Investigating Selection Criteria of Constrained Cluster Analysis: Applications in Forestry

Gavin Richard Corral

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Statistics

JP Morgan
Jeffrey Birch
Pang Du

10/29/2014
Blacksburg, VA

Keywords: Forestry, Cluster Analysis, Redundancy Analysis, Simulation

Investigating Selection Criteria of Constrained Cluster Analysis: Applications in Forestry

Gavin Richard Corral

ABSTRACT

Forest measurements are inherently spatial. Soil productivity varies spatially at fine scales and tree growth responds by changes in growth-age trajectories. Measuring spatial variability is a perquisite to more effective analysis and statistical testing.

In this study, current techniques of partial redundancy analysis and constrained cluster analysis are used to explore how spatial variables determine structure in a managed regular spaced plantation. We will test for spatial relationships in the data and then explore how those spatial relationships are manifested into spatially recognizable structures. The objectives of this research are to measure, test, and map spatial variability in simulated forest plots.

Partial redundancy analysis was found to be a good method for detecting the presence or absence of spatial relationships (~95% accuracy). We found that the Calinski-Harabasz method consistently performed better at detecting the correct number of clusters when compared to several other methods. While there is still more work that can be done we believe that constrained cluster analysis has promising applications in forestry and that the Calinski-Harabasz criterion will be most useful.

Acknowledgements

I would like to express my deep appreciation for the continued support of my major advisors, Dr. Burkhart and Dr. Morgan. I would also like to thank my committee members Dr. Birch and Dr. Du for their continuous support and encouragement throughout my graduate experience in the Department of Statistics. Also, a special thanks to Dr. Legendre for helping me develop the ideas necessary for this project.

I am lucky to have tremendous support from my entire family. In particular to my best friends in the world, Ry and Kahlia, thank you for always believing in me and being there for me. Last but certainly not least, to my wonderful girlfriend, partner in crime, code writer extraordinaire, and the sweetest woman I have ever known, Anita. Thank you all for making this possible.

Table of Contents

List of Figures

List of Tables

Introduction

Forests are an invaluable part of our environment. They provide ecosystem services like carbon sequestration, clear air and water, fuel, food, and have aesthetic value and recreational use. Furthermore, products from forests can have great economic value. Timber products and livestock grazing, for example, are economic commodities. Managing forests is complex and requires advanced knowledge of tree growth and stand dynamics.

Understanding stand dynamics requires knowledge of how the stands vary through time and space. Many studies in forestry have looked at how trees grow and stands develop through time. Statistical models and methods for examining the effect of time are commonplace in ecological studies. Studied to a much lesser extent, but equally as important is the study of how space is related to tree growth and stand development. Spatial heterogeneity is a fundamental concept to many ecological theories including patterns of succession, stand disturbance, habitat structure, and site productivity.

A need for an investigation into fine scale spatial effects on tree development exists. Microsites are patches within a stand (also called a site) that have distinct levels of productivity. These discontinuities in site productivity cause measureable changes in growth. The differences in growth rate can result in trees becoming either suppressed or dominant. The structural differences caused by these microsites are called spatially recognizable structures. These structures can be spatially analyzed.

Testing for spatial relationships, measuring spatial variance, and mapping spatial structures leads to a rich understanding of underlying ecological processes. Many interrelated processes such as depth and composition of soil layers, water and nutrient availability, decomposition rates, and macro and microfauna composition govern site and microsite productivity. All these processes affect the growth response of trees. Recognizing and understanding these spatial patterns is integral in being able to model tree growth and understand stand dynamics.

There are many benefits of studying the spatial patterns of tree growth. Testing for spatial heterogeneity of growth will allow practitioners to make more informed decisions into site preparation and treatments by refining delineation of treatment areas. Investigating and eventually mapping spatial structures can have economic benefits. Application of treatments to more precise locations can save time and money. Most importantly, spatial patterns play a key role in stand development. Forest data is inherently spatial. Testing spatial heterogeneity, measuring spatial variance, and mapping spatial patterns is fundamental to understanding forest processes at a stand level.

Our methods will be a logical first step in a focused effort to better understand spatial patterns. We will use contemporary statistical techniques and provide results that can act as a starting point for further analyses and experimentation.

Literature Review

Forests develop through time and across space. Stand development and dynamics are examined, tested, and analyzed with respect to time. Temporal dependences are accounted for and its effects are measured. Spatial effects receive less attention, but are of equal importance. Forest measurements are inherently spatial. Soil productivity varies spatially at fine scales and tree growth responds by changes in growth-age trajectories. Measuring spatial variability is a perquisite to more effective analysis and statistical testing.

Environmental scientists have a major interest in spatial analysis (Legendre & Legendre 1998). Forest cover and tree distributions are in part determined by spatial variation (Leduc et al. 1992). In particular, soil patterns within forests are seen to mostly affect tree distributions (Leduc et al. 1992). This phenomenon is prevalent in most environmental studies because ecological measurements taken by sampling geographic space are affected by spatial components (ter Braak & Prentice 1988, Leduc et al. 1992, Legendre & Legendre 1998, Peres-Neto et al. 2006). Legendre & Legendre (1998) note that ecological processes create spatially recognizable structures, which may display spatial patterns and be the subject of spatial analysis. Hulbert (1984), acknowledges that through mensurative experiments, which include measurements made at multiple points in space, will allow one to test hypotheses about patterns in space.

Discontinuities in soil productivity often happen at fine scales. The patches formed by these discontinuities are called microsites. These microsites cause trees to grow along different height-age trajectories (Oliver & Larson 1996). Weber (1983) recognizes that the height-age curves for these microsites can be polymorphic (i.e curves are not proportional). Polymorphism occurs when one microsite may favor rapid early growth and another may favor rapid late growth (Oliver & Larson 1996). Unlike anamorphic differences where curves develop proportionally, Polymorphic patterns imply that structural differences may develop at different points in time and not necessarily be exhibited evenly through time. Oliver & Larson (1996) explain that trees on good microsites are more likely to become dominant if they grow next to trees on poor microsites. This gives rise to spatially recognizable structures, which develop spatial patterns, that may be spatially analyzed (Legendre & Legendre 1998). A reasonable approach to investigate the phenomena of spatially recognizable structures in tree growth would be to first test if spatial relationships exist between growth and space. The methods of canonical analysis and ordination are well suited to analyze spatial relationships.

Ordination is used in canonical analysis as a powerful tool known to many ecologists as redundancy analysis (RDA) (Peres-Neto et al. 2006, Borcard et al. 2011). Ordination is a common technique that is used widely by ecologists (Borcard et al. 2011, Legendre & Legendre 1998). At its most basic, ordination is simply the arrangement of units into some order (Goodall 1954). Bray (1957) observed that a rise in the use and application of ordination techniques began in the 1950's. Work such as Motyka et al. (1950), Curtis & McIntosh (1951), Brown & Curtis (1952), Vries (1952), Webb (1954) and Poore (1955) began applying and building quantitative techniques for plant community classification. Through the years ordination has been used in forestry as a useful tool in studying the variance of some response, generally vegetation composition, across a landscape (see Greig-Smith 1967, Peet 1981, Lahti 1987, Martel et al. 2007, Grimaldi et al. 2014). Borcard et al. (1992) proposed a new method that utilized ordination techniques. This method utilized preexisting techniques in canonical analysis to partition variation into independent components. Borcard et al. (1992) used partial RDA (pRDA) to perform this task.

Redundancy analysis is the investigation of explained variance (Gittins 1985, Legendre & Legendre 1998). Rao (1964) was the first to describe RDA and it was later rediscovered and presented by Van den Wollenberg (1977). Since then, RDA has been applied widely in many different aspects of ecology (Peres-Neto et al. 2006). The search for causes of spatially recognizable structures is of great importance to environmental scientists and RDA is an important tool for this (Peres-Neto et al 2006). Redundancy analysis provides the means for conducting direct explanatory analysis in which the association among growth measures can be studied according to their shared relationship with environmental and spatial variables (Peres-Neto et al. 2006). Peres-Neto et al. (2006) notes that over 1500 studies have been published applying canonical correspondence analysis (an extension of RDA) and RDA in modeling species-environment relationships. Borcard et al. (1992) introduced variation partitioning using two sets of explanatory variables. One set of variables was called environmental. The environmental variables were described as descriptors that are not spatially structured. The other set of explanatory variables were spatially structured (Borcard et al. 1992). Borcard et al. (1992) described applications of pRDA that could be used to partition the variation of observed responses into components of variation. The partitioning included pure environmental, pure spatial, environmental-spatial, and undetermined components of the variation. This method allows a partition of explained variation relative to the total amount of variation. This will also allow for significance testing.

Hypothesis testing is possible with RDA, because the canonical axes (eigenvectors) are orthogonal to each other (Borcard et al. 2011). The details of hypothesis testing go beyond the scope of this research, but it is an important concept to know because it allows for testing of spatial correlation in the data set and for testing of fractions in partial pRDA. Permutation F-tests are commonly used in ecological data sets to deal with the complex nature of the data (Legendre & Legendre 1998, Borcard et al. 2011). ter Brakk & Smilauer (2002) constructed an F statistic that can examine fractional analysis via pRDA. This analysis is done in the presence of a second explanatory matrix of spatial coordinates (Legendre 2011 et al.). These methods can bring insight into whether or not spatial variables are related growth variables of trees. Once it is determined that spatial relationships exists it would be logical to address how they are manifested. This insight can be gained by cluster analysis.

Cluster analysis is the discovery of groups in data (Everitt et al. 2011). This is a very simple way to think about an incredibly diverse subject. Legendre & Legendre (1998) explain that clustering is a family of techniques that are undergoing rapid development. This development is in large part due to that fact that clustering is not a traditional statistical method in that there are no formal procedures to follow or hypotheses to test. There are several examples in the literature of cluster analysis with forest data, but a notable lack of literature pertaining to regular spaced, managed forest. In the 1980's, a pulse of forestry literature utilizing cluster analysis emerged. Lorimer (1985) described how cluster analysis could be used to improve his sampling design to better understand disturbance history in the forest. Guevara et al. (1986) used cluster analysis to examine the spatial analysis of forest succession and habitat patterns for bird species. Applications of cluster analysis stayed consistent through the years with focus on forest patch diversity, canopy patterns, or disturbance patterns (see Fraver 1994, Oliveira-Filho et al. 2000, Plotkin et al. 2002, Steane et al. 2006). These applications are appropriate as spatial heterogeneity of populations and communities play a central role in many ecological theories, in particular theories of succession, adaption, maintenance of species diversity, community stability, competition, parasitism, epidemics, and natural catastrophes (Legendre & Fortin 1998). There are a great deal of algorithms and methods of clustering so, for brevity, we describe our general approach to clustering by introducing some terms:

Sequential- This algorithm works by applying a recurrent sequence of operations to the objects (trees).

Agglomerative- Agglomerative procedures begin with all objects being considered separate from one and another. This method successively groups the objects into larger and larger clusters until a single all encompassing cluster is obtained.

Hierarchical- Hierarchical methods allocate members of inferior ranking clusters to larger, higher ranking clusters. Most of the time and in our case, this method will produce non overlapping clusters.

Non-Probabilistic- do not use parametric or non-parametric methods for estimating density functions in multivariate space. There are no probabilities linked with the association matrices.

The Lance & William general model (Borcard et al. 2011) for clustering encompasses many agglomerative methods and is easy translatable to many computer packages (Legendre & Legendre 1998). The Lance & Williams algorithm is also appropriate for constrained cluster analysis when the objects in a cluster are contiguous (Legendre & Legendre 1998). This method is based on a similarity matrix whose elements are values that describe how related two objects are. The Lance & Williams algorithm is known as a combinatorial agglomerative method because it utilizes a similarity matrix (Legendre &

Legendre 1998). This clustering procedure is used widely for classification of landscape and cover type (see Drewa et al. 2002, Urban et al. 2002, Snelder et al. 2004, Perrin et al. 2006, Divíšek et al. 2014). Through the clustering process and analysis we will be able to develop maps of spatial correlation (Legendre et al. 2009, Legendre 2012).

Currently, there are several criteria to determine the appropriate number of clusters. Among these are the adjusted R^2, AIC, and cross validation residual error (CVRE). These criteria are probably not going to be very effective (pers. comm. Pierre Legendre September 8th 2014). The most informative criterions will probably be the Calinski-Harabasz statistic (CH) and the number of groups in the pruned tree (PT) (pers. comm. Pierre Legendre September 8th 2014). There are many ways to select the number of clusters but it is most important that cross validation be preformed. This can be statistical (CH and PT) or nonstatistical such as visual inspection of ordination graphs (Legendre & Legendre 1998).

This investigation will use current techniques of pRDA and constrained cluster analysis to explore how spatial variables explain structure in a managed regular spaced plantation. We will test for spatial relationships in the data and then explore how those spatial relationships are manifested into spatially recognizable structures. The objectives of this research are to measure, test, and map spatial variability in forest plots. This entails:

1. Simulate stands with varying levels of within plot heterogeneity.
    a. Test effectiveness of pRDA to detect no spatial correlations in a control plot with no microsites. This will be measured by success or failure to detect spatial relationship.
        i. Simulate 100 times at each level of variability

        ii. Estimate probability of detecting no spatial pattern

b. Test effectiveness of constrained cluster analysis to detect number of microsites (*k*) of our specified spatial patterns with known *k*. This will be measured as a success or failure to detect correct number of groups.

    i. Simulate 100 times for each combination of variability and difference in mean.

    ii. Estimate probability of successful detection by each criterion.

c. Preform a misclassification test to estimate percent of trees correctly allocated to its each microsite when using a specified criterion.

    i. Estimate probability of misclassification by each criterion.

Methods

1. Simulation

Sites representing 625 trees (25x25) were simulated. Within each site, patterns of microsites were embedded. The simulations replicated 5 sites with distinct spatial patterns (figure 1), each with different microsites. Response and explanatory variables are generated randomly. The response variables are diameter at breast height (DBH) and total tree height (H). The DBH values are generated separately for each microsite. R software is used to draw randomly from a normal distribution of a specified mean DBH and coefficient of variation (CV) for each microsite. For *K* microsites we have $\mu_1 < \mu_2 < \cdots < \mu_K$, where μ_K is the mean DBH for the jth microsites (subplot). For each simulation the CV was equal among all microsites. That is, $DBH_{ijk} \sim N(\mu_j, \sigma^2)$, where σ^2 is chosen to achieve a prespecified value of $CV = (\sigma \div \mu) * (100\%)$. Heights are calculated based on known diameter-height relationships. Equation (1) from Sabatia and Burkhart (2013) can be used to model heights.

$$H_{ijk} = \beta_0 e^{\beta_1 DBH_{ijk}^{-1}} + \theta_{ijk} \qquad (1)$$

The parameter estimates from Sabatia and Burkhart (2013) are from a loblolly pine study of similar planting conditions. Where H_{ijk} is the total height and DBH_{ijk} is the diameter at breast height of the *k*th tree in the *j*th microsite of the *i*th simulation. The index values range from, *i*=1…100, *j*=1…*k*, and *k*=1…K β_0 is the upper asymptote parameter and β_1 is the rate parameter, and θ_{ijk} is is the random error due to the *k*th tree

$[\theta_{ijk} \sim N(0, \sigma_\theta^2)]$ (Sabatia & Burkhart 2013). The estimated equation from Sabatia and Burkhart (2013) is then:

$$H_{ijk} = 20.382 e^{-7.309 DBH_{ijk}^{-1}} + \theta_{ijk}\ ,\ \theta_{ijk} \sim N(0, 0.482) \tag{2}$$

We now have our response matrix $\mathbf{Y_{ijk}}$=[$\mathrm{DBH}_{\mathrm{ijk}}$, $\mathrm{H}_{\mathrm{ijk}}$]. Next, the environmental variables were determined.

We used site index (SI) as a proxy variable to measure soil productivity. Each microsite has a unique site index value that relates to its productivity. The less productive microsites have smaller SI values. We assigned SI values to each microsite within each spatial pattern. The SI values remained fixed for each microsite throughout the simulation process. The SI values were generated randomly from R, $\mathrm{SI}_{\mathrm{ijk}} \sim \mathrm{N}(\mu_j, \sigma_j^2)$ where j represents the microsite *j*=1…k and where σ_j^2 satisfies the conditions $CV_j = (\sigma_j \div \mu_j) * (100\%)$. The mean SI for each microsite is positively related to the mean DBH of the microsite. Increasing values of mean SI result in increased values of mean DBH. This is because increased values of SI represent higher levels of productivity that produce larger trees. Assigning mean SI values was based on known relationships between H and SI. The specific mean SI is less consequential for this analysis. What is important is that the SI values are higher for more productive sites. These SI values represent our matrix of environmental variables (**X**). For *k* microsites we simulate *k* SI distributions.

A second explanatory matrix of spatial variables is built, where **W**=[$\mathrm{X}_{\mathrm{cord}}$ $\mathrm{Y}_{\mathrm{cord}}$]. We are simulating a managed plantation stand so values of **W** remain fixed grid points, where

X_{cord} Y_{cord} are grid points measured in feet. The next step is to allocate value to spatial patterns. The designed spatial patterns are:

1. Control – Stand structures are spatially homogenous. Response variables are generated from the same distribution.
2. Biplot – The whole plot is divided into two equal microsites. Response variables and explanatory variables are drawn from distinct distributions to reflect productivity of each microsite.
3. Triplot - The whole plot is divided into three equal microsites. Response variables and explanatory variables are drawn from distinct distributions to reflect productivity of each microsite.
4. Quadplot - The whole plot is divided into four equal microsites. Response variables and explanatory variables are drawn from distinct distributions to reflect productivity of each microsite.
5. Free plot – Five microsites are created by hand. Microsites are irregular sizes and shapes. Response variables and explanatory variables are drawn from distinct distributions to reflect productivity of each microsite.

For each spatial pattern (figure 1) we varied differences in mean DBH among the microsites and plot CV. The reason for this is to study how our spatial analysis will handle different spatial structures. All four spatial patterns (excluding the control) were simulated following the parameters of Table 1. This means that for each spatial pattern eight different parameters were used for the simulation. Each cell in table 1 represents assigned differences among microsites.

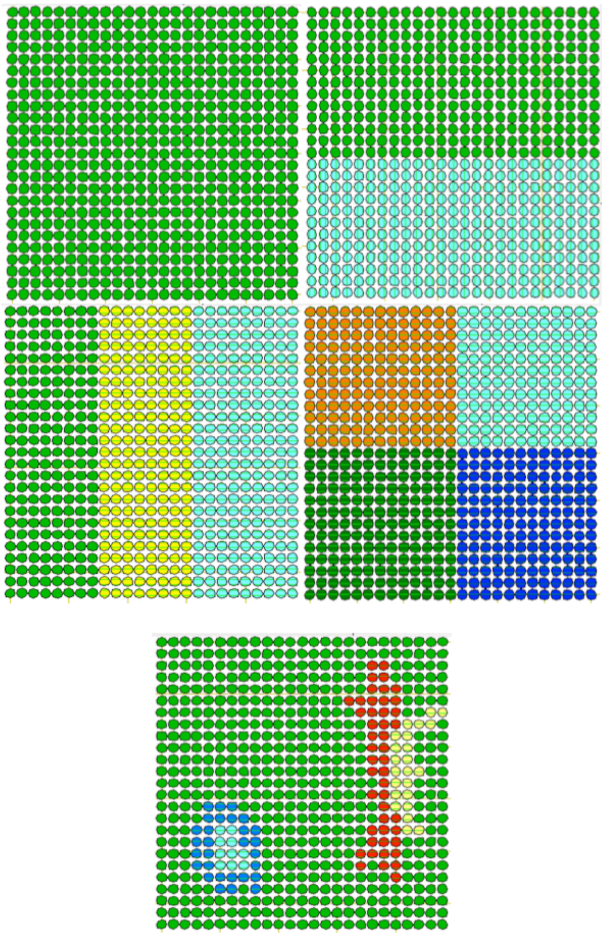

Figure 1 The five spatial patterns used in the simulations. From left to right: Control (k=1), Biplot (k=2), Triplot (k=3), Quadplot (k=4), and Free plot (k=5).

Using the triplot as an example, row two and column two has (2,10%), so for this simulation microsites differed in mean DBH values successively by 2 inches and each microsite had a coefficient of variation of 10%. Three sets of DBH values were then generated randomly using R and these were $DBH_1 \sim N(5, .25)$, $DBH_2 \sim N(7, .49)$, $DBH_3 \sim N(9, .81)$. Using equation (2) heights were then calculated and each of these values were assigned to a spatial coordinate within its appropriate microsite. The appropriate microsite was determined by its mean SI value, so the matches were SI=70 with $DBH_1 \sim N(5, .25)$, SI=75 with $DBH_2 \sim N(7, .49)$, and SI=80 with $DBH_3 \sim N(9, .81)$. Microsite number was assigned based on when it first appeared in the tree order. Tree 1 is located on the upper left corner of the plot, tree 2 and 3 follow from left to right of the row, then down each row successively. This continues down to tree 625 that is located in the bottom right corner of the plot. Using the free plot as an example, microsite 1 is green (tree 1), microsite 2 is red (starts at tree 69), microsite 3 is yellow (starts at tree 174), up to microsite 5 that is light blue (starts at tree 406). Each spatial pattern with k>1 underwent 100 simulations for each cell in table 1. The control plot (k=1) underwent 1000 simulations for the four levels of CV. The control plot was tested with partial redundancy analysis and not clustering analysis so it was much more time efficient.

Table 1 Shows the difference in mean DBH and CV for microsites.

		Coefficient of Variation			
		5%	10%	15%	20%
Difference in Mean DBH	1	(1, 5%)	(1, 10%)	(1, 15%)	(1, 20%)
	2	(2, 5%)	(2, 10%)	(2, 15%)	(2, 20%)

The simulations were all run using the R software. We used the “vegan” and the “const.clust” packages in R to analyze the data and “nb2listw(tri2nb())” function to apply spatial constraints. Constrained clustering methods take into account more information than other types of clustering. In our analysis spatial information is used to build clusters and to make results more interpretable. For spatial contiguity, the only admissible clusters are those that obey a contiguity relationship (Legendre & Legendre 1998). Spatial contiguity is described by a connection scheme. We spatially constrained the cluster analysis by Delaunay triangulation methods (discussed later). The spatial constraint forces the clusters to be restricted the same way a microsite is restricted-spatially. This is a more interpretable form of cluster analysis in our case because we want our cluster map to mirror our spatial patterns (map of known microsites). A microsite is a spatially constrained patch, containing two or more trees, that has dissimilar site productivity from the surrounding area. Two microsites that have the same productivity but are not spatially connected are considered to be two distinct microsites and therefore we would expect 2 distinct clusters to identify them.

For each run of the constrained cluster simulation, we examined the selection criteria for c=2,…,10 clusters in the data. The sites can be investigated for c=2,…,n-1 clusters in the data, but for our analysis c=2,…,10 is sufficient. We then saved the number of clusters estimated by each criterion. For example, when using the CH statistic as a criterion we choose the number of clusters with the highest CH value (higher values are better). The number of clusters picked by the CH was then stored into a new matrix with the other “best” selections by PT, AIC, R, and CVRE. We are using our cluster maps to estimate locations of microsites (our known spatial pattern). Since the microsites are known and generated by us, we can compare the effectiveness of the cluster maps to estimate or number of microsites.

After completing the constrained cluster simulations, we computed the probability of successful (POS) detection of known microsites for each spatial pattern by each criterion. This was done by counting the number of times each criterion correctly identified the known number of microsites for each simulation. Each time the criterion was correct we counted 1, otherwise we counted zero. This was done a hundred times for each simulation (n=100). The POS ($\hat{p}$) is the probability a constrained clustering criterion will correctly determine the number of microsites in the data. The calculations of POS are a series of Bernoulli trials:

$x_i = \begin{cases} 1 \; if \; correctly \; identified \; by \; criterion \\ 0 \; otherwise \end{cases}$ and $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$

For all POS values that exceeded 80% we ran misclassification simulations. The misclassification simulations would examine how each tree is allocated based on the criterion. For a single run of "const.clust" trees were all assigned a cluster based on the criterion. If the assigned cluster of each tree matched the known microsite then it was a successful grouping. If the assigned cluster did not match the known microsite then the tree was misclassified. Ideally, cluster arrangements mirror microsite arrangements. Each misclassification simulation was run 50 times. At the end of the simulation the number of misclassified trees were summed and divided by the total number of trees involved. The resulting value was the probability of misclassification by criterion. The misclassification simulations estimated the probability of misclassification by criterion.

2. Analysis

All trees in the control plot come from the same distribution. Therefore the spatial structure should be homogenous. To test spatial homogeneity we use the redundancy equation (3).

$$\left(\boldsymbol{S}_{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}} - \boldsymbol{\lambda}\mathbf{I}\right)\boldsymbol{u} = \mathbf{0} \tag{3}$$

Where $\boldsymbol{S}_{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}} = \left[\frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}}\right]\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{S}_{\boldsymbol{YX}}\boldsymbol{S}_{\boldsymbol{XX}}^{-\mathbf{1}}\boldsymbol{S}'_{\boldsymbol{YX}}$. Here $\boldsymbol{S}_{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}$ is the covariance matrix corresponding to fitted values $\hat{Y}$, where $\hat{Y}$ is the centered matrix of the estimated response, $\boldsymbol{\lambda}$ is the vector of eigenvalues of the covariance matrix, $\boldsymbol{u}$ is the matrix of eigenvectors of the covariance matrix, and **I** is an identity matrix. The steps below use our generated data **X** and **W.** Where **X** is a matrix of environmental variables (SI) and **W** is a matrix of spatial coordinates, **W**=[X_{cord} Y_{cord}]. We want to partition the variation in Y such that we can explain the total variation through the sum of different fractions. These are fraction (a) that represents the variation explained by the environmental variables. Fraction (b) represents the variation explained by the confounded variation of the environmental variables and spatial variables. Fraction (c) represents the variation explained by the spatial variables. The last fraction (d) is the residual variation not explained by the other components. Figure 2 illustrates the different fractions. We run the following steps to partition the variation:

1. Run an RDA of the response data **Y** by **X**. This yields the first fraction [a+b].
2. Run an RDA of the response data **Y** by **W**. This yields fraction [b+c].
3. Run an RDA of the response data **Y** by **X** and **W** together. This gives fraction [a+b+c].
4. Compute the R^2_{adj} of the three RDA's above:

$$R^2_{adj}=1-\frac{n-1}{n-m-1}(1-R^2) \tag{4}$$

   where *n* is the number of objects and *m* is the number of explanatory variables.

5. Compute the fractions of adjusted variation by subtraction:

   For example, fraction [a]=[a+b+c] – [b+c]. Repeat for fraction [b] and [c].

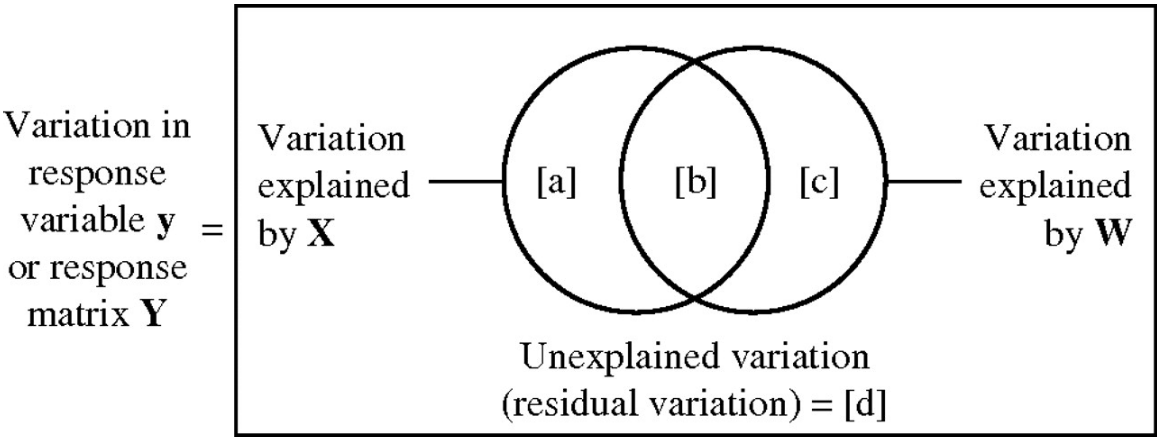
Variation in response variable **y** or response matrix **Y**
=
Variation explained by **X**
[a]
[b]
[c]
Variation explained by **W**
Unexplained variation (residual variation) = [d]

Figure 2 Illustrates the partitioning method. (Legendre & Legendre 1998).

Next, we tested the effectiveness of pRDA to detect spatial relationships in a control plot. We expect to accept the null hypothesis of no spatial relationship. We iterated this process 1000 times and measured its success rate. In our case, we considered it a success to accept the null hypothesis. Here, we used the permutation F test (Borcard et al. 2011) to test significance of explanatory and spatial components. Since we are most interested in fraction [c], we will show the process for it here. We use an ANOVA like F test (eq. 5) to investigate the effectiveness of **W** on explaining the variation in **Y**.

$$F = \frac{\sum_1^p \lambda_i}{RSS/(df)} \tag{5}$$

The numerator is the contribution to the variance of **Y** from **W** after removing the contribution of **X**. The denominator (RSS) is the sum of the unconstrained eigenvalues (fraction [d] of figure 2) and *df* are the degrees of freedom. The next step is to test the success of constrained clustering on maps with spatial heterogeneity.

To analyze spatial structures we implement spatially constrained cluster analysis using the R software package "const.clust". The purpose of this analysis was to identify spatially recognizable structures in tree growth of our known maps. For this procedure we had to specify which distance metric we would use and which connection network.

Distance metrics are used to measure the association between two objects (trees). The smaller a distance value or closer it is to zero the more related object (trees) are structurally. In our data, two trees that are identical would have a distance value of 0. The most common metric measure is the Euclidean distance (Legendre & Legendre 1998). We used Euclidean distances (eq. 6) among objects using non-geographic information to create our dissimilarity matrix (**D**).

$$D(y_r, y_{r+1}) = \sqrt{\sum_{c=1}^{p}(y_{rc} - y_{(r+1)c})^2} \qquad (6)$$

For equation (6), *r*=row of matrix **Y**, *c*=column of matrix **Y**, and *p* is the number of variables in matrix **Y**. For one of our simulations, *r*=1…625, and *p*=2. This step is typical in many clustering algorithms, but in the next steps we impose spatial constraints on the dissimilarity matrix (Figure 3), which is information typically not incorporated into clustering analysis.

Prior to preforming spatially constrained clustering it is important to state which trees are neighbors in space. In order for a tree to enter a cluster it has to be a neighbor to it in space. The only admissible clusters in a spatially constrained analysis are those that obey the contiguity scheme. We relate clusters to microsites by constraining clusters so they are spatially defined in the same way as a microsite. A cluster is then a contiguous patch or group of trees that are structurally unlike the rest of the trees. Microsites create spatially recognizable structures in tree growth due to differences in productivity. This is the link from clusters to microsite and because of this we expect cluster location to parallel microsite location. Microsites create the structural differences which clustering recognizes.

The Delaunay triangulation uses spatial coordinates to identify neighbors. This is how we define contiguity. To determine neighbors we produce a list of connection edges to create a contiguity matrix containing 1's for connected and 0's elsewhere (based on spatial coordinates of plot map). The contiguity matrix is how we spatially constrain our cluster analysis. The 1's and 0's are how we define neighbors and create connections among the trees. The Delaunay triangulation method states that for any triplet of non-collinear points A, B, and C the three edges connecting these points are included if and only if the circle passing through these points (figure 3) include no other point (Legendre & Legendre 1998). This criterion is a robust method for defining contiguity. This connection scheme works well with regular grids and is adaptable to various patterns of planting grids and will transfer well to real plots that are slightly irregular.

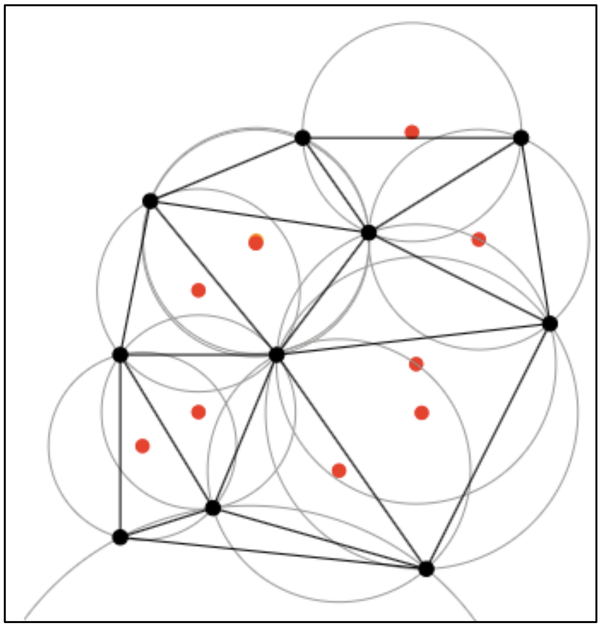

Figure 3 Illustrates the circles used in Delaunay triangulation. Black dots represent the objects (trees in our data), red dots represent the center of each circle used to circumscribe three points, and the thick black lines connect "neighbors" (Wikipedia, November 3rd, 2014).

The spatial constraint allows only connected trees to be clustered together. This prevents a scattering of cluster assignments on the map. Instead, clusters form distinct clumps. The cluster analysis results can be mapped with the spatial coordinates of the trees. The resulting map shows cluster assignment of each tree. When compared to our map of known microsites, we expect clusters to form over microsites and for trees within a microsite to be assigned the same cluster number.

Figure 4 illustrates the general framework for how a dissimilarity matrix interacts with the contiguity matrix to create a spatially constrained dissimilarity matrix suitable for constrained clustering. The Hadamard product between the dissimilarity matrix and contiguity matrix creates a constrained dissimilarity matrix where distance values exist only where neighbors were previously defined by the contiguity matrix. Our data file consists of growth information on 625 trees. A 625x625 dissimilarity matrix is created from equations (6). The more akin any two trees are in structure the closer to 0 is their dissimilarity value. We then create a 625x625 contiguity matrix of 1's and 0's where 1's mark neighbors as defined by Delaunay triangulation and 0's elsewhere. The Hadamard product for our data is the dissimilarity in growth among neighboring trees.

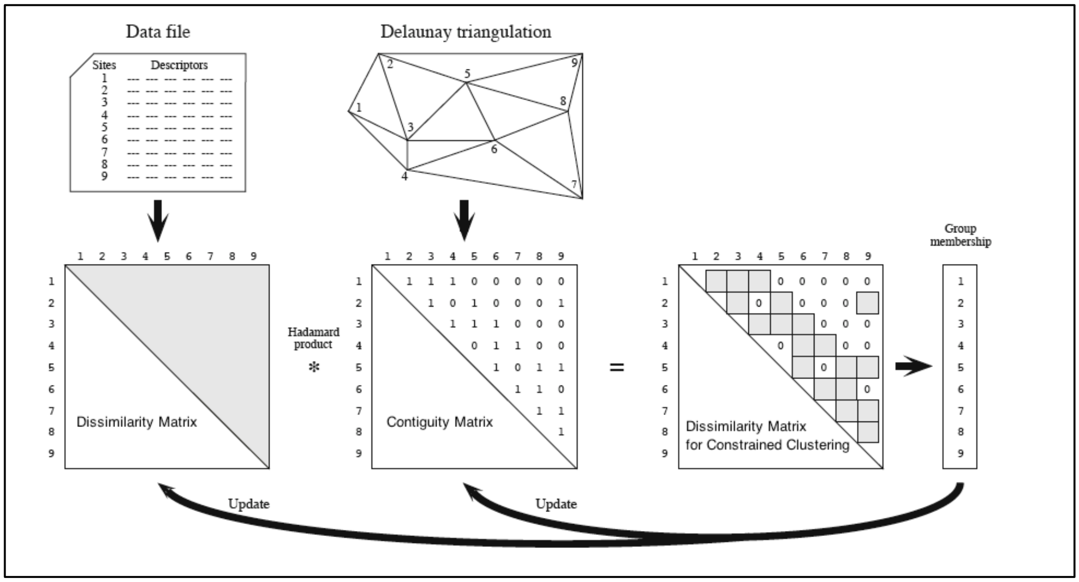

Figure 4 illustrates how spatial constraints are imposed in the clustering process (modified from Legendre & Legendre 1998)**.**

We iterate this process 100 times for each combination of spatial pattern and parameter values. After the iteration process is completed we can examine how successful the criteria were at detecting spatial patterns. We explored the probability of successfully allocating a tree to the correct microsite. The cluster map should reflect the microsite map if each tree is assigned correctly from the clustering algorithm. This is a simple process where we count 1 if tree *i* is correctly grouped into a cluster and 0 otherwise. Probability of success is *p* and *n=625* is the number of trees per simulation. By the central limit theorem:

$x_i = \begin{cases} 1 \text{ } if\ correctly\ grouped \\ 0 \text{ } otherwise \end{cases}$ and $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$. Therefore $\hat{p} \sim N(p, \frac{p(1-p)}{n})$.

We can then compute the confidence interval: $\hat{p} \pm z_{\propto/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Results

Simulations for the control plot indicated that pRDA is an effective method for testing of spatial homogeneity. For these simulations a “success” was considered to be those situations where the F test (equation 5) failed to reject the null hypothesis of no linear relationship between the response and spatial coordinates. We found a high success rate at each level of CV. Table 2 shows the success rates for our simulations.

Table 2 Shows the probability for detection of spatial homogeneity for the control plot.

Coefficient of Variation	Probability of Success
5%	0.95
10%	0.96
15%	0.96
20%	0.94
25%	0.95

The CH statistic and PT were the only two criteria to successful detect the correct number of microsites. The CH statistic is an F statistic comparing the among cluster sum of squares to the within cluster sum of squares (Borcard et al. 2011). The PT is a cross validation procedure that determines the best number of groups based on a relative error ratio of the dispersion unexplained by the cluster tree divided by the overall dispersion of the response data (Borcard et al. 2011). For evaluation, the algorithm produces a specified range of clusters to evaluate. We examined 2 through 10 clusters for each simulation. For each number of clusters we would get a value from each criterion. The number of clusters corresponding to the highest CH statistic is best, or in the case of the PT it was the cluster number with the lowest relative error ratio. The CH statistic was the most successful in that it detected the correct number of microsites most often and when the PT detection rate was high (>80%) the CH statistic was still the more successful. With this in mind, we are reporting probability of successful detection of both the CH and PT, but only the misclassification rates of the CH statistic.

There were two main expectations. The first is that less complex spatial patterns (biplot being the least and free plot the most complex) would be correctly detected more often than the complex patterns. The second was that for both the 1 inch and 2 inch difference in mean DBH, probability of detection would drop with increased amounts of variation.

As expected, the biplot was consistently detected the most by the CH statistic and the free plot the least (figures 5 & 6). Interestingly, the triplot and quadplot alternate in their relative success between figures 5 & 6. The low probability of successful detection using the PT was unexpected (figures 7 & 8). The more complex patterns were generally detected more often than the less complex spatial patterns. The PT criterion was not successful in detecting the biplot in all scenarios, but was relatively successful with the quadplot.

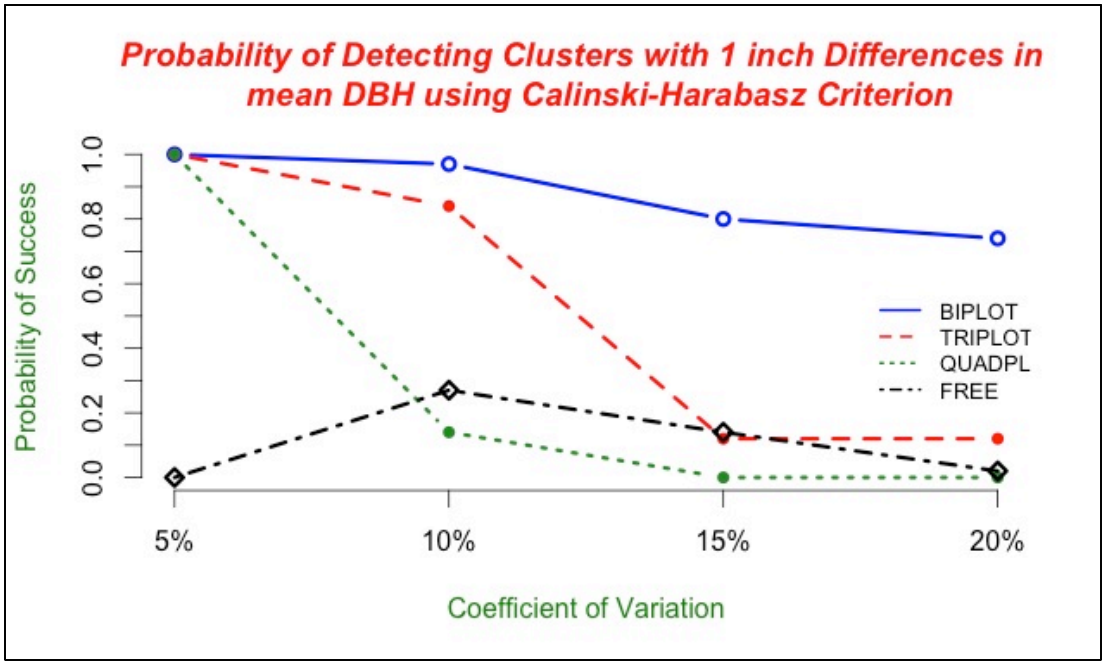

Figure 5 Illustrates the probability of successful detection by the CH statistic when the difference in successive mean DBH values between k groups is 1 inch.

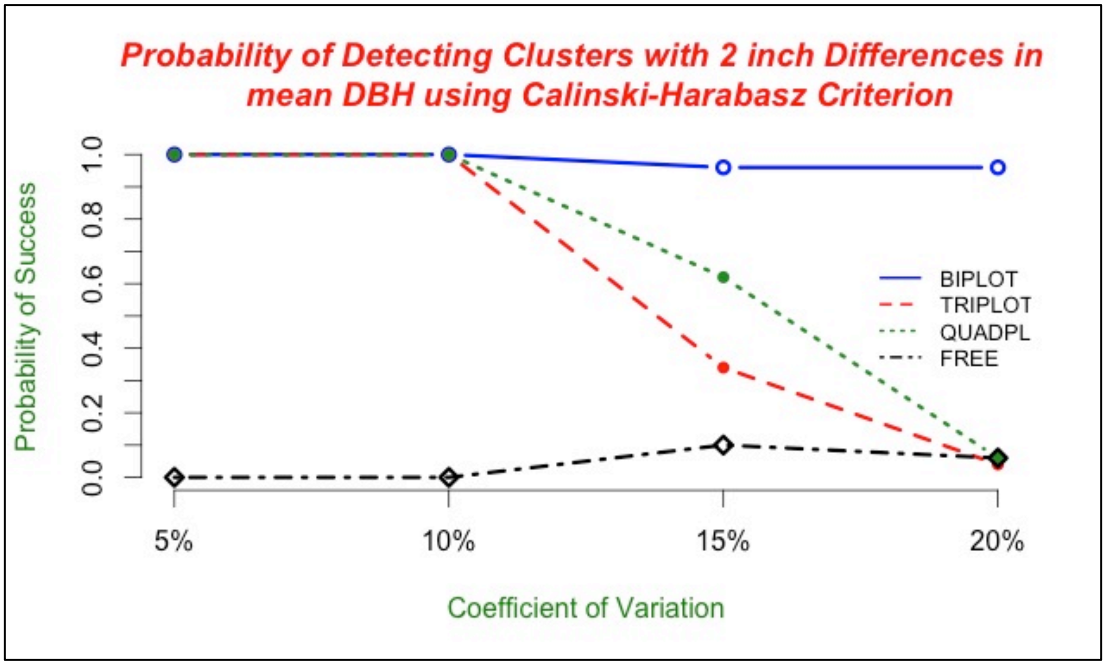

Figure 6 Illustrates the probability of successful detection by the CH statistic when the difference in successive mean DBH values between k groups is 2 inch.

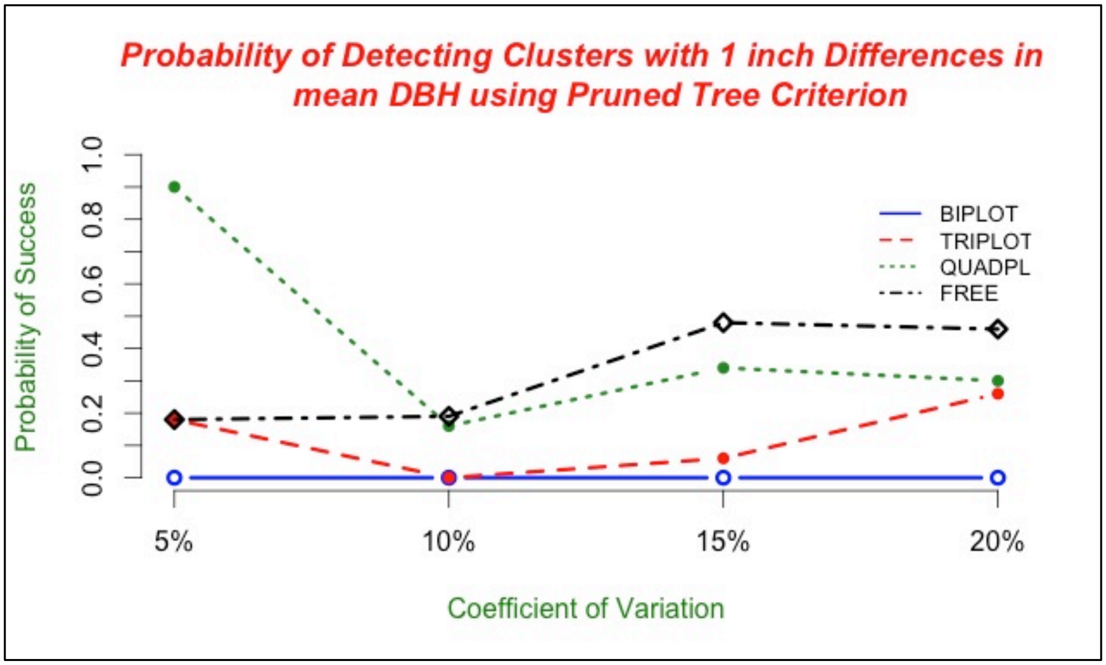

Figure 7 Illustrates the probability of successful detection by the PT when the difference in successive mean DBH values between k groups is 1 inch.

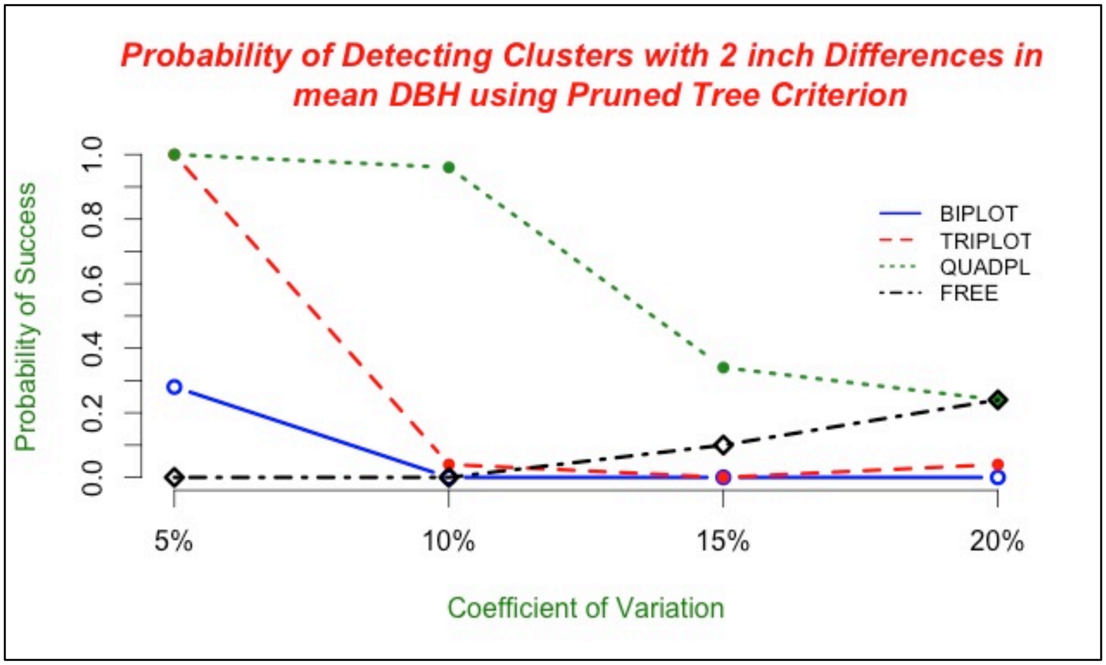

Figure 8 Illustrates the probability of successful detection by the PT when the difference in mean DBH values between successive k groups is 2 inch.

Table 3 shows the 95% confidence intervals for each estimate of the POS for the correct number of clusters. The values of *p* range from 0 to 1. It is apparent that CH values are consistently more useful. This can most easily be noticed by examining successive rows. The rows of the table alternate from the CH to the PT.

Table 3 95% confidence intervals for the probability of successful detection of the number of microsites.

Coefficient of Variation	*BIPLOT*	*TRIPLOT*	*QUADPLOT*	*FREE*	*Criterion*	*Mean Difference*
5%	1,1	1,1	1,1	0,0	CH	1
5%	0,0	.10,.26	.84,.96	.1,.26	PT	1
5%	1,1	1,1	1,1	0,0	CH	2
5%	.19,.37	1,1	1,1	0,0	PT	2
10%	.94,1	.77,.91	.07,.21	.18,.36	CH	1
10%	0,0	0,0	.09,.23	.11,.27	PT	1
10%	1,1	1,1	1,1	0,0	CH	2
10%	0,0	0,.08	.92,.99	0,0	PT	2
15%	.72.88	.06,.13	0,0	.07,.21	CH	1
15%	0,0	.01,.11	.25,.43	.38,.58	PT	1
15%	.92.99	.25,.43	.52,.72	.04,.16	CH	2
15%	0,0	0,0	.52,.72	.04,.16	PT	2
20%	.65,.83	.06,.18	0,0	0,.05	CH	1
20%	0,0	.17,.35	.21.39	.36,.56	PT	1
20%	.90,1	0,.08	.01,.11	.01,.11	CH	2
20%	0,0	0,.08	.16,.32	.16,.32	PT	2

The CH statistic was consistently the best criterion. Since the CH was the best choice for selection criterion we preformed a misclassification test on all situations were the successful detection by the CH was above 80% (our threshold for an adequate success rate). The interpretation of the values in tables 4 and 5 are the probability that a tree will be correctly allocated when using the CH statistic. This is not a conditional probability (i.e. given CH identified 2 microsites). The misclassification rates were generally expected with the probability of misclassification generally increasing from left to right and from top to bottom. Range of misclassification was 0% for the biplot, triplot, and quadplot at 5% CV and 2 inch difference to 16% for the biplot at 15% CV and 1inch difference.

Table 4 Shows the probability of tree misclassification when using the CH statistic at a mean difference in successive DBH values at 1inch.

Coefficient of Variation	*BIPLOT*	*TRIPLOT*	*QUADPLOT*
5%	0.004	0.012	0.009
10%	0.080	0.276	*
15%	0.159	*	*
20%	*	*	*

Table 5 Shows the probability of tree misclassification when using the CH statistic at a mean difference in successive DBH values at 2inch.

Coefficient of Variation	*BIPLOT*	*TRIPLOT*	*QUADPLOT*
5%	0.000	0.000	0.000
10%	0.007	0.013	0.036
15%	0.084	*	*
20%	0.136	*	*

Discussion

The POS for the free draw pattern is anomalous. For both the PT and CH statistics, the POS of the free draw pattern seems to peak at approximately 15% CV. What is most peculiar is that the POS appears to rise from 5% to about 15% CV, like the other patterns we expect a decrease in POS from low to high values of CV. There are three distinct features about the free plot that are plausible explanations for this occurrence. First is that the free plot has irregular shaped microsites. The irregular pattern of the microsites can influence which trees are usurped into a cluster. Connection schemes are carefully chosen before clustering is done in order to mitigate for possible influences from how objects (trees) and groups (microsites) are spatially dispersed. As described earlier, we choose Delaunay triangulation for our connection scheme that provides at least 2 neighbors for each tree in our simulated stand. Based on review of our contiguity matrix this seems an unlikely cause of free plot POS behavior. Possible neighbors include within and among microsite trees. Even for the smallest microsites, trees were neighbors (spatially constrained) with other trees in their microsite. Second, the free plot has a fifth microsite and is the most complex stand we simulate. This is a cause for change in the POS, but not a factor that will cause the POS to rise from low to high CV values. Like the other patterns, as we add an additional microsite we see a general decrease in POS compared to the previous less complex pattern. This too is an unlikely candidate. Third, the five microsites are all different sizes. Size is not always equal among microsites due to the odd number of rows and columns in our stand, but up to the free plot they were as close as possible. The number of trees per microsite in the free plot ranged from 10 trees to 506 trees. This means that the five distributions for each microsite have a varying number of trees, as well as different means and variances. It is likely that to minimize the CH statistic and the CVRE (the statistic minimized for the PT) trees from other distributions were simply split or taken into other distributions. It is most likely that the erratic POS across the CV values for the free plot were due to the large range in sizes of each microsite. This could have likely caused an overplotting effect that creates uncertainty in cluster assignment due to ranges in microsite values. The erratic behavior of the POS for the free plot is subject for further investigation and may require additional simulations and analysis.

Cluster analysis is difficult to validate. In some instances, mostly with the biplot, we measured a success rate of 100%. Although we do not know the spatial patterns in practice, we are still able to apply our algorithm. Through our simulations we were able to mimic real situations and gather information that will allow us to make more informed decisions. Given our findings, the CH statistic in our opinion is the best choice. There are a few interesting topics to mention pertaining to our results. These topics are applications to real data, validating with multiple methods, and extensions of this work.

Forest data is inherently complex. There are a plethora of variables that can alter growth. These variables include soil chemical reactions to stochastic weather including ice storms and lightning. We have known that productivity varies at very fine scales as a result of many processes. When we examine a forest plot we do not know how many groups are in the data so we can follow two procedures. First, we would want to know if there is a spatial component to our data, we could use pRDA to test for this. If there is a significant spatial relationship we can proceed to using cluster analysis. Spatially constrained cluster analysis is well suited for describing its structure and the CH statistic is the best criterion to follow. Once the microsites are located it is important to verify the microsites by other means. This reduces the chance of error. Once we locate the microsites there are a variety of methods to check if we have reliable results.

In order to check the validity of clustering results we can implement different types of clustering. For example, K means clustering and constrained clustering can be used together. If the constrained clustering indicates that 2 microsites are present in the data then you would expect another clustering method to come to similar results if there are in fact 2 microsites in the data.

Clustering analysis has some promising applications in forestry. Even close approximations of structural differences may give foresters better ideas of how to apply expensive fertilizer and herbicidal treatments. It is fully expected that refinements in clustering techniques will improve management by foresters on the ground. Further investigation will need to ultimately be done. This includes modeling trees with software that includes complex competition interactions in estimating growth. Also, these methods will need to be compared with fine scale soil maps that measure productivity between trees.

There is further work that needs to be done, but this is certainly a first step in developing a richer understanding of growth dynamics of forest plots. For the first time we have information on how to more effectively measure clusters in a forest plot. Our investigation indicates that the CH statistic is best suited for cluster analysis in forestry applications.

References

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, *73*(3), 1045-1055.

Borcard, D., Gillet, F., & Legendre, P. (2011). *Numerical ecology with R*. Springer.

ter Braak, C. J., & Prentice, I. C. (1988). A theory of gradient analysis. *Advances in ecological research*, *18*, 271-317.
Chicago

ter Braak, C. T., & Šmilauer, P. (2002). CANOCO reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (version 4.5). *Section on Permutation Methods. Microcomputer Power, Ithaca, New York*.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs*, *27*(4), 325-349.

Brown, R. T., & Curtis, J. T. (1952). The upland conifer-hardwood forests of northern Wisconsin. *Ecological Monographs*, 217-234.

Curtis, J. T., & McIntosh, R. P. (1951). An upland forest continuum in the prairie-forest border region of Wisconsin. *Ecology*, *32*(3), 476-496.

Divíšek, J., Chytrý, M., Grulich, V., & Poláková, L. (2014). Landscape classification of the Czech Republic based on the distribution of natural habitats. *Preslia* 86, 209-231.

Drewa, P. B., Platt, W. J., & Moser, E. B. (2002). Community structure along elevation gradients in headwater regions of longleaf pine savannas. *Plant Ecology*, *160*(1), 61-78.
Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. Wiley Series in Probability and Statistics.

Fraver, S. (1994). Vegetation Responses along Edge‐to‐Interior Gradients in the Mixed Hardwood Forests of the Roanoke River Basin, North Carolina.*Conservation Biology*, *8*(3), 822-832.

Gittins, R. (1985). Canonical correlations and canonical variates. In *Canonical Analysis* (pp. 13-36). Springer Berlin Heidelberg.

Goodall, D. W. (1954). Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. *Australian Journal of Botany*, *2*(3), 304-324.

Greig-Smith, P., Austin, M. P., & Whitmore, T. C. (1967). The Application of Quantitative Methods to Vegetation Survey: I. Association-Analysis and Principal Component Ordination of Rain Forest. *The Journal of Ecology*, 483-503.

Grimaldi M, Oszwald J, Dole´dec S, Hurtado MP, Miranda IS, de Sartre XA, de Assis WS, Castan˜eda E, Desjardins T, Dubs F, Guevara E, Gond V, Lima TTS, Marichal R, Michelotti F, Mitja D, Noronha NC, Oliveira MND, Ramirez B, Rodriguez G, Sarrazin M, da Silva Jr ML. (2014). Ecosystem services of regulation and support in Amazonian pioneer fronts: searching for landscape drivers. *Landscape Ecology*, *29*(2), 311-328.

Guevara, S., Purata, S. E., & Van der Maarel, E. (1986). The role of remnant forest trees in tropical secondary succession. *Vegetatio*, *66*(2), 77-84.

Hallmann, F. W., & Amacher, G. S. (2012). Forest Bioenergy adoption for a risk-averse landowner under uncertain emerging biomass market. *Natural Resource Modeling*, *25*(3), 482-510.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological monographs*, *54*(2), 187-211.

Lahti, T., & Väisänen, R. A. (1987). Ecological gradients of boreal forests in South Finland: an ordination test of Cajander's forest site type theory. *Vegetatio*, *68*(3), 145-156.

Leduc, A., Drapeau, P., Bergeron, Y., & Legendre, P. (1992). Study of spatial components of forest cover using partial Mantel tests and path analysis. *Journal of Vegetation Science*, *3*(1), 69-78.

Legendre, P., & Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, *80*(2), 107-138.

Legendre, P., & Legendre, L. (1998). Numerical ecology: second English edition. *Developments in environmental modelling*, *20*.

Legendre, P., & Legendre, L. F. (2012). *Numerical ecology* (Vol. 20). Elsevier.

Legendre, P., Mi, X., Ren, H., Ma, K., Yu, M., Sun, I. F., & He, F. (2009). Partitioning beta diversity in a subtropical broad-leaved forest of China. *Ecology*, *90*(3), 663-674.

Legendre, P., Oksanen, J., & ter Braak, C. J. (2011). Testing the significance of canonical axes in redundancy analysis. *Methods in Ecology and Evolution*, *2*(3), 269-277.

Lorimer, C. G. (1985). Methodological considerations in the analysis of forest disturbance history. *Canadian Journal of Forest Research*, *15*(1), 200-213.

Martel, N., Rodriguez, M. A., & Berube, P. (2007). Multi‐scale analysis of responses of stream macrobenthos to forestry activities and environmental context. *Freshwater Biology*, *52*(1), 85-97.

Motyka, J., Dobrzanski, B., & Zawadzki, S. (1950). Preliminary studies on meadows in the south-east of Lublin province. *Ann. Univ. Mariae Curie-Sklodowska 5E.*, 367-447.

Oliveira‐Filho, A. T., & Fontes, M. A. L. (2000). Patterns of Floristic Differentiation among Atlantic Forests in Southeastern Brazil and the Influence of Climate1. *Biotropica*, *32*(4b), 793-810.

Oliver, C. D., & Larson, B. C. (1996). Forest stand dynamics: updated edition. *Forest stand dynamics: updated edition.*

Peet, R. K. (1981). Forest vegetation of the Colorado front range. *Vegetatio*, *45*(1), 3-75.

Peres-Neto, P. R., Legendre, P., Dray, S., & Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, *87*(10), 2614-2625.

Perrin, P. M., Martin, J. R., Barron, S. J., & Roche, J. R. (2006, January). A cluster analysis approach to classifying Irish native woodlands. In *Biology & Environment: Proceedings of the Royal Irish Academy* (Vol. 106, No. 3, pp. 261-275). The Royal Irish Academy.

Plotkin, J. B., Chave, J., & Ashton, P. S. (2002). Cluster analysis of spatial patterns in Malaysian tree species. *The American Naturalist*, *160*(5), 629-644.

Poore, M. E. D. (1955). The use of phytosociological methods in ecological investigations: III. Practical application. *The Journal of Ecology*, 606-651.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, 329-358.

Sabatia, C. O., & Burkhart, H. E. (2013). Height and Diameter Relationships and Distributions in Loblolly Pine Stands of Enhanced Genetic Material. *Forest Science*, *59*(3), 278-289.

Steane, D. A., Conod, N., Jones, R. C., Vaillancourt, R. E., & Potts, B. M. (2006). A comparative analysis of population structure of a forest tree, Eucalyptus globulus (Myrtaceae), using microsatellite markers and quantitative traits. *Tree Genetics & Genomes*, *2*(1), 30-38.

Urban, D., Goslee, S., Pierce, K., & Lookingbill, T. (2002). Extending community ecology to landscapes. *Ecoscience*, *9*(2), 200-212.

Van Den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, *42*(2), 207-219.

Vries, D. D. (1952). Objective combinations of species. *Acta botanica neerlandica*, *1*(4), 497-499.

Weber, C. D. (1983). *Height growth patterns in a juvenile Douglas-fir stand, effects of planting site, microtopography and lammas occurrence* (Doctoral dissertation, University of Washington).

Webb, D. A. (1954). Is the classification of plant communities either possible or desirable. *Bot. Tidsskr*, *51*, 362-370.

Delaunay Triangulation. (2014, September 19). Retrieved November 3, 2014, from http://upload.wikimedia.org/wikipedia/commons/1/1f/Delaunay_circumcircles_centers.svg.