly increases

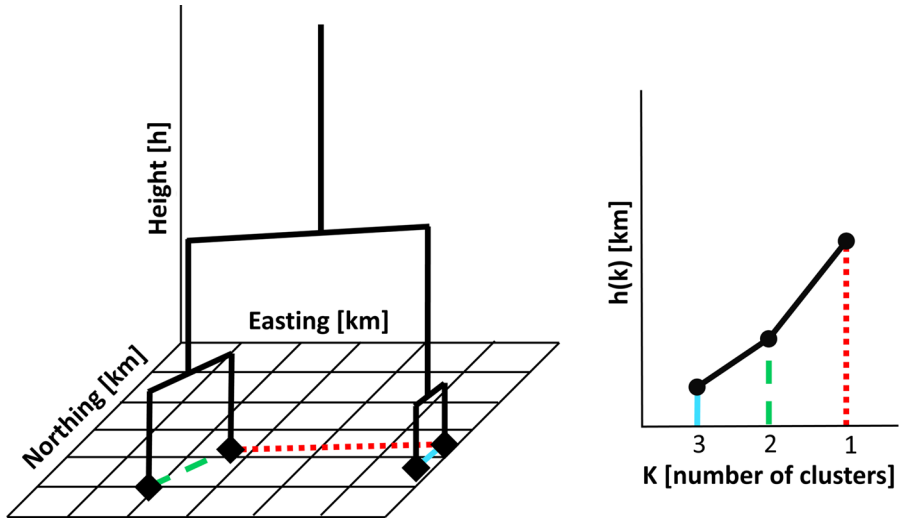

Fig. 1 Left, a simple example of agglomerative hierarchical clustering using single-linkage distance measure (Reproduced after Fig. 1, Takai et al. (2017). Note however that we have reversed the x-axis direction to coincide with the sequence of cluster merging). The colored lines on left, show the distances between nearest points, that correspond to the heights for successive cluster merges on the right

nearing $k = 1$, indicates a pattern of points that tend to be aggregated into groups. *Aggregated* point patterns tend to plot like the lower dotted line in Fig. 2. On the other hand, if the curve lies above the CSR envelope and tends toward a horizontal topology, this indicates underlying dispersal processes that produce *regularly* spaced point patterns (at the scale of the horizontal line). Finally, if the curve falls entirely within the envelope of CSR then the pattern is said to be indiscernible from *randomness*. Thus, what category the point pattern falls into is determined by where the disturbance AG-curve plots relative to the CSR envelope. To illustrate how spatial point patterns translate to AG-curves, Fig. 3 provides a specific example for AG-curve interpretation. Synthetic data were generated for the three pattern categories (aggregated, regular, random) and then categorized using the AG-curve.

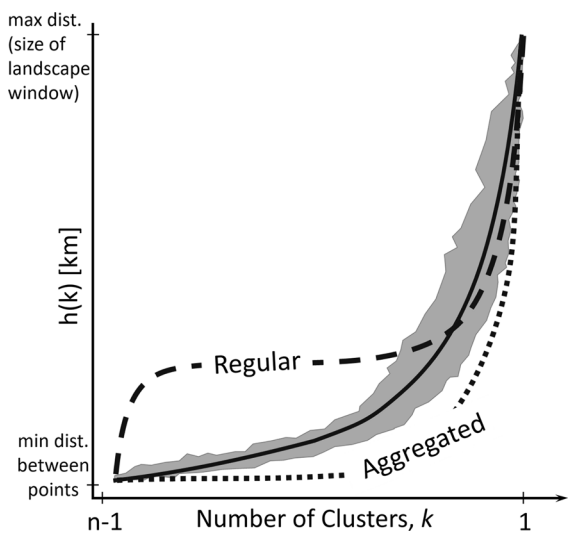

Fig. 2 Diagram of a hypothetical AG-curve showing the curve for three different types of patterns: (1) random point pattern, solid line, (2) aggregated point pattern, dotted lower line and (3) regularly spaced point pattern, dashed upper line (note the x-axis is reversed to coincide with the sequence of cluster merging)

AG-curves of landscape disturbance

To evaluate the suitability of the AG-curve as a preliminary classification method for landscape disturbances observed through satellite imagery, MODIS NDVI (normalized difference vegetation index) imagery (Spruce et al. 2016) for a set of three prototypical disturbances and one undisturbed landscape were analyzed. Each disturbance landscape has an NDVI history reflecting a pronounced loss in vegetation density. Rather than analyzing a scene of NDVI at one point in time, which convolves disturbance patterns with patterns of stable but heterogeneous vegetation communities, an integrated image (raster) representing the total record of 18 years of NDVI change (2000–2017) was analyzed. Net change in NDVI was calculated for each image point (raster pixel), the same as is net change in elevation, as the sum in every NDVI gain minus every loss. The result is an image that is standardized across all locations despite vegetation type, so patterns from landscapes as different as

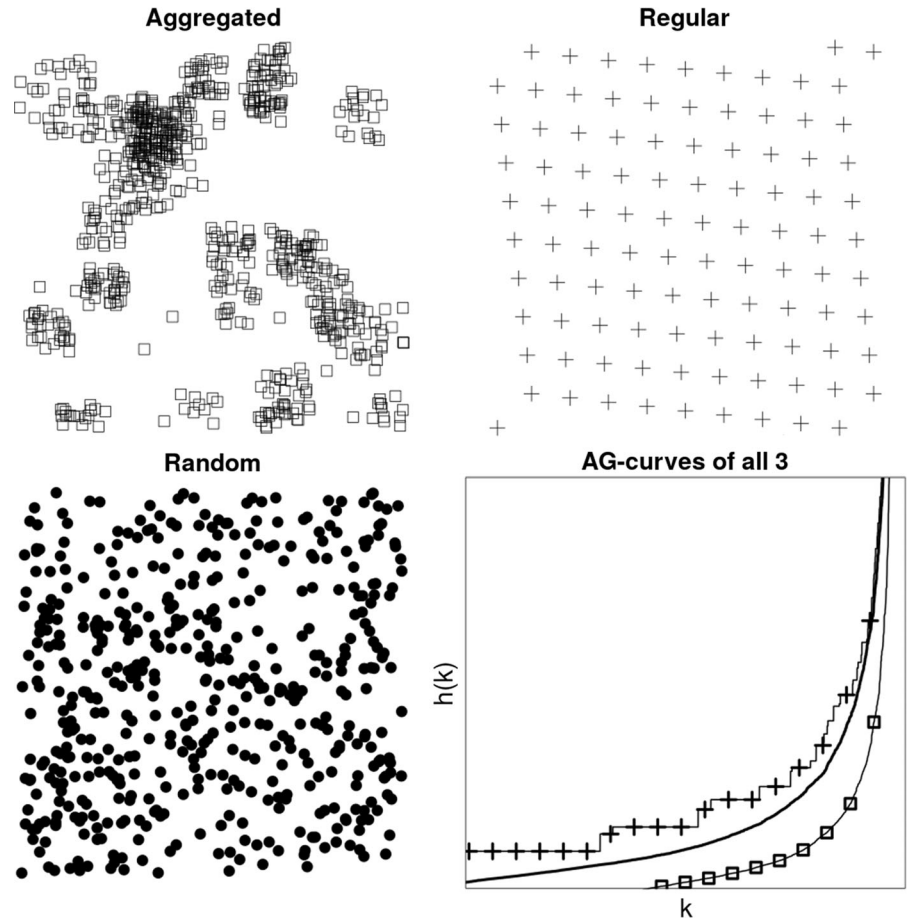

Fig. 3 Aggregated, regular and random patterns are plotted separately in x, y space, along with their corresponding AG-curves in the lower right sub-plot

deserts and swamps can be compared equitably, and based solely on their change in vegetation density.

The four prototypical landscapes are illustrated in Fig. 4. They are as follows: a landscape with multiple timber harvests (2002–2016) in Oregon, the Cascade Complex Wildfire (2007, 2008) in Idaho, and two landscapes in West Virginia including a mountaintop mining site (2008) and an undisturbed mixed-use forest for reference. In creating the AG-curve graph for each landscape, a 40 km × 40 km window was analyzed. The AG-curve of disturbance for each landscape was developed from the subset of n points (pixels) that showed a net-loss in NDVI, which are shaded red in Fig. 4. The CSR envelope, on the other hand, also consisted of n points, but was sampled from all points whether they showed a net-gain or loss in NDVI, or neither. Note that by setting the threshold for disturbance to include all points with negative net-NDVI values, we are also lumping together the most extreme losses in NDVI (e.g., severe wildfire) with points that witnessed only slight losses (e.g., gradual shifts in community composition). As can be seen in Fig. 4, roughly half of the conterminous US landscape seems to fall into the broadly inclusive "disturbance" category. A more nuanced approach, for example, could have filtered for moderate to severe disturbances by setting a more specific threshold for net loss in NDVI.

The disturbance case for logging is in Umpqua National Forest in Southern Oregon. Prior to logging activities, the forest suffered the effects of a mountain pine beetle epidemic, killing large numbers of lodgepole pines. Through the Healthy Forest Restoration Act, a project has allowed for forest thinning and timber sales of various tracts of land in Umpqua N.F. since 2002. The forest gaps, better seen here (http://bit.ly/2Y6NroJ) on the LanDAT map viewer, are characterized by abrupt declines in NDVI that occurred in different years depending on when the timber sale occurred.

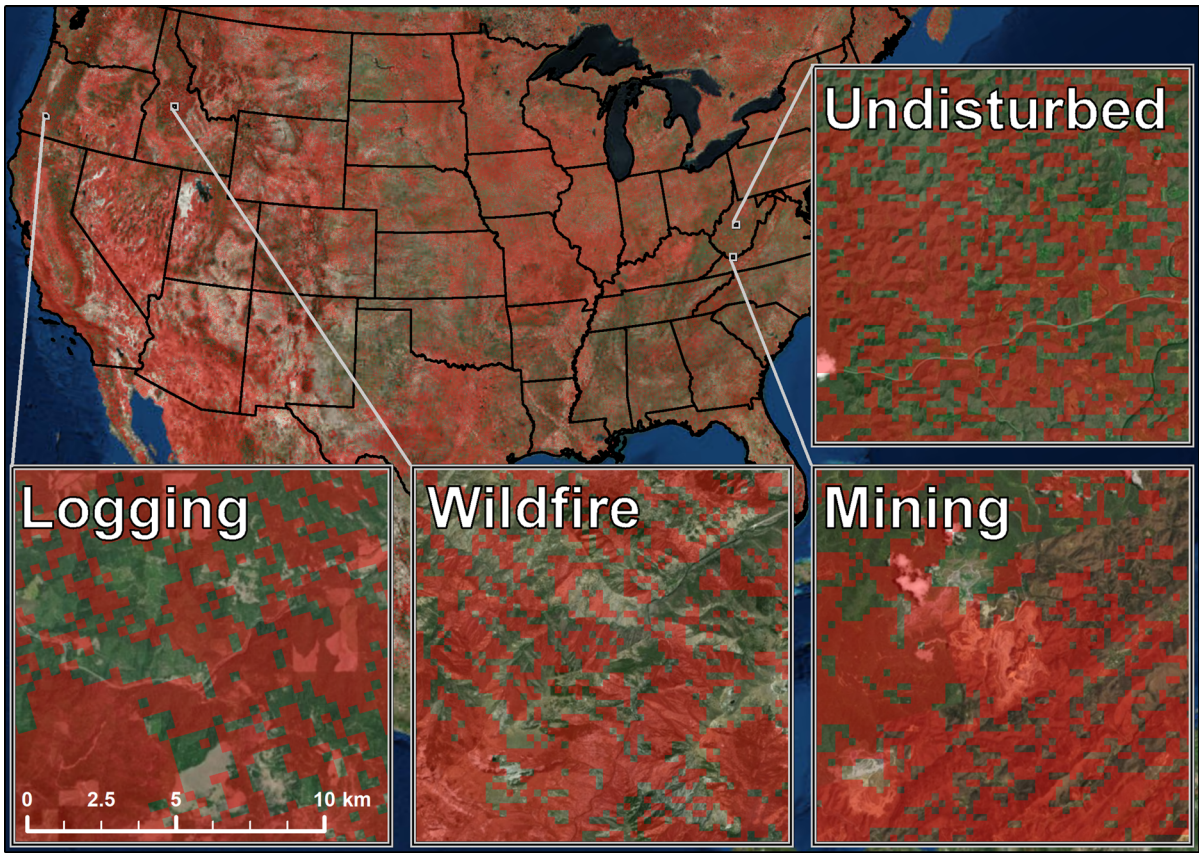

Fig. 4 Visual imagery map showing the locations of the four landscapes. Transparent red overlay are MODIS NDVI data showing points (pixels) that have a net loss in NDVI vegetation density. Note that NDVI-loss is larger than the mine footprint. This highlights that there are factors outside of the mine's footprint that are driving some level of decline there as well. Interestingly, when examined at a continental scale the region with the least NDVI-loss is the southern Rocky Mountains

The wildfire disturbance in Fig. 4 comes from the Cascade Complex Wildfire, which occurred over several years, particularly 2007 and 2008. Massive mortality of lodgepole pine has shifted the landscape phenology signature from evergreen dominant to the current mixture of deciduous seral communities (see time series here, http://bit.ly/2vS5jIb). The spatial pattern of wildfire is spread over a broad area, and has mottled the landscape with patches of differential recovery.

The final prototypical disturbance, the mountaintop removal case, occurred mid-way through the NDVI record (circa 2008–2009). Many of the points within the 40 km × 40 km area reveal an abrupt loss in NDVI with little recovery and even further decline in NDVI (see time series here, http://bit.ly/2YpHsMm). Much of the area outside of the mining footprint shows little appreciable change in NDVI.

For each of the three disturbance cases, the mapped data were translated to AG-curves by first extracting the x,y spatial coordinates of the n points (raster pixels) showing a net loss in NDVI. The coordinates were transformed to a distance matrix, which was then clustered. The dendrogram output $h(k)$ and k were extracted and used to generate AG-curve plots. The CSR envelope that accompanies each curve was formed from repeated simple random samples, also of n points, from the complete set of all points (disturbance and its complement) within the 40 × 40 km area. Note that each subset (disturbance or CSR) was hierarchically clustered using the average point distance metric (unweighted pair group method with arithmetic mean). Thus cluster height, $h(k)$, at each cluster merge, k, represents the average distance between point pairs between the two cluster groups.

Compared to curves presented in the original paper by Takai et al. (2017), the most prominent feature of the AG-curves in Fig. 5 is that the CSR envelope has a surprisingly compact range (Fig. 5). The AG-curve envelopes here differ because the underlying data have different distributions and the sample sizes differ by orders of magnitude. The sample size of points analyzed in Takai et al. was on the order of 10^2, whereas here 10^4–10^5 points were analyzed for each AG-curve. By increasing our sample size, incidentally we decreased the sampling distribution of the sample variance, which is why the CSR envelopes here are more constrained. However, this in no way reduces the reliability of the results presented here, but it does suggest that even small divergences of AG-curves from the CSR envelope do signify important spatial patterns.

The AG-curves of all three prototypical disturbance landscapes in Fig. 5 do show that spatially aggregated patterns of disturbance exist. In further examining the graphs of logging, wildfire and mining in Fig. 5, the divergence of the disturbance curve from the CSR envelope reveals the spatial scales at which these disturbances are significant (*i.e.*, the average distance between cluster merges as mentioned above). Among the disturbance cases, the AG-curve of mining diverges from the CSR envelope at the largest scale near 4.5 km, followed by logging near 3 km and wildfire near 2.5 km. On the other hand, the AG-curve of the undisturbed forested landscape remains completely within the envelope.

The AG-curve is a representation of the *rate* at which points merge in a dendrogram. If the AG-curve and dendrogram are examined side-by-side, intervals where more dendrogram branching occurs will correspond to flatter slopes in the AG-curve. In Fig. 5, all three prototypical disturbances diverge below their envelopes, indicating that points are more aggregated than would be expected under complete spatial randomness. Note that the AG-curves of logging and

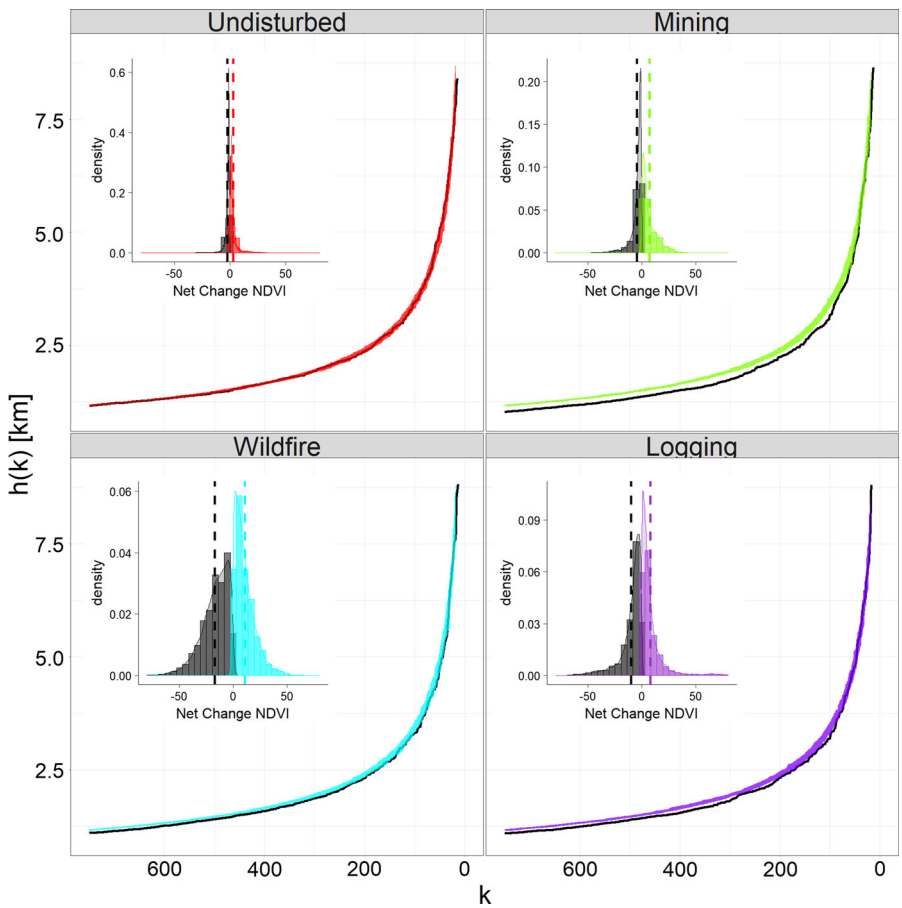

Fig. 5 AG-curves for four example landscapes. The curve of the disturbance points are represented by the black line, and the curve of the complete sets of points (CSR envelope) are represented by the colored bands. (note the x-axis is reversed so as to coincide with the sequence of cluster merging) Each AG-graph has a subplot showing the distribution of the disturbance values in black and their complement, represented by the colored distributions. (The CSR envelopes were sampled across both black and colored distributions) Their means are given by the vertical dashed line in the distribution plots

mining diverge more strongly than does wildfire, which means that their patterns will appear more non-random on a map. This does not mean that the degree of disturbance from wildfire is weak (Note the wide range in NDVI values in the distribution subplot of wildfire, Fig. 5), only that the pattern contrast between disturbed versus non-disturbed (or regrowth) is less pronounced. This points to a wildfire disturbance pattern that is more scattered, and more random, than that of logging or mining. This comes as no surprise, as the pattern of red (NDVI loss) in Fig. 4 is visibly more scattered than the other two disturbance cases. Following wildfire it is not uncommon for recovery to be rather heterogeneous across space, at least by comparison to the checkerboard pattern of clearcuts from logging and mining.

Summary

Pattern identification in satellite imagery can be a difficult task. Chief among the difficulties is the sheer volume of data to search. Even at 250 m MODIS resolution there are more than 146 million pixels covering the Conterminous US. Our intent is to improve capacities to monitor landscapes by filtering windows that do not contain spatial patterns. In this work we evaluated the suitability of the AG-curve as a technique for categorizing whether a window of remote sensing pixels: contains patterns that do not differ significantly from a null model (e.g., complete spatial randomness), contains patterns of pixels exhibiting clustering, or contains patterns of pixels that are non randomly dispersed. The example landscapes of mining, wildfire and logging serve as prototypical cases of disturbances that we expect any pattern analysis technique to correctly classify.

In this paper each AG-curve graph includes a CSR envelope representing complete spatial randomness. Unlike the AG-curves in the original paper by Takai et al. (2017), the CSR envelopes for these data are surprisingly compact (Fig. 5). This is a consequence of the large number of pixels sampled in each spatial window ($\sim 10^4$ to 10^5), which is inversely related to the sampling distribution of the sample variances. Nonetheless, the curves of the prototypical disturbances are still distinguishable enough from CSR to indicate the presence and scale of patterns.

The AG-curve is a descriptive statistical test that is easier to apply to remote sensing imagery than spatial pattern analyses approaches that include inferential test statistics, such as a p value. We do not mean to imply that statistical significance testing of patterns can be bypassed (sensu Amrhein et al. 2019; Wasserstein et al. 2019). Rather we point out that the AG-curve can be a powerful tool for surveying remote sensing imagery, and that it also could be paired with a test statistic if statistical inference is necessary. To demonstrate this we used a raster representing the net change in NDVI over 18 years, and searched for patterns within the negative pixels, comparing them to CSR envelopes sampled from the complete set of negative and positive net-change pixels. This targeted the question: *Is there a significant pattern of disturbance or decline present?* We could have further constrained the threshold to sample a specific type of disturbance or even searched a different input raster that instead separated pixels that experienced an abrupt loss in NDVI (e.g., extreme disturbance). While both of these alternatives would have improved the specificity in identifying disturbances, for this paper, we only intended to describe a minimal test case given a roughly equally divided input data set into loss and gain subsets.

As a method for detecting coherent disturbance patterns, such as clearcut logging, the AG-curve was able to correctly identify spatial patterns of disturbance. If the appropriate input data is analyzed, the AG-curve could be useful in binary classification of windows for presence/absence of spatial patterns (and clustering/dispersal in patterns). Windows with significant patterns could then be flagged for later inspection to classify the exact kind of pattern (e.g., tornado, wildfire, mining, urbanization). The tests in this work say nothing about the confidence in identification of the AG-curve technique for landscape pattern classification (nor do we test for error of commission/omission), but they do offer a window through which one can assess landscape shifts.

Acknowledgements We would like to thank the anonymous reviewers whose comments significantly improved the quality of this paper. This research was supported in part by an appointment to the United States Forest Service (USFS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA). ORISE is managed by ORAU under DOE Contract Number DE-SC0014664. All opinions expressed in this paper are the

author's and do not necessarily reflect the policies and views of USDA, DOE, or ORAU/ORISE.

References

Amrhein V, Greenland S, McShane B (2019) Scientists rise up against statistical significance. Nature 567(7748):305–307.

Besag J, Diggle P (1977) Simple Monte Carlo tests for spatial pattern. J R Stat Soc Ser C (Appl Stat) 26(3):327–333.

Diggle PJ (2003) Statistical analysis of spatial point pattern, 2nd edn. Academic Press, New York

Frazier Emerging trajectories for spatial pattern analysis in landscape ecology. Landscape Ecol.

Gustafson EJ (2018) How has the state-of-the-art for quantification of landscape pattern advanced in the twenty-first century? Landscape Ecol. https://doi.org/10.1007/s10980-018-0709-x

R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ripley BD (1977) Modelling spatial patterns. J R Stat Soc Ser B (Methodol) 39(2):172–212

Spruce JP, Gasser GE, Hargrove WW (2016) MODIS NDVI data, smoothed and gap-filled, for the conterminous US: 2000–2015. ORNL DAAC, Oak Ridge, Tennessee, USA. https://dx.doi.org/10.3334/ORNLDAAC/1299. Accessed 01 March 2019

Takai T, Tamura Y, Motoyama H (2017) A new graphical approach to classify spatial point patterns based on hierarchical cluster analysis. J Jpn Soc Comput Stat 30(1):1–14.

Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond "$p < 0.05$". Am Stat 79:1–10.