WHITEBARK PINE DATA ANALYSIS PLAN

**Preparation for Analyses (Andy’s lab)**

1. Please check a small subset of values for each variable against the original data to ensure that the variables have been appropriately compiled (read, re-formatted, merged, transformed to metric units, etc.) and are error-free.

2. I know that for an earlier dataset we confirmed that all of the variables were in metric units, but please check that all of the variables for the dataset we will actually analyze are in metric units.

3. With the latest dataset having a sample size of 2359 observations, our earlier plan to obtain a 70% / 30% split of the total available sample for the train and test datasets, respectively, is still fine. But the plan to choose some number of variables such that there are at least 10 observations per variable in the train dataset is no longer relevant. If we were to take 70% of 2359, we would get about 1650 observations. If we chose the number of variables such that we would have 10 observations per variable in the candidate set, we would have about 165 variables, which would constitute the classic “fishing expedition.”

4. Using knowledge (from experience/literature) of the ecology of WBP and its competitors, as well as the likely influence of certain physical variables on all tree species, choose explanatory variables that are expected to be highly relevant for a given response variable. List the candidate explanatory variables for each response variable so I know what to include in the model. Identification of this initial set of variables is probably the most important step in the entire analysis.

5. I do not think we want to start out with more than roughly 20 variables, and each of these will need a reasonable justification. With about 20 variables, we will have about 80 observations per variable for the train data, which may help us reduce the effect of ecological noise (high variation in natural systems) on the results. As you consider what variables to include in the candidate variable sets, include relevant two-way interaction effects. I recommend that we do not consider higher-order interactions (e.g., three-way interactions) because they can be difficult to interpret.

6. Develop a backup list of variables that includes several more variables than is discussed in step 5. That way, if some of the variables chosen initially need to be deleted because of information revealed in subsequent phases of the analyses, we will still have plenty of relevant variables to consider. In the event that some of the initially selected explanatory variable need to be discarded, it will be valuable to know beforehand how to include other variables in the analysis. To facilitate this process, I suggest that you rank the utility of the variables in a very general way (e.g., “highly likely to be important,” and “potentially important”) that will help you decide which variables should be added to the analysis, if needed. This ranking can be made on the basis of knowledge and experience you have about plant ecology in general and WBP ecology in particular.

7. Record the criteria (e.g., ecological relevance, interpretability of the variable, relevance to management or conservation, ease of collecting future data, etc.) that you use to select the candidate explanatory variables. We will need this information later for the Methods section of manuscripts to explain our selection of variables.

**Preliminary Analyses (Kevin)**

1. Examine histograms, scatter plots, and summary statistics to identify outliers, possible errors, and nonlinear relationships.

2. Add quadratic terms to the initial candidate variable set as needed based on scatter plots. If a quadratic term is included in the candidate set of variables, be sure to include its associated main effect too.

3. Use BY statements in SAS Proc Plot to generate coplots to assess whether the data are appropriate for analyzing selected interaction effects. Are sample sizes large enough for various factor-level combinations to give reliable results? Delete from the candidate variable list interactions whose data are inadequate for sound analyses.

4. Run a VIF analysis on the set of explanatory variables that emerges from previous steps (steps 4-6 of Preparation for Analyses; steps 1-3 of Preliminary Analyses) and discard the variable with the highest VIF > 3. If a variable is discarded, add other variable (from step 6 under “Preparation for Analyses”), and re-evaluate the VIFs for this new set of variables. Continue this process, discarding one variable at a time as needed, until you have a set in which the VIFs for all of the variables are < 3.

**Statistical Modeling (Kevin)**

1. For a given response variable, fit a full model, which will include all main and interaction effects identified in the previous preparation and preliminary analysis steps.

2. If the data were obtained via nested sampling protocols, include random effects that will accommodate the possibility that some observations within a hierarchical sampling level may have responses that are correlated. In the case of hierarchical sampling, use SAS Proc Mixed to fit a model with random effects (for the nested variables) and fixed effects. If no clear hierarchical sampling was involved, use generalized linear modeling.

3. Apply the method of backward elimination. Start with a full model and delete non-significant explanatory variables one at a time using Type III sums of squares from SAS Proc GLM and a sequential Bonferroni adjustment of a family-wide alpha = 0.10. At each step, delete the variable that is least significant (i.e., has the highest *P* value). The family of explanatory variables will be the set of variables being considered for a given response variable at each step of the backward elimination process.

4. Continue the backward elimination process until you have a model that contains only variables that are statistically significant based on a sequential Bonferroni adjustment of a family-wide alpha = 0.10.

5. If the response variable is occurrence (presence/absence), use most of the same analyses described above, except apply logistic regression (implemented via SAS Proc Logistic and SAS Proc Genmod). Use −2LogL statistics (instead of *F*, *t*, or Wald statistics) to draw inferences about the model and the explanatory variables. For occurrence data and logistic regression, it will be important to weight the observations in the analysis so that the prevalence (sample size × observation weight) of each group of the binary response variable (presence/absence) is equal (0.5). Implement this step to avoid potential biases in logistic regression coefficients that can have important effects on inferences about explanatory variables.

6. If the response variable involves count data, Poisson regression can be used, but it is almost always affected adversely by overdispersion, which leads to underestimated standard errors of the coefficients and hence overestimated test statistics. To avoid the overdispersion problem of Poisson regression, use negative binomial regression implemented via SAS Proc Genmod. It also is possible that a zero-inflated negative binomial probability distribution will be appropriate for the data; zero-inflated negative binomial regression can be implemented via SAS Proc Genmod or SAS Proc Countreg.

**Checking Statistical Assumptions (Kevin)**

1. Once a model whose explanatory variables are all statistically significant has been identified, assess whether the model adequately meets statistical assumptions. Transformations of variables, and the addition of spatial eigenvectors to account for spatial autocorrelation at different spatial scales, may be needed.

2. Plot the residuals versus each explanatory variable and versus y-hat to assess homogeneity of variances and the appropriateness of model structure. Examine a histogram and normal probability plot of the residuals to check normality. For logistic regression, check the magnitude of the coefficients and their standard errors; aberrantly large standard errors (especially relative to the magnitude of the coefficients themselves) signal failure to meet several assumptions of logistic regression.

3. To assess whether there is spatial autocorrelation in the residuals of a model, use SAM4.0 software to compute Moran’s *I* and plot correlograms. If │Moran’s *I*│ for the residuals > 0.10 for any distance class, use the Spatial Eigenvector Module and partial regression in SAM4.0 to identify the spatial eigenvectors that need to be added to the model for │Moran’s *I*│ of the residuals to be < 0.10 for all distance classes. For logistic and negative binomial regression, check the deviance-to-df ratio (from SAS Proc Genmod); values close to 1.0 indicate no evidence of spatial autocorrelation, outliers, or an inappropriate distributional model.

4. If any changes to the variables or the model are needed to satisfy statistical assumptions, re-evaluate the significance and VIFs of all variables in the model, eliminate variables that are non-significant or whose VIF > 3, and then re-evaluate whether all statistical assumptions have been satisfied. Continue this process until all statistical criteria and assumptions specified above have been met.

5. Use Cook’s distance to identify observations that may have undue influence on the regression coefficients. If the F-distribution percentile of a Cook’s distance value is > 50%, run the model with and without the corresponding observation to determine whether conclusions about the statistical significance of any of the explanatory variables changes. If any observations are deleted, re-fit the model and re-confirm that all of the statistical criteria and assumptions discussed above have been satisfied.

**Interpretation of Results (Kevin)**

1. To conclude that an explanatory variable has potential to exert a biologically important influence on a WBP response variable, that explanatory variable will need to have a coefficient of partial determination (*r* 2) > 10%, or improve a model’s *R*2 by > 10%.