International Journal of Geospatial and Environmental Research

Volume 8 | Number 3 Article 4

November 2021

Characterizing Clustering Models of High-dimensional Remotely Sensed Data Using Subsampled Field-subfield Spatial Cross-validated Random Forests

Andrew B. Whetten
University of Wisconsin-Milwaukee, whettenandrew@gmail.com

Characterizing Clustering Models of High-dimensional Remotely Sensed Data Using Subsampled Field-subfield Spatial Cross-validated Random Forests

Abstract

Clustering models are regularly used to construct meaningful groups of observations within complex datasets, and they are an exceptional tool for spatial exploratory analysis. The clusters detected in a recent spatio-temporal cluster analysis of leaf area index (LAI) in the Columbia River Basin (CRB) require further investigation since they are only derived using a single greenness metric. It is of great interest to further understand how greening indices can be used to determine separation of sites across an array of remotely sensed environmental attributes. In this prior work, there are highly localized minority clusters that were detected to be most dissimilar from the remaining clusters as determined by annual variation in remotely sensed LAI. *The objective of this study is to discern what other environmental factors are important predictors of cluster allocation from the mentioned cluster analysis, and secondarily, to construct a predictive model that prioritizes minority clusters.* A random forest classification is considered to examine the importance of various site attributes in predicting cluster allocation. To satisfy these objectives, I propose an application-specific process that integrates spatial sub-sampling and cross-validation to improve the interpretability and utility of random forests for spatially autocorrelated, highly-localized, and unbalanced class-size response variables. The final random forest model identifies that the cluster allocation, using only LAI, separates sites significantly across many other environmental attributes, and further that elevation, slope, and water storage potential are the most important predictors of cluster allocation. Most importantly, the class errors rates for the clusters that are most dissimilar, as detected by the cluster model, have the best misclassification rates which fulfills the secondary objective of aligning the priorities of a predictive model with a prior cluster model.

Keywords

Random Forests, Cross-validation, Cluster Analysis, Leaf Area Index, Remote Sensing Data

Acknowledgements

I would like to thank Zengwang Xu (UWM Department of Geography) for his support over the semester of Spring 2021 as I worked on this project as well as Hannah Demler for her continued collaboration on these projects.

1.Introduction: The need for further validation of a clustering model

The general objective of cluster analysis models is to group observations as determined by some measure of dissimilarity across a collection of explanatory variables. In standard models, such as k-means or k-mediods cluster analysis, the optimization of clusters is done by allocating observations to the nearest measure of center for a cluster, which is then adjusted recursively until the optimization criteria is satisfied, namely that within-group variance is minimized while across group variance is maximized (Hastie et al. 2009). If a homogenous group of observations has dramatic separation from the remaining observations, as measured by the input explanatory variables, then it follows intuition to expect a stable cluster to form even for a low number of clusters. It would not be alarming for this cluster to remain as such with minimal perturbations as the number of clusters permitted in the model increases.

The k-mediod functional cluster analysis performed in the paper, *Detection of Multidecadal Changes in Vegetation Dynamics and Association with Intra-annual Climate Variability in the Columbia River Basin*, is characterized in this way (Whetten Demler 2021). As demonstrated in the supplementary applet (refer to S1 Applet Access), a small but substantial coastal evergreen cluster of sites is separated from the remaining sites with k=2 clusters, and as the number of sites increases, this cluster remains intact. In the final cluster model (with k=5), this is Cluster 4 labeled in purple in Figure 1.

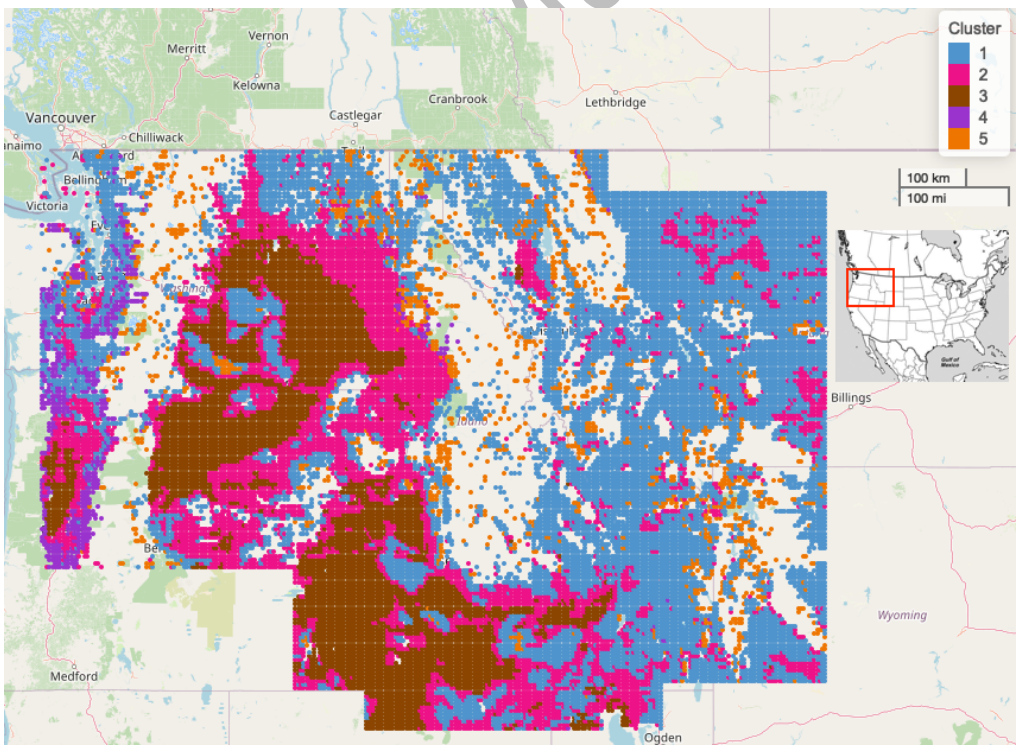

Figure 1. The Clustering Model. K-mediod cluster analysis of the pairwise correlation matrix of the 27191 B-spline smoothed LAI profiles (Whetten Demler 2021, Cheng et. al. 2019)

The Columbia River Basin (CRB) is located in the north-western United States and south-western British Columbia, Canada. The drainage basin is bounded by the Rocky Mountains to the east and the Cascade and Coast ranges to the west and covers an area of 670,000 km : 568,000 km of which are spread across the US states of Washington, Oregon, Idaho, Montana, Wyoming, Utah, and Nevada. Climate in the CRB varies from humid and maritime along the western parts of the basin to semi-arid and arid in the southeast. The CRB hosts a range of diverse natural ecosystems as well as large agricultural regions consisting largely of forestry, dairy and cattle farming, and production of apples, potatoes, wheat, and other small grains (USGS River Basins of the United States Columbia Report). The diversity of this region prompted an exploration of regions that have similar annual trends in greening with the objective of detecting multidecadal shifts in plant phenology in these regions. Many studies of environmental change have been performed using greenness indices although there are few pertaining to the assessment of regional level phenological and atmospheric (such as temperature and precipitation) shifts in the CRB (Berner et. al. 2020; Tawatchai et. al. 2017; Queen et. al. 2021; Hopkinson et.al 2020; Hamlet et. al. 2013; Knowles et. al. 2006).

Phenology refers to periodic and seasonal reproductive events in biological life cycles. Vegetative phenological phenomena are sensitive to annual climate conditions and therefore changes in phenology, such as the timing, rate, duration, and magnitude of annual vegetative growth, can signal important effects of climate change on plants (Piao 2019). Leaf Area Index (LAI), a widely utilized measure of plant growth and activity, is a unit-less measurement of leaf area (m) per ground area (m). LAI provides a key measure of plant cover in a given area and is defined as an essential climate variable (ECV) by the Global Climate Observing System (GCOS) due to its critical contribution to the characterization of Earth's climate (Bojinski 2014). Satellite-derived LAI products offer multidecadal records of terrestrial plant cover around the world, allowing for analysis of inter-annual variability in vegetation dynamics which provides key insight to how plants respond to global change.

This clustering model, used to investigate changes in phenology in the CRB, is derived solely from spline smoothed multidecadal NOAA AVHRR Leaf-Area Index (LAI) curves, a single spatio-temporal attribute measured across the 27,191 sites, and, as opposed to the traditional approach of constructing a dissimilarity matrix of sites across an array of variables, the vast number of replications at each site is advantageously used to measure dissimilarity using pairwise temporal correlations of LAI between sites. As illustrated in the supplementary Figure S2, Cluster 4 is distinguished from the other clusters across all of these attributes. More specifically these sites experience “double-peaks” in LAI in the Spring and Fall seasons, and they have smaller annual variation in temperature and higher cumulative precipitation. Since sites that are spatially closer to each other tend to have stronger correlations in LAI, the cluster model indirectly retains some information about spatial proximity of sites. Figure 1 reveals that the Cluster 2 and Cluster 3 follow major river ways in the CRB, dominantly the Snake and Columbia Rivers, and since Cluster 1 and Cluster 5 comprise of many sites in the mountainous regions of Eastern Idaho and the Western regions of Montana and Wyoming, it was a preliminary hypothesis that elevation may implicitly play a large role in the separation of sites. Most of these described features

are best to observe in the previously mentioned applet (S1 Applet Access). This work ultimately confirmed that the detected clusters had clear differences in timing and magnitude of peak seasonal LAI as well as temperature and precipitation profiles, but most importantly across all clusters, the majority of variance in regionally averaged LAI was characterized by earlier and higher peaks in LAI as time progressed from 1996 to 2017.

The objective of this work is to justify that clustering using a greenness index, such as LAI, should also involve a follow-up assessment to identify which other remotely sensed environmental factors are strong predictors of cluster allocation of a remotely sensed site. Although, it may appear unnecessarily retrospective or unconventional to study a set of detected clusters, it is crucial to confirm that clustering solely by a single attribute, such as LAI, meaningfully distinguishes the sites across a larger set of attributes. Each site is characterized by a complex set of dynamic and static (or mostly static) attributes, and reporting only difference in clusters across LAI, temperature, and precipitation is an unsatisfying simplification. Clear trends have been detected across clusters towards earlier and higher regional/cluster average annual LAI profiles. However, the trends for annual temperature and precipitation profiles and their associative relationships are not homogenous (Whetten Demler 2021). Further characterization of each cluster will aid in the interpretation of these results. In particular, the secondary objective of this work is to ensure that the clusters which are previously identified to be most dissimilar have the best misclassification rates. Harmonizing the objectives of a cluster analysis and a predictive model is essential in this work in order to correctly interpret the result of the first objective which rely on the predictive model correctly classifying as many sites as possible in highly dissimilar minority clusters.

The characterization of cluster allocation in the cluster model lends naturally to standard machine learning classification methods, such as random forests (Breiman 2001; Kiely et. al. 2020). The use of random forest for spatial data is of growing interest and several adaptations have been developed in recent years (Stefanos et. al. 2021; Hengl et. al. 2018; Geremia et. al. 2013; Hee et. al. 2006). To address the challenges of spatial-autocorrelated, highly localized, and unbalanced class sizes, several techniques for subsampling (Khalilia et. al. 2011; O'Brien et. al. 2019; Chen et. al. 2004) and spatial cross-validation (Adams et. al. 2020; Meyer et. al. 2019; Ramenzan et. al. 2019; Valavi et. al. 2019; Brenning 2012; Stum 2010) of predictive models have been proposed. I rely on many of these techniques and methods in the construction of the proposed model.

I accomplish the outlined objective by proposing a field-subfield spatial cross-validation subsampling procedure (FSSCV) applied to random forest classification that addresses spatial autocorrelation and improves the classification of localized and high-priority minority clusters (which are the smallest class sizes of the response variable used in a random forest model). The choice of prioritizing minority classes in a response variable is application-specific, but generalizable to any problem where the minority class is of great importance or is known to be substantially different from other classes.

Spatial cross-validation and sub-sampling methods are generally implemented disjointly to handle issues of spatial autocorrelation and unbalanced class response

levels since in many applications one of these issues may not be substantial. The FSSCV algorithm proposed in this work combines the subsampling and spatial cross-validation processes by ensuring that within any collection of locations that could be removed during cross-validation, an appropriate distribution of the response variable is preserved in the removed fold and remaining training folds. This approach is a natural choice for increasing the the value of prediction highly localized minority classes since the subsampling requires consideration of spatial distribution of the response variable across each fold in the cross-validation procedure.

Of the site attributes considered for predictors in the final random forest model (elevation, slope, aspect, water storage potential, soil hydrologic unit, and land cover), the FSSCV random forest results indicate that the cluster model separates sites significantly across many of these attributes and further that elevation, slope and water storage potential are the most important predictors of cluster allocation. Most importantly, the class errors rates for the most dissimilar clusters, as detected by the cluster model, have the best misclassification rates. This work provides a foundation for developing methods to assess clustering model results from remote-sensing derived data, and further, provides evidence of the need to holistically consider spatial autocorrelation and unbalanced class levels within an integrated process.

2. Materials and Methods

2.1. Data Products

A general exploration of most of these data products is provided in our applet (refer to S1), and the use of this applet alongside this manuscript is strongly encouraged.

2.1.1. LAI AVHRR Climate Data Record

The LAI Climate Data Record (LAI CDR) produces a daily product on a 0.05 0.05 degree grid dating back to 1981 derived from Advanced Very High Resolution Radiometer (AVHRR) sensors using data from eight NOAA polar orbiting satellites: NOAA -7, -9, -11, -14, -16, -17, -18 and -19. The highest resolution of AVHRR sites is approximately 1km per pixel (Claverie et. al. 2016, Claverie et. Al. 2014). In this analysis, the data is subset from January 1st, 1996, until December 31st, 2017, and the spatial domain is restricted to 37,110 sites in the US portion of the CRB. In this 22-year period, daily LAI measurements are summarized on a weekly resolution, by taking weekly average LAI across a 7-day period. The resulting data product has 1,152 weeks. In this product, there are thousands of sites that report high volumes of missing values.

The LAI CDR required some further pre-processing steps in order to construct the presented cluster model. The construction of spline smoothed curves on the 22-year period required a maximum missing value threshold: 28 percent of weeks in the 22-year period needed to have at least one weekly recording of LAI. This filtering process leaves 27,196 sites. By inspection, it was clear that many of the removed sites are barren/sparsely vegetated regions and high-altitude sites, and there are some systematic errors in the data product that inhibit the detection of LAI readings across

large sections of the high mountain ranges in the region. From these results, the previous work identified that the smoothing process implemented was robust enough to handle sites with higher occurrences of missing values (towards a threshold of 15 to 20 percent), although this was left to future work (Whetten Demler 2021). The functional clustering model used in the previous work is constructed solely from this data product. The results of the cluster analysis prompted further investigation into site characteristics that may be driving factors in the separation of clusters, and the following data products are used in this work to explore this objective.

2.1.2. BaseVue 2013 Land Cover Product

BaseVue 2013 Land Cover, which is a commercial global, land-use/land cover product developed by MDA. BaseVue is independently derived from roughly 9,200 Landsat 8 images and has a spatial resolution of 30 meters. The capture dates for the Landsat 8 imagery range from April 11, 2013, to June 29, 2014, and contain 16 classes of land use/land cover (MacDonald 2014).

2.1.3. USGS National Elevation Product

Site elevations, slopes, and aspects were extracted at each site using the USGS National Elevation product. This dynamic image service provides numeric values on a 30-meter resolution representing orthometric ground surface heights (sea level = 0) which are based on a digital terrain model (DTM) (National Elevation Dataset 2002).

2.1.4. USA Soils Hydrologic Group Product

The value for hydrologic group is derived from the 30-meter resolution (contiguous U.S.) produced by the Natural Resources Conservation Service (NRCS) using the gSSURGO map unit aggregated attribute table field Hydrologic Group - Dominant Conditions (Soil Service Staff 2020).
The seven classes of hydrologic soil group followed by definitions:

1. Group A - Group A soils consist of deep, well drained sands or gravelly sands with high infiltration and low runoff rates.
2. Group B - Group B soils consist of deep well drained soils with a moderately fine to moderately coarse texture and a moderate rate of infiltration and runoff.
3. Group C - Group C consists of soils with a layer that impedes the downward movement of water or fine textured soils and a slow rate of infiltration.
4. Group D - Group D consists of soils with a very slow infiltration rate and high runoff potential. This group is composed of clays that have a high shrink-swell potential, soils with a high-water table, soils that have a clay pan or clay layer at or near the surface, and soils that are shallow over nearly impervious material.
5. Group A/D - Group A/D soils naturally have a very slow infiltration rate due to a high-water table but will have high infiltration and low runoff rates if drained.

6. Group B/D - Group B/D soils naturally have a very slow infiltration rate due to a high-water table but will have a moderate rate of infiltration and runoff if drained.
7. Group C/D - Group C/D soils naturally have a very slow infiltration rate due to a high-water table but will have a slow rate of infiltration if drained.

2.1.5. USA Soils Available Water Storage

The amount of water in soil is dependent on rainfall volume, proportion of rain infiltration into the soil, and the soil storage capacity. Available water storage is the maximum amount of plant available water a soil can provide, and it is an indicator of a soil's ability to retain water and make it sufficiently available for plant use. Available Water Storage capacity estimate for the top 150 centimeters of soil is calculated from the difference between soil water content at field capacity and the permanent wilting point adjusted for salinity and fragments. Data from the gNATSGO database was used to create this product for the contiguous United States and is derived from the 30-meter resolution raster produced by the Natural Resources Conservation Service (NRCS) (Soil Service Staff 2020).

2.2. Spatial Classification with Random Forests

2.2.1. Baseline RF Model

To assess the importance of site attributes in the prediction of cluster allocation, a baseline multinomial random forest classification is implemented on the entire dataset of 27196 sites (Liaw 2002). Using a basic grid search for optimal parameters for "Number of variables randomly sampled as candidates at each split" (mtry) and "Number of Trees" (ntree) with mtry = 2,3,4, ntree = 100, 250, 500, 1,000, the results expose several issues that must be addressed. The training set accuracy in all considered modifications of the tuning parameters yields alarmingly high in-bag error rate of 0.00 to 0.25 with an out-of-bag (OOB) error rate of 0.37-0.45. In Table 1, I report the results of the OOB confusion matrix 500 tree model with mtry = 2. There is an over-prioritization of classifying majority cluster correctly at the expense of dismal prediction performance of the minority classes. With such high OOB error and a complete disregard for classifying minority sites, significant adjustments to the modeling process are required before moving to any further interpretation. Altogether these results are not surprising, and I outline the major issues that must be addressed.

2.2.2 Spatial Autocorrelation

In Figure 2, I present an entropy-based local indicator of spatial association (ELSA) plot of the cluster response variable. The ELSA statistic is a measure of the magnitude of spatial association of a variable at each location relative to its neighboring locations (Naimi et. al. 2019). This indicator simultaneously incorporates both spatial and attribute aspects of spatial association into account for both categorical and

continuous data. The following explanation of the measure focuses on the categorical setting. Assume $X = (x_1, \ldots, x_n)$ are a list of *n*-observations realized from a spatial process at *n*-locations $U = (u_1, \ldots, u_n)$. For a given categorical variables, such as cluster allocation, $A=(\alpha_1, \ldots, \alpha_2)$ represents the possible values of observation x_i (Naimi et. al. 2019). The dissimilarity between a site and its neighboring sites is quantified with the ELSA statistic which is defined by

$$E_i = E_{ai} \times E_{ci} \quad (1)$$

$$E_{ai} = \frac{\sum_j \omega_{ij} d_{ij}}{max\{d\} \sum_j \omega_{ij}}, i \neq j \quad \text{and} \quad E_{ci} = \frac{\sum_{k=1}^{m_w} p_k \, log_2(p_k)}{log_2(m_i)}, i \neq j$$

$$m = \begin{cases} m & \text{if } \sum_j \omega_{ij} > m \\ \sum_i \omega_{ij}, & \text{otherwise} \end{cases}$$

$$d_{ij} = |c_i - c_j|,$$

where w_{ij} is a binary (0 or 1) weight defining the neighborhood size, and it specifies whether the site *j* is within a specified distance of site *i.* The value of m denoted the number of categories of the variables of interest, d_{ij} is the dissimilarity between categories at two sites *i* and *j* which is binary metric for nominal categorical variables or for ordinal variables it is the difference in rank. The value p_k is the probability that the kth category from the $m\omega$ categories within the neighborhood distance from site *i*, and *mi* is the possible number of categories in this same neighborhood. The value Eai is the attribute dissimilarity of the i^{th} from all sites in its neighborhood, and Eci is the normalized Shannon Entropy of the neighborhood (Naimi et. al. 2019). The product of these two measures is a measure of entropy-weighted dissimilarity within a neighborhood of the site.

It is clear from Figure 2, that large swaths of the CRB region have high spatial autocorrelation. Spatial autocorrelation in regression models (including random forest) tends to yield over-fitted and poor performing models, and jeopardized model interpretation (Sinha2019, Dubin 1998). In this analysis, a spatial lag term for the response variable is incorporated as a predictor variable in the final RF procedure proposed in the following sections. The spatial lag terms are derived using row standardized weight matrices of the euclidean distances between all sites. High spatial autocorrelation is present for elevation, slope, and water storage potential, and spatial lag terms for each of these respective variables are incorporated in the RF procedure.

2.2.3. Unbalanced response class levels

The frequency of site cluster allocation is provided in Table 2. The disparity of cluster size between Cluster 1 (n=12,545) and Cluster 4 (n=742) yields challenges for random forest models. The random forest approach although boasting numerous advantages over other predictive models is considered a "greedy" algorithm since by default the objective of a random forest classification is to minimize overall misclassification errors. This inherently places greater emphasis on correctly placing most of the 12,545 sites in Cluster 1 (error = 0.123) while sacrificing the placement of Cluster 4 sites (error = 0.701).

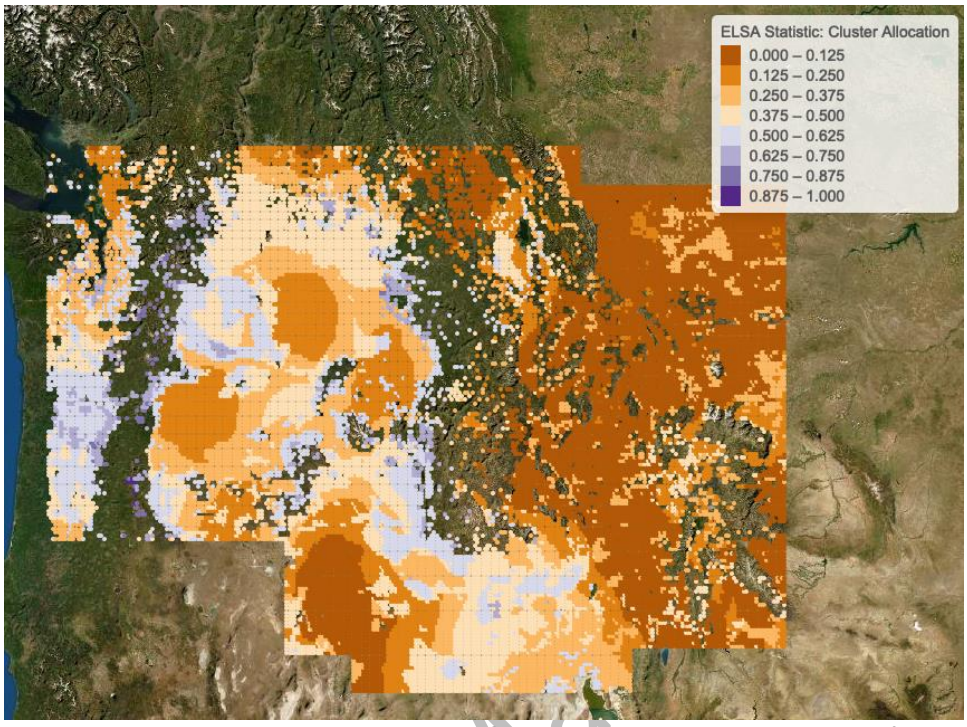

Figure 2. Spatial Autocorrelation of Cluster Allocation. Measure of Spatial autocorrelation in LAI cluster allocation shown in Figure 1 using the ELSA Statistic. A Satellite view is used as the background map layer in this image to increase contrast with the gradient scale of autocorrelation.

Cluster	1	2	3	4	5	Class Error	Class Total
1	11,010	689	719	105	25	0.1226	12,548
2	3,389	1,711	1,562	71	4	0.7460	6,737
3	925	684	3,973	11	0	0.2896	5,593
4	307	80	131	222	2	0.7008	742
5	1,422	56	44	8	41	0.9739	1,571

Table 1. Baseline RF Performance. OOB Confusion Matrix for our standard tuned RF model with numbers of variables tried at each split set to 2 and the number of trees set to 500.

Cluster	1	2	3	4	5
Frequency	12,548	6,737	5,593	742	1,571

Table 2. Site distribution by cluster allocation.

This prioritization is the antithesis of our research objectives in this analysis. The sites in Cluster 4 are identified in previous work to be the most distinguished sites by multidecadal vegetation dynamics and annual climate profiles. Cluster 3 also has substantially different characteristics from most of the other clusters. Further, although there are differences between the other Clusters, they are not as dissimilar from each other, and as such, it is a lower priority to correctly classify these sites. This work is aimed at exploring what site attributes are driving the effective detection of sites that are previously known to be different with respect to the prior attribute, LAI. More simply, the objective is to harmonize the construction predictive model and the prior cluster analysis so that the most dissimilar sites, by some previous metric, have prioritized classification. This priority will assist in indicating which site attributes are most important in predicting the differences detected in the prior cluster analysis.

2.2.4. Highly-localized spatial grouping of minority Clusters

In Figure 1, Cluster 4's sites are exclusive to the narrow region paralleling the Oregon and Washington coast. Additionally, Cluster 3's sites closely follow the major river ways. This phenomenon causes issues with standard cross-validation procedures as well as generic spatial cross-validation procedures. Spatial cross-validation is performed by dividing the data spatially into a collection of n regions and recursively removing one of the n regions as a test set while training on the n-1 regions. The choice of spatial sampling in this application must be carefully considered to avoid complete or "near-complete" removal an entire subgroup of the response variable in the cross-validation process, especially when the classification of these sites is of such great importance.

2.2.5. Transition and Adjacent Classes

Although this is not a primary focus of this manuscript, it is appropriate to mention the misclassification "transition classes." Transition classes are class-level responses that are in some way "in-between" two other classes by some measure of ordinality, and as such, they are prone to misclassification in the adjacent class-levels. This problem is apparent in Table 2. The sites allocated to Cluster 2 in the 5-cluster model were dominantly grouped with Cluster 1 and Cluster 3 for preliminary cluster models of smaller size. Refer to S1 for a visualization of this in the supplementary applet. It is shown that Cluster 2 has one of the worst class error rates of 0.746 and the misclassification of these sites is primarily Clusters 1 and 3.

Adjacent classes simply refer to a class that is similar to another class (and are not necessarily between two classes). Cluster 5 and Cluster 1 are adjacent classes, and

Cluster 5 sites were almost exclusively grouped with the Cluster 1 sites for candidate models with lower numbers of clusters. Refer to the applet again to see this phenomenon. In this work, it is not a top priority to correctly classify transition classes since they may inherently have strong similarities and overlapping site characteristics with other classes.

2.3. Proposed FSSCV RF Model

I propose a spatial subsampling cross validation procedure for the random forest algorithm that addresses the issues listed in the last section. This procedure is referred to as field-subfield spatial cross validation (FSSCV). The steps of this procedure are as follows:

1. Divide the geographic region using k-means clustering on the coordinates of the 27191 sites in the data. We call these groups *fields* to reduce confusion, and refer to them also by $F_1, \ldots, F_n$ where n is the number of fields that the k-means algorithm is set to detect.
2. For levels of the categorical response variable that do NOT span all fields, randomly sub-sample a proportion of these sites, p_{lc}, from the data. The purpose of this is to permit some of the data to be left as a test set if further model validation is desired. This also prevents over-representation of minority clusters.
3. For levels of the categorical response variable that have sites distributed across all fields, sub-sample a collection of sites from each field. If the minimum number of sites in a given field is less than 1/3 the number of sites sampled from the minority cluster in step (2), then sample the minimum number of sites present from any field. The result is an equal amount of randomly selected sites from a single cluster across each field. Otherwise, sample a proportion of the minimum number of sites, p_{gc}, in a field across all fields.
4. Divide the subsamples of the fields $F_1, \ldots, F_n$ into h-subfields referred to as $S_{1i}, \ldots, S_{ni}$ where $i = 1, \ldots, n$. The combinations of field-subfield indices yield $(n \times h)$ folds that are used for spatial cross-validation.

The steps of the FSSCV procedure are also visualized as a flowchart in Figure 3. This subsampling procedure can be tuned in several ways: the number of fields n, the percentage of sampled local cluster sites p_{lc}, the percentage of sampled global cluster sites p_{gc}, and the 1/3 cutoff defined in step (3) of the algorithm. The parameters chosen in our sub-sampling procedure are selected based on application-specific objectives of this analysis, but the following general goals are met: (1) Sub-sampled cluster sizes are more balanced, but the sub-sampling prevents over-representation of highly localized minority clusters, (2) adequate number of sites from each cluster are sampled from each field except for highly localized clusters (where this is not necessarily possible). Over-representation here refers to the risk of losing model interpretability by excessively down-sampling of majority classes while retaining a large quantity of minority classes. This approach,

although relatively simple, has potential adaptations to improve flexibility which are mentioned in the discussion section.

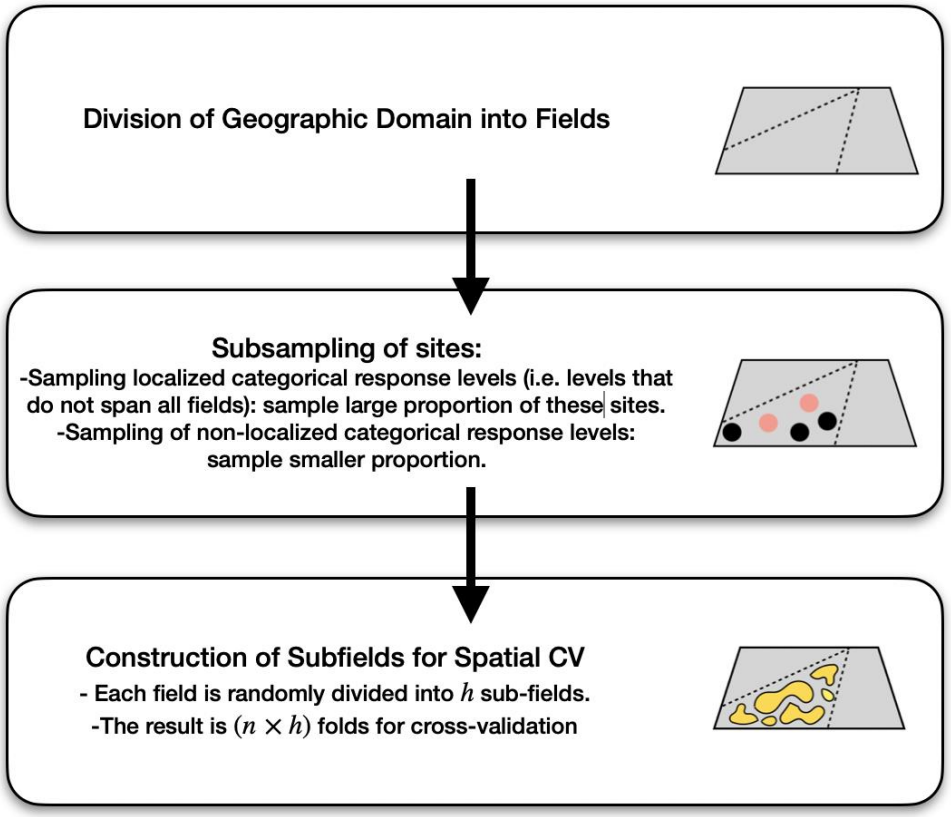

Figure 3. FSSCV Process Map

Figure 4 provides a visualization of our sub-sampling procedure where n=5 and h=3. The specific sizes of the subfield and folds used are provided in the results section. The sub-fields are not shown in this figure to avoid over-complicating the image, but the sub-field sampling would divide each field into 3 (not necessarily equal) parts of which representation from most clusters (response variable classes) in this application is observed. For other applications, the choice of subfield may require adjustment to get desired diversity within subfields.Standard subsampling is not considered here because of the highly localized nature of the clusters and the implications of using generic leave-one-field-out cross-validation. When simple subsampling is used, it is possible to acquire a subsample with adequate representation and spatial distribution across the levels of the response, but it still leaves the problem of removing a field that may have all or almost all of an important minority cluster which is the case in this application. It is cumbersome to subsample and then find appropriate fields to meet these circumstances, and as such the FSSCV method proposed here is intended to "unify the process" so that the sub-fields defined have good subsampling properties.

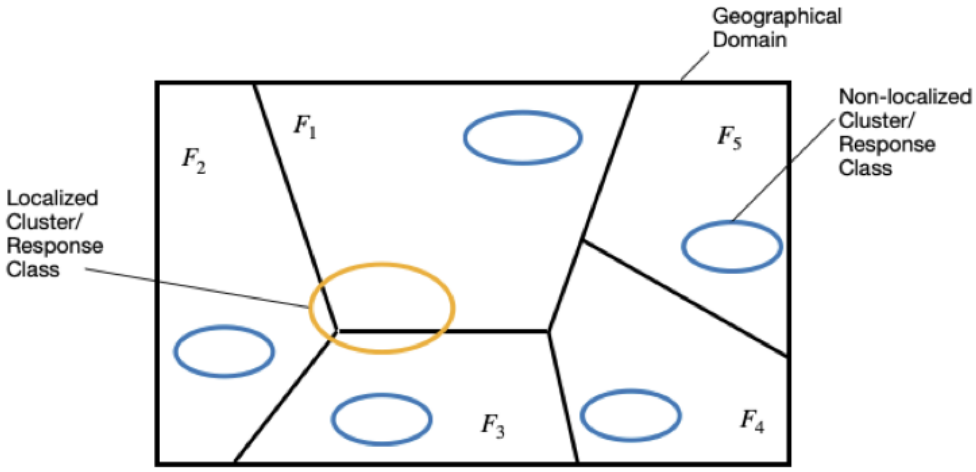

Figure 4. FSSCV Visualization. Example of Geographical field construction for the FSSCV procedure, with an illustration of potential localization of clusters within a sub-group of fields.

3. Results

3.1. Site Attribute exploration by cluster and Field Sub-field Sampling of Site

Since the LAI CDR is the only thoroughly explored data product used in this work, it is instructive to visually explore the spatial distribution and autocorrelation of all variables used in the model. The magnitude of separation of the explanatory variables induced the cluster model is assessed since this will be strongly tied to the classification model to distinguish clusters successfully. In Figure 5 and 6, the box plots of elevation, slope, aspect and water storage potential and the bar charts of 2013 Land Cover and Soil Hydrologic Unit characterize the distribution of these predictors by cluster (Wickam 2016, Baptiste 2015). Significant differences in these distributions are detected across all variables using 1-way ANOVA and non-parametric ANOVA (Kruskal-Wallis) models with Tukey-adjusted comparison of multiple group means and Chi-square goodness of fit tests for Land Cover and Soil Hydrologic unit. It is important to emphasize the drastically lower elevations levels found in Cluster 4 and the lower slopes found in Cluster 3. Visually these differences lend to the intuition that the model should be able to get the most accurate predictions for these clusters in a global RF model (as opposed to performing localized RF models) even though this is not accomplished in the original RF model.

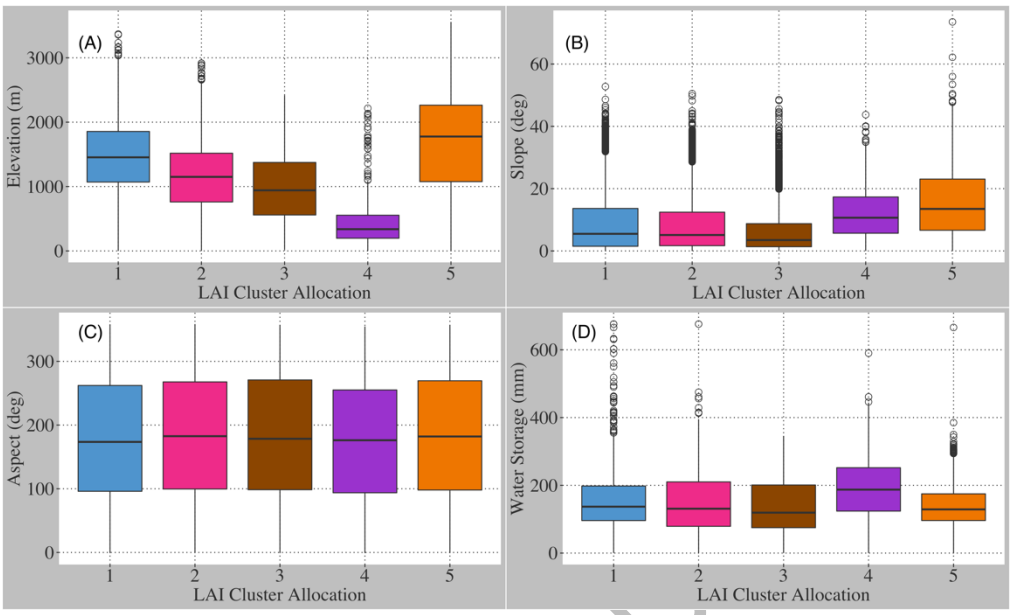

Figure 5. Continuous Site Attributes (A) Elevation, (B) Slope, (C) Aspect, (D) Water Storage Potential. Site Attribute Distributions by Cluster. Boxplots are colored to match the cluster model results in Figure 1.

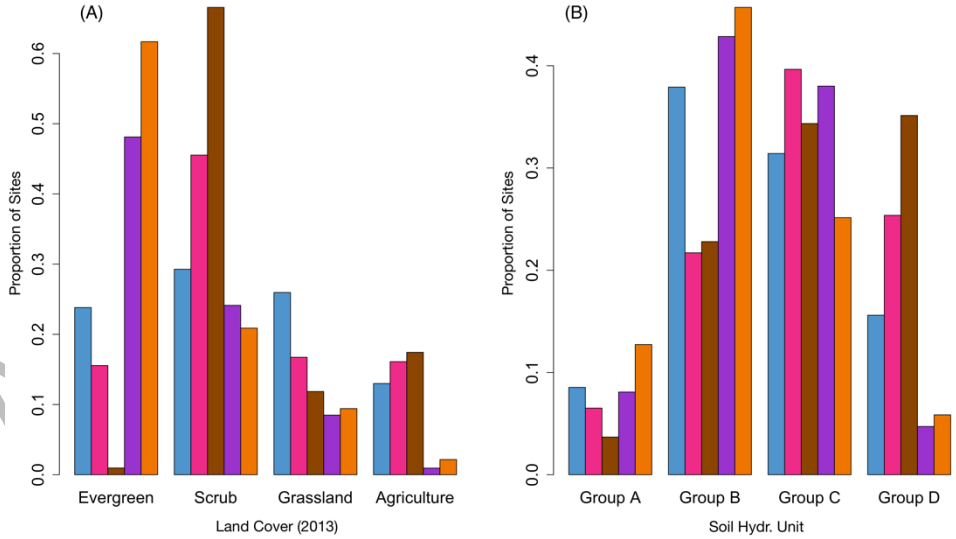

Figure 6. Categorical Site Attributes. Site Attribute Distributions by Cluster. Barcharts are colored to match the cluster model results in Figure 1. The barcharts are scaled to assess the proportion of (A) Land Cover and (B) Soil Hydrologic attribute by cluster. Any collection of bars from the same cluster will have a net area 1 since not all values of each predictor variable are depicted in the picture.

In Figure 7, the fields are shown for the FSSCV procedure using k-means clustering on the latitude and longitude coordinates. The left-most field almost completely contains Cluster 4 and the three left-most fields contain almost sites from Cluster 3. Although this work does not define a measure of localization, it is evident that these clusters are moderately or highly localized in the geographic domain. The full FSSCV procedure subsamples 3,590 sites from the data and the distribution of the fields by cluster is shown in Table 3. Table 4 and Table 5 reports the sub-field distribution of sites by cluster and field respectively. The subsampling and sub-field site selection procedure achieves well-balanced folds to use in the cross-validation of the random forest model. The distribution of all subfield selected sites is visualized Figure 7.

In the construction of our final RF model, spatial lag terms are incorporated for the response variable, as well as elevation, slope, and water storage potential. Spatial lag terms are utilized to account for the spatial autocorrelation in our response variable, and although similar information is contained in the Latitude and Longitude as predictors are shown to improve prediction rates, but these geo-spatial attributes are removed since previous work has shown that highly auto-correlated predictors (such as geolocation variables, e.g. latitude, longitude) can lead to considerable overfitting and result in models that can reproduce the training data but fail in making spatial predictions (Kiely 2020). This was not apparent this application, but these guidelines are followed to avoid complicating the interpretation of this model.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Field 1	187	137	278	0	158	760
Field 2	187	137	170	4	158	656
Field 3	187	137	0	9	158	491
Field 4	187	137	24	6	158	512
Field 5	187	137	154	535	158	1,171

Table 3. Sub-sampled site distribution by field and cluster allocation.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
Sub-field 1	302	254	187	182	272	1,197
Sub-field 2	340	212	206	190	249	1,197
Sub-field 3	293	219	197	218	269	1,196

Table 4: Sub-sampled site distribution by sub-field and cluster allocation. Sites were assigned randomly to temporary subfields 1, 2, and 3. Then the division of the final subfields

as subregions of each of the fields is done by collecting all of the sites with the same subfield values within a respective field.

	Field 1	Field 2	Field 3	Field 4	Field 5	Total
Sub-field 1	241	214	177	186	379	1,197
Sub-field 2	262	233	158	171	373	1,197
Sub-field 3	257	209	156	155	419	1,196

Table 5. Sub-sampled site distribution by field and sub-field allocation. The cells of this table are the folds used in the FSSCV procedure.

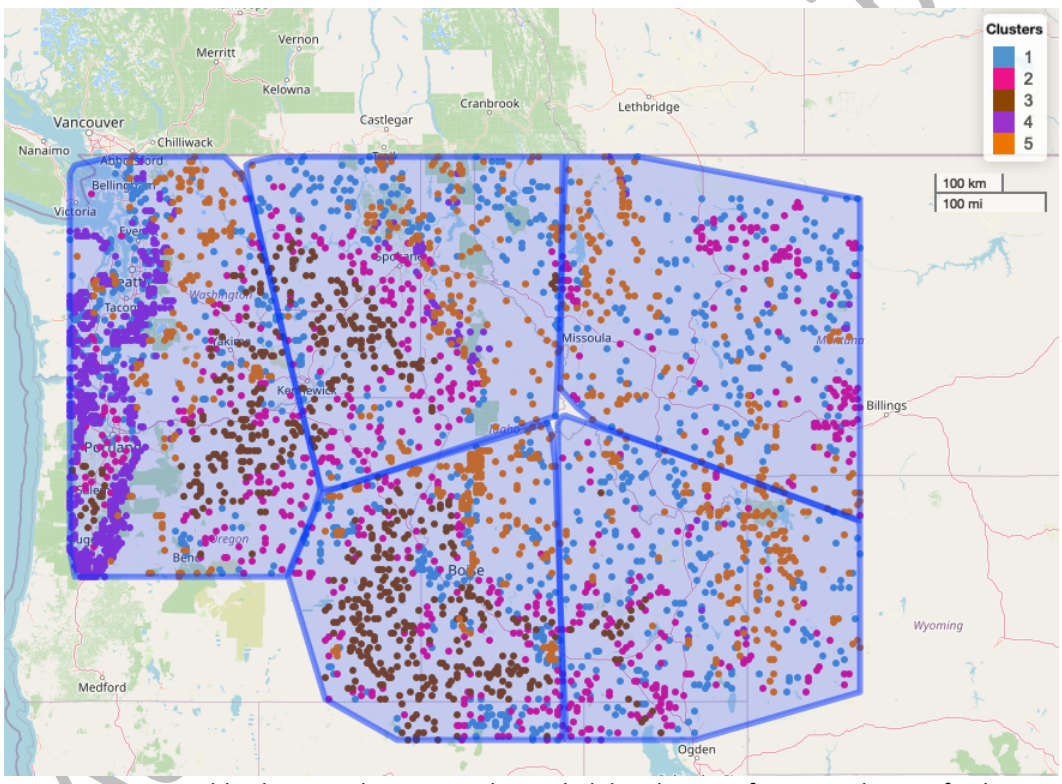

Figure 7. Field Selection. The FSSCV subsampled distribution of sites used in our final RF model as colored by cluster. The polygons are the k-mean selected fields used in this analysis. K-means is a common choice for field generation in spatial cross-validation.

3.2. FSSCV RF Performance Summary

The $(n \times h)$ FSSCV procedure constructs 15 folds that divide the 3590 sampled sites using $n=5$, $p_{lc} = 0.80$, $p_{gc} = 0.20$, and the RF performance summary is provided in Figure 8. Although the global mean FSSCV error is not exceptional, the misclassification error by class confirms that the objectives have been satisfactorily fulfilled. The

misclassification rates of Cluster 4 and Cluster 3 sites across the 15 folds are centered at 0.15 and 0.20 respectively. The variable importance across all 15 folds shown Figure 8 identifies that site elevation is the most important predictor of cluster allocation in the model. The spatial lag term for water storage potential is the second most important predictor, and the second and third most important non-lag term predictors are slope and water storage potential. Since there is a clear relationship between water storage and the spatial lag for water storage, it is evident that some predictive information is shared between these attributes, but it is noteworthy that the geographic information implicitly contained in the lag term made this attribute a highly important predictor. I hypothesize that this is because there are climatic differences across the region, which are geographically dependent, that have effects on the type of vegetation at a site.

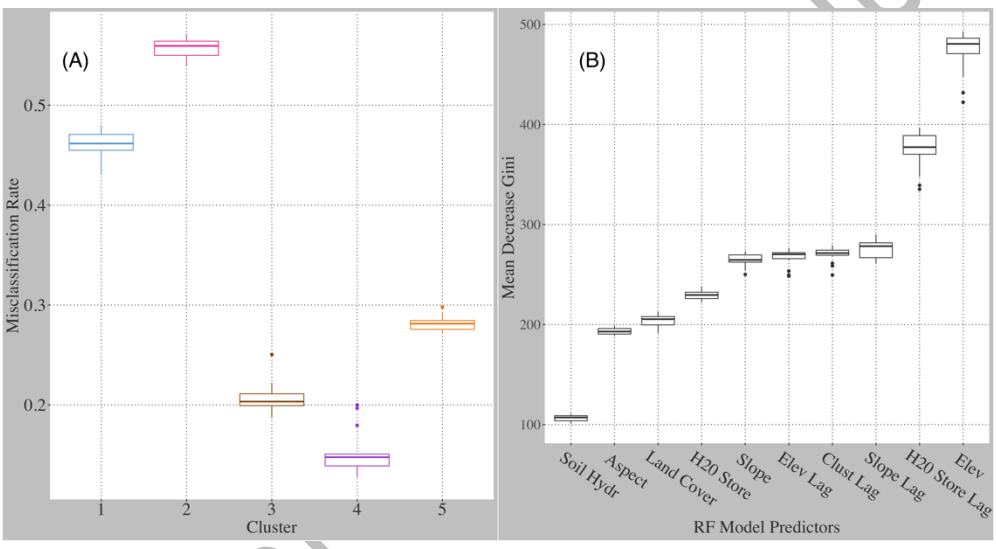

Figure 8. FSSCV General Model Assessment. (A) FSSCV RF performance summary. The misclassification rates for cluster allocation are plotted using boxplots that show the misclassification rates for the removed fold across all 15 folds. (B) FSSCV RF variable importance plot. The distribution of an attribute's predictive performance across all 15 folds is plotted as a boxplot.

The relationships of the previously mentioned attributes to the probability of prediction of sites to the Cluster 3 and Cluster 4 allocation are explored using partial dependence plots (PDPs). Since the final model is effectively making predictions for these attributes, it is of interest to understand further how their attributes increase or decrease the probability of being classified in these clusters. In Figure 9, the PDPs for Elevation, Slope, and Water Storage Potential are presented. Sites with elevation less than 1000m have higher probability of a Cluster 4 site prediction, and sites with elevation approximately between 0 to 1700m have higher probability of a Cluster 3 site prediction. The clear overlap between elevation classification probability is not surprising as both clusters have the lowest elevation distributions of the 5 clusters.

It is across slope and water storage potential that the greatest disparities are observed between the prediction probabilities for these clusters. Sites with slopes closer to 0 degrees have a high probability of a Cluster 3 prediction, and the probability of a Cluster 3 prediction declines drastically as the slope of the site increases. Sites that have slopes between 5 and 25 degrees have the highest probability of a Cluster 4 prediction. For water storage potential, sites with observations greater than 150mm had the highest probability of a Cluster 4 prediction whereas site observations less than 200m have a higher predicted probability of a Cluster 3 prediction.

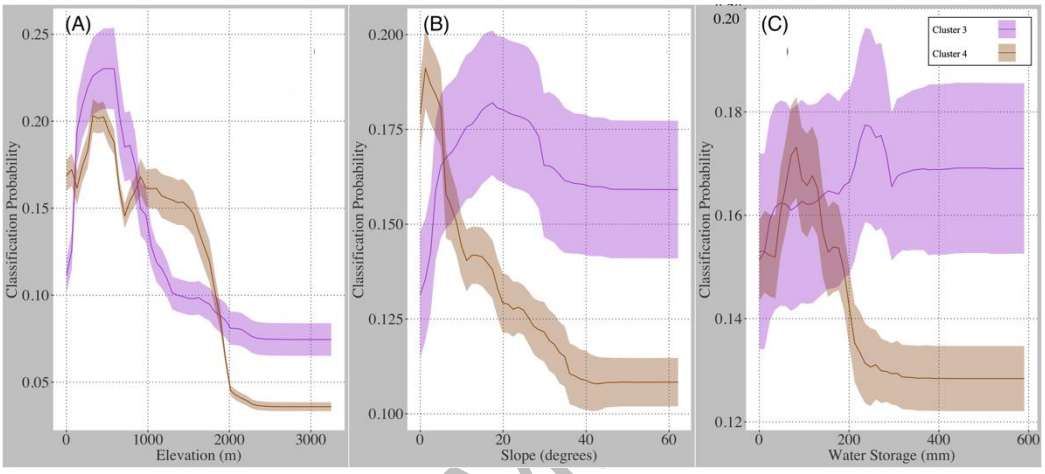

Figure 9. Probabilistic Relationship of Cluster Allocation to sight attributes. PDPs for Elevation (A), Slope (B), and Water Storage (C): We emphasize that the interpretation of the magnitude of classification probability is not to be interpreted directly since PDPs by construction assess the changes in prediction probability across changes in the variables of interest while all others are held constant at their average. Although Elevation is the most important predictor of cluster allocation, Slope and Water Storage Potential are vital in correct Cluster 3 and Cluster 4 model predictions (Wickam 2016, Baptiste 2015, Greenwell 2017). The partial dependence is evaluated for each fold, and the average partial dependence across all folds is plotted as the solid line in each of these plots. The colored ribbon marked standard error of folds from the average partial dependence line.

The magnitude of classification probability is not to be interpreted directly since PDPs by construction assess the changes in prediction probability across changes in the variables of interest while all others are held constant at their average. The differences in Cluster 3 and Cluster 4 classification probabilities highlight that a clear interaction exists between these attributes in order to make cluster allocation predictions. In this case, it is clear that although elevation is the most important predictor in the model, the magnitude of water storage potential and slope are driving factors in the correct classification of Cluster 3 and Cluster 4 sites, namely that Cluster 4 sites generally have higher water storage potential and higher slopes, which are characteristic of coastal evergreen forests, and Cluster 3 sites have lower water storage potential and lower slopes, which are characteristics of major river valleys in the CRB dominantly covered by scrub, grasslands, and agriculture. Interactions can be investigated further using various methods, but this is left to future work.

It is also of interest to assess which regions these predictors are providing the greatest aid to cluster allocation predictions. Geographical local variable importance plots (LIMPs) are examined in Figures 10 through 13. In Figure 10, the explanation of high misclassification rates for Clusters 1,2 and 5 is apparent. Site elevation is highly important for classifying coastal site cluster allocation, river way cluster allocation and pockets of the national park region of Western Wyoming. The importance of elevation in other regions, that are dominantly characterized by these cluster assignments, is notably lower. In Figure 11, the slope of a site is highly important for classifying sites in the Magic Valley of region of Southern, Idaho, and some other small pockets of sites along the Snake and Columbia Rivers. In Figure 12 and 13, the joint role that water storage potential and the spatial lag term for water storage potential play in making cluster allocation predictions is identified. The spatial lag term for water storage is important for classifying coastal sites and sites along the lower Columbia River, while the original water storage variable is important for classifying sites along the entire Idaho border and large sections of the Snake River in Southern Idaho. This provides meaningful insight into the cluster allocation since most sites neighboring the Pacific Coast and lower Columbia River have similar water storage properties in spite of many clusters being present across these areas. The spatial weighting of this attribute aids in making prediction in this region, whereas across a large swath of the CRB, the cluster allocation has a remarkably similar shape to the spatial distribution of water storage potential which is a driving factor in the predictive importance of water storage in these areas.

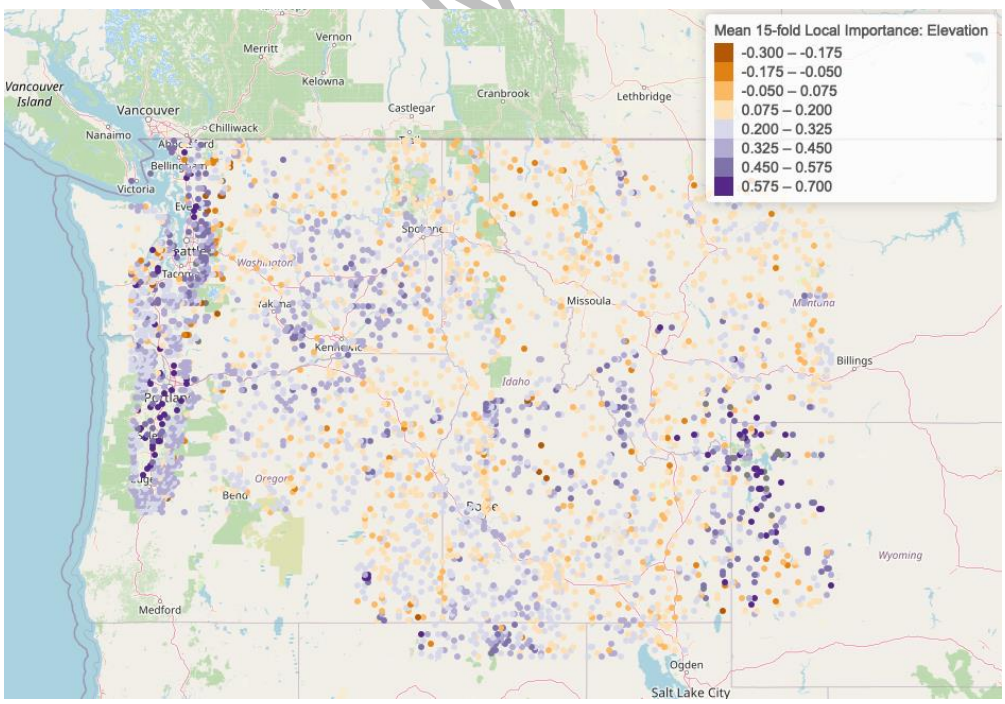

Figure 10. Local Importance of Elevation in the FSSCV random forest model. Similar to our visualization of partial dependence in Figure 9, local importance is averaged across all 15 folds. Variability from the average local importance for each site is not provided.

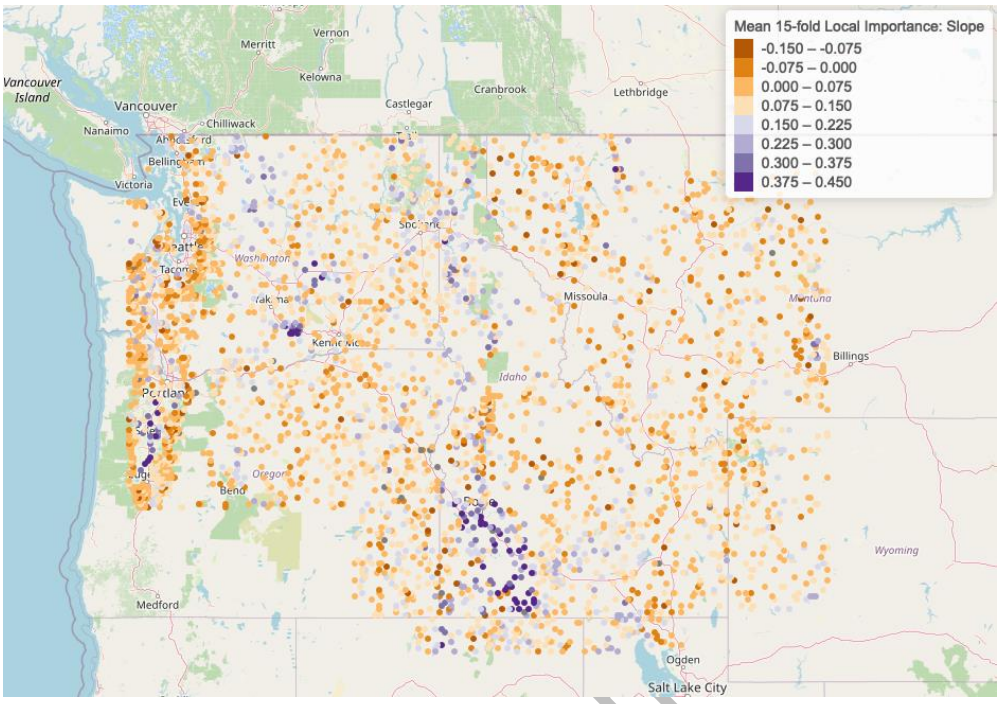

Figure 11. Local Importance of Slope in the FSSCV random forest model.Similar to our visualization of partial dependence in Figure 9, local importance is averaged across all 15 folds. Variability from the average local importance for each site is not provided.

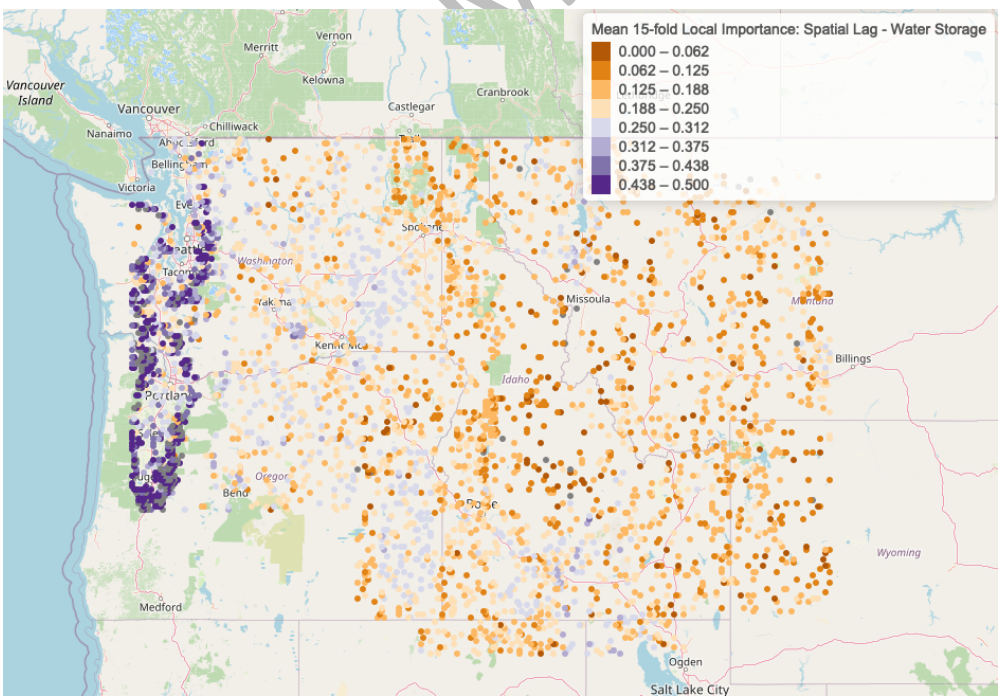

Figure 12. Local Importance of Water Storage Spatial Lag Term in the FSSCV random forest model.Similar to our visualization of partial dependence in Figure 9, local importance is averaged across all 15 folds. Variability from the average local importance for each site is not provided.

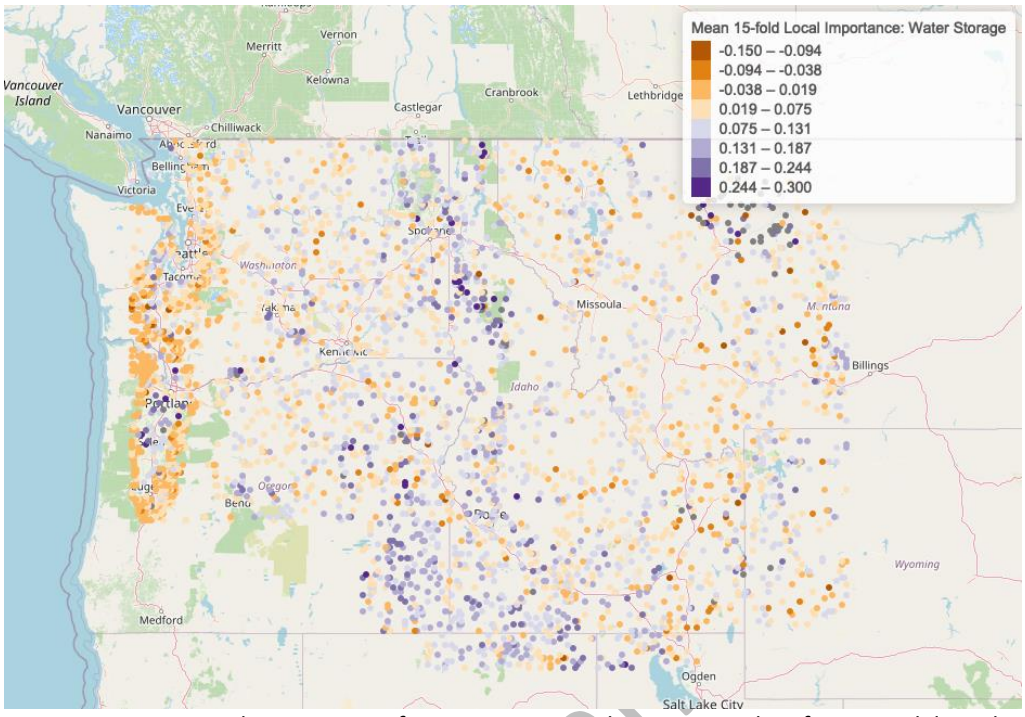

Figure 13. Local Importance of Water Storage in the FSSCV random forest model.Similar to our visualization of partial dependence in Figure 9, local importance is averaged across all 15 folds. Variability from the average local importance for each site is not provided.

4. Discussion

This work, motivated by the need to explore cluster analysis results, provides a foundational method for combining the sub-sampling and spatial cross-validation procedures. The focus of this work was to prioritize highly localized minority clusters that are known by some other metric to be the most different from the other clusters. The derived exploratory classification model agrees with the objectives of the cluster analysis, and further, the model accounts for spatial autocorrelation, has acceptable misclassification rates for high-priority minority clusters. Consequentially, the examination of site attribute importance in predicting cluster allocation is more interpretable. The use of information from each fold also improves our interpretation of variable importance and probabilistic relationships of site attributes to cluster allocation.

There are multiple disadvantages to this approach that should be acknowledged, and potential improvements to the subsampling and modeling procedure are mentioned here. As currently presented, the proposed approach is only considering 1/9 of sites using the FSSCV subsampling. This was strategic, but a somewhat arbitrary choice. For this data product and approach, it is plausible that this approach could incorporate to 1/3 to 1/2 of sites, and this would need to be shown in future work. This subsampling procedure also has the capability in its current form to account for more sites using a *recursive subsampling* strategy. As an example, the

results of our analysis are from a single subsample of 3,590 sites. This could repeat times with replacement of the sites, and the variable importance and partial dependence can be averaged and visualized across the realized subsamples. Since the model only considers 3590 sites, it is also of great interest to utilize the extensive supply of test set data not considered in the subsample of 3590 to assess the stability and performance of our model further.

As mentioned in the Proposed FSSCV section, this approach can be tuned across several parameters to balance the clusters to the preference and judgement of the user. Beyond the scope of this work is the fascinating question of grid-searching for the optimal parameters to improve accuracy of all classes. It is apparent when comparing the baseline RF to the FSSCV RF used for this application that our sub-sampling choice has “reversed" the priority of cluster classification to the point of poorly classifying the Cluster 1 and Cluster 2 sites. This is shown in Table 6. For this application, this is expected and appropriate since it is known from previous work that these clusters are two of the most similar clusters. The grouping of these two clusters into a single cluster would eliminate substantial cross-misclassification of these clusters and improve global accuracy, although this is not of primary interest here. There is no current way of optimizing error rates of majority and minority classes across all folds of such an approach.

Cluster	Baseline RF Class Error	FSSCV RF Class Error	Class Total
1	0.1226	0.4611 +/- 0.0125	12,548
2	0.7460	0.5574 +/- 0.0098	6,737
3	0.2896	0.2063 +/- 0.0155	5,593
4	0.7008	0.1525 +/- 0.0222	742
5	0.9739	0.2812 +/- 0.0078	1,571

Table 6. Performance summary of Baseline and FSSCV random forests models.

There is also no explicit assessment of interactions shown in this analysis. This is left to future work using essential random forest packages such as the randomForestSRC package, and it is considered out of the scope of this project (Ishwaran 2021). Although the response variable in this analysis is a dynamic variable, meaning that it has grouped sites using a temporal measure of association, all predictors in this model are considered static for this work. This assumption is not free of flaws, since attributes such as land cover are known to change over time. However, the inclusion of dynamic variables is ultimately of great interest and could be performed using difference indices marking the change in an attribute over time. As a result of the unique characteristics of a dynamic variable, there are several possibilities, and for brevity, they are not discussed further here.

I advocate for more complex techniques for detecting highly localized and minority clusters. This is done strictly by inspection in our work, as the application and

geographical visualizations lends naturally to this simple approach, but important future work lies in detecting locality using some measure of geographic entropy and relative frequency of site cluster allocation (Leibovici 2009).

5. Conclusion

This work is a furthers the ongoing discussion of spatially conscious machine learning models that celebrate and account for the phenomena of Tobler's Law of Geography. Clustering models are generally used as a means to study structure in the data, but it is less common to study the detected structure with the use of predictive modeling. A cluster analysis naturally requires careful selection of method, cluster size, numbers of clusters, and, in most cases, a careful choice of dissimilarity measure. The development of application-specific assessments for validating clustering model quality and interpretability are important, and this work outlines such an approach for high-dimensional remotely sensed gridded or ungridded data, and the results of this work justify future work in improving spatial cross-validation strategies for machine learning models.

Supporting Information

S1 Applet Access: CRB Exploratory Applet
Use https://abwhetten.shinyapps.io/CRB_LAI_1996_2017/ to access the supplementary applet.

S2: Supplementary Figure
Interannual profiles of regional average weekly maximum LAI, maximum Temp, and cumulative precipitation. The curves are colored on a gradient scale where greener curves are closer to 1996 and pinker curves are closer to 2017. Noticeable time-dependent changes in LAI were identified where no clear trend in temperature and precipitation is visually observed.

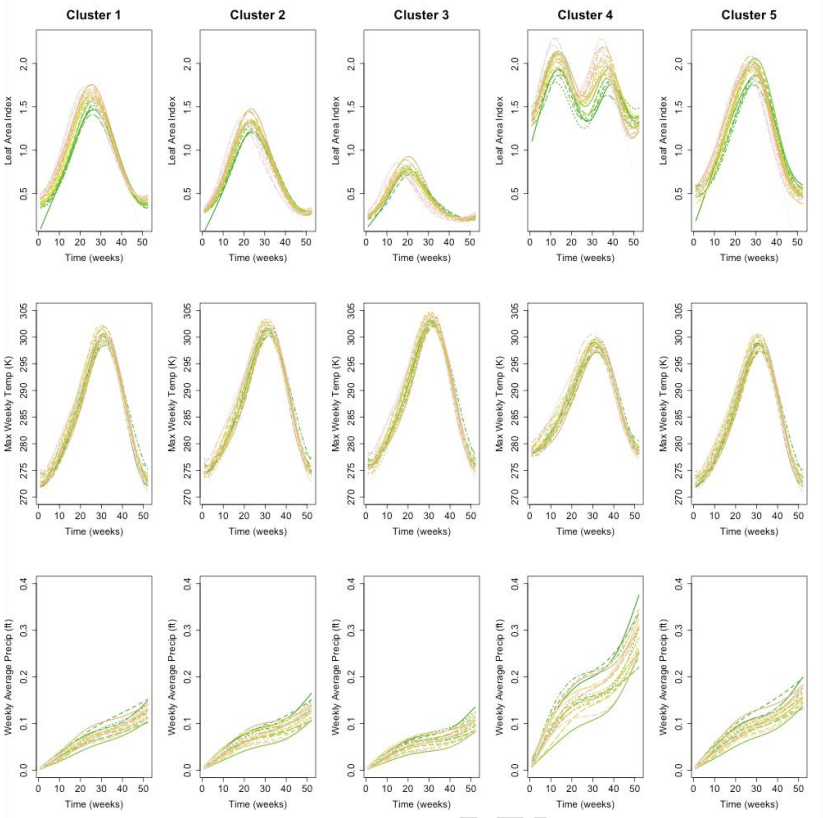

Figure S2: This is a supplementary figure since it is only used to briefly summarize the evidence found in prior work that this cluster model is effectively separately sites beyond differences in NOAA AVHRR satellite measured LAI.

References:

Liaw, A., and Wiener, M. (2002) Classification and Regression by randomForest. *R News*. 2(3), 18–22.

Adams M.D., Massey F., Chastko, K., Cupini, C. (2020) Spatial modelling of particulate matter air pollution sensor measurements collected by community scientists while cycling, land use regression with spatial cross-validation, and applications of machine learning for data correction. Atmospheric Environment. Volume 230. https://doi.org/10.1016/j.atmosenv.2020.117479.

Baptiste A. (2015) gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.0.0.http://CRAN.R-project.org/package=gridExtra

Berner, L.T., Massey, R., Jantz, P. et al. (2020) Summer warming explains widespread but not uniform greening in the Arctic tundra biome. *Nat Commun* 11, 4621. https://doi.org/10.1038/s41467-020-18479-5

Greenwell, B.M. (2017) pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1), 421–436.https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

Breiman, L. (2001) Random Forests. *Machine Learning* 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brenning A. (2012) "Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest," IEEE International Geoscience and Remote Sensing Symposium pp. 5372-5375. DOI: 10.1109/IGARSS.2012.6352393.

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., and Zemp, M. (2014) The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. Bulletin of the American Meteorological Society, 95(9), 1431–1443. https://doi.org/10.1175/BAMS-D-13-00047.1

Chen C., Liaw A., Breiman L. (2004) Using random forest to learn imbalanced data. Discovery (University of California Technical Report 666). 1999:1-12. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Cheng, J., Karambelkar, B., and Xie Y. (2019) leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.3. https://CRAN.R-project.org/package=leaflet

Claverie, M., Matthews, J., Vermote, E., Justice, C. (2016) A 30+ Year AVHRR LAI and FAPAR Climate Data Record: Algorithm Description and Validation. Remote Sensing. Vol 8, Issue 3: 263.

Claverie, M., Vermote, E., NOAA CDR Program. (2014) NOAA Climate Data Record (CDR) of Leaf Area Index (LAI) and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Version 4. [indicate subset used]. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5M043BX. Accessed [04/25/21].

Dubin R.A., Spatial Autocorrelation: A Primer. Journal of Housing Economics. Volume 7(4): 304-327. 1998. https://doi.org/10.1006/jhec.1998.0236.

Forzieri G., Duveiller G. (2018) Evaluating the Interplay Between Biophysical Processes and Leaf Area Changes in Land Surface Models. JAMES: 10(5) pp 1102-1126. https://doi.org/10.1002/2018MS001284

Geremia E., B. H. Menze and N. Ayache, "Spatially Adaptive Random Forests," 2013 IEEE 10th International Symposium on Biomedical Imaging, 2013, pp. 1344-1347, doi: 10.1109/ISBI.2013.6556781.

Stefanos, G., Tais Grippa, T., Gadiaga A.N., Linard C., Lennert, M., Mboga S.V.N., Kalogirou E.W.S. (2021) Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling, Geocarto International, 36:2, 121-136, DOI: 10.1080/10106049.2019.1595177

Hamlet, A.F., Elsner, M.M., Mauger, G.S., Lee, S.-Y., Tohver, I., Norheim, R.A. (20130 An Overview of the Columbia Basin Climate Change Scenarios Project: Approach, Methods, and Summary of Key Results. Atmos. Ocean. 51, 392–415.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

Hopkinson, C., Fuoco, B., Grant, T., Bayley, S.E., Brisco, B., MacDonald, R. (2020) Wetland Hydroperiod Change Along the Upper Columbia River Floodplain, Canada, 1984 to 2019. *Remote Sensing*. 12(24):4084. https://doi.org/10.3390/rs12244084

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler B. (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. 6, 5518. DOI: 10.7717/peerj.5518. PMID: 30186691; PMCID: PMC6119462.

Hee Wai, T., Young, M.T., Szpiro A.A., Random Spatial Forests. ArXiv. https://arxiv.org/abs/2006.00150

Kogalur I.H. (2021) Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.11.0, https://cran.r-project.org/package=randomForestSRC.

Khalilia, M., Chakraborty, S. and Popescu, M. (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11, 51. https://doi.org/10.1186/1472-6947-11-51

Kiely, TJ, Bastian, ND. (2020) The spatially conscious machine learning model. Stat Anal Data Min: The ASA Data Sci Journal. 13:31–49. https://doi.org/10.1002/sam.11440

Knowles, N., Dettinger, M.D., Cayan, D.R. (2006). Trends in Snowfall versus Rainfall in the Western United States. J. Clim. 19, 4545–4559.

Leibovici D.G. (2009) Defining Spatial Entropy from Multivariate Distributions of Co-occurrences. In: Hornsby K.S., Claramunt C., Denis M., Ligozat G. (eds) Spatial Information Theory. COSIT 2009. Lecture Notes in Computer Science, vol 5756. Springer, Berlin, Heidelberg.

MacDonald, Dettwiler and Associates Ltd. (MDA). (2014) BaseVue 2013. Available at: http://www.arcgis.com/home/item.html?id=1770449f11df418db482a14df4ac26eb [Last accessed 23 March 2021].

Meyer, H., Reudenbach. C., Wöllauer, S., Nauss T. (2019) Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecological Modelling. Volume 411. 108815.ISSN 0304-3800.

Naimi, B., Hamm, N.A., Groen, T.A., Skidmore, A.K., Toxopeus, A.G., Alibakhshi S. (2019) "ELSA: An Entropy-based Local indicator of Spatial Association." Spatial Statistics, 29, 66-88. https://doi.org/10.1016/j.spasta.2018.10.001

Naimi, B., Hamm, N. A., Groen, T. A., Skidmore, A. K., Toxopeus, A. G., and Alibakhshi, S. (2019) ELSA: Entropy-based local indicator of spatial association. Spatial statistics, 29, 66-88.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

National Elevation Dataset. (2002) Web site; U.S Geological Survey

O'Brien, R., Ishwaran, H. (2019) A Random Forests Quantile Classifier for Class Imbalanced Data. Pattern recognition, 90, 232–249. https://doi.org/10.1016/j.patcog.2019.01.036

Piao, S., Fang J. (2003) Interannual variations of monthly and seasonal normalized difference vegetation index (NDVI) in China from 1982 to 1999. JGR Atmospheres: 48(14). https://doi.org/10.1029/2002JD002848

Queen, L. E., Mote, P. W., Rupp, D. E., Chegwidden, O., and Nijssen, B. (2021) Ubiquitous increases in flood magnitude in the Columbia River basin under climate change, Hydrol. Earth Syst. Sci., 25, 257–272, https://doi.org/10.5194/hess-25-257-2021.

Ramezan, A.C., A. Warner, T.E., Maxwell, A. (2019) Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. Remote Sens. 11, 185. https://doi.org/10.3390/rs11020185

Sinha, P., Gaughan, A.E., Stevens, F.R., Nieves, J.J., Sorichetta, A., Tatem, A.J. (2019) Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. Computers, Environment and Urban Systems. 75: 132-145. https://doi.org/10.1016/j.compenvurbsys.2019.01.006.

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available online at https://websoilsurvey.nrcs.usda.gov/. Accessed [04/25/21].

Soil Survey Staff. Gridded National Soil Survey Geographic (gNATSGO) Database for the Conterminous United States. United States Department of Agriculture, Natural Resources Conservation Service. Available online at https://nrcs.app.box.com/v/soils. December 1, 2020 (FY2020 official release).

Stum, A.K., "Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah." (2010) All Graduate Theses and Dissertations. 736. https://digitalcommons.usu.edu/etd/736

Sumida, A., Watanabe, T., Miyaura, T. (2018) Interannual variability of leaf area index of an evergreen conifer stand was affected by carry-over effects from recent climate conditions. Sci Rep 8, 13590. https://doi.org/10.1038/s41598-018-31672-3

Tawatchai N., and Xingguo M. (2017) "Detecting Spatial and Temporal Change of NDVI Dynamics in the Mekong River Basin: Relationship with Anthropogenic Effects,"

International Journal of Environmental Science and Development vol. 8, no. 10, pp. 719-723.

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G. (2019) blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol. 10: 225–232. https://doi.org/10.1111/2041-210X.13107

Whetten A.B., Demler H. (2021). Detection of Multidecadal Changes in Vegetation Dynamics and Association with Intra-annual Climate Variability in the Columbia River Basin. In *ArXiv e-prints*. arXiv: 2105.08864 [q-bio.QM]