# WHITEBARK PINE: Capability Methods and Regression Analysis Applied to the Natural Sciences.

Eva Marquez
Department of Mathematical Sciences
Montana State University

September 21, 1999

A writing project submitted in partial fulfillment
of the requirements for the degree
Master of Science in Statistics

# APPROVAL

of a writing project submitted by

EVA MARQUEZ.

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

9/22/99

Date

John J Borkowski

John Borkowski
Writing Project Director

# CONTENTS

# I. Abstract

Whitebark pine *(Pinus albicaulis)* is a valuable wildlife resource in the Western United States and Southwestern Canada. Its large seeds are a preferred food for a variety of birds and mammals, especially Clark's nutcrackers *(Nucifraga columbiana)*, red squirrels *(Tamiasciurs hudsonicus)*,and bears *(Ursus spp.)*.Whitebark pine communities provide food and shelter for non-granivorous species as well. (Kendall and Arno, 1989)

The importance of whitebark pine as a wildlife food arises from the large size and high lipid content of its seeds. The seeds are a concentrated, high quality food source that can be stored for 12 months or more in squirrel middens or nutcracker caches; other high elevation foods are more ephemeral. Typically, birds and mammals harvest almost all the viable seeds produced. (Kendall and Arno, 1989)

Whitebark pine seed consumption by grizzly bears in the Yellowstone area is closely correlated with cone crop size. During good cone crop years, Yellowstone bears feed almost exclusively on pine seeds in Autumn. Good cone crops appear to be positively correlated with grizzly bear cub production and early weaning of young; poor whitebark pine cone crops are associated with increased grizzly bear mortalities and conflicts with humans. (Kendall and Arno, 1989)

## II.1 Presentation of the Problem.

The following research is based on data provided by David Spector and Adam Morril for their Master's Thesis *"The Influence of Tree-form, Competition, and Stand Characteristics on Cone Production in Whitebark Pine throughout the Greater Yellowstone Ecosystem"* directed by Kathy Hanson, Associate Professor of Geoghraphy, M.S.U.

For many reasons related to the bears, environment, economy and areas surrounding Yellowstone National Park, the interest in food resources for grizzly bears has been increasing, and with it, the interest in whitebark pine due to the importance of their seeds for the bears' diet. One of the goals in this study was to find a measure which could determine whether or not a tree was a good cone producer. This measure had to be based on the amount of cones produced over several years and had to characterize both *yearly cone production* and *regularity in the production process.*

The other research question presented was related to the regression analysis. Data corresponding to different variables influencing cone production were collected for each tree. The goal here was to determine which of those variables were appropriate for explaining this measure of *"goodness of cone production* " and be able to use this information to predict whether or not a stand would have trees with good cone production. Another important reason was to determine the potential of managing stands in order to improve cone production, benefiting both a potentially threatened tree and endangered grizzly bears in the Greater Yellowstone Ecosystem.

## II.2 Background.

The data were collected throughout the **Greater Yellowstone Ecosystem**. Several measurements corresponding to different aspects of the tree, such as crown area, cone production and competition with other trees, were taken to assess the structure and health of whitebark pine trees.

The **Interagency Grizzly Bear Study Team** has established 19 whitebark pine cone count transects (or stands) within this enviroment. Of these transects, 12 were selected for this study based on factors such as:
location within Greater Yellowstone, number of years of cone counts per site, habitat type and cover types (structure of the forest, dominant vegetation). Sites with a greater number of cone counts were given preference over those having data from fewer years. However, sites with fewer years were also chosen based on their habitat and/or cover types. All habitats and all cover types were represented in the study. This allowed for 120 trees from 12 sites distributed throughout the Greater Yellowstone Ecosystem, with 9 to 18 years of cone counts, in a variety of habitat types and cover types, to be studied.

The field methodology for each measurement will now be described. The diameter at breast hight (dbh) was measured for each stem on the subject tree using a standard dbh measuring tape. Breast height is approximately 4.5 feet from the ground, measured from the uphill side of each stem. Stems were defined as individual trunks within a tree cluster, and branches that diverged from any of these trunks below breast height.

Approximately 60 tree cores were taken from trees that had not already been aged at the time of transect establishment. Tree cores were taken at breast height from the largest diameter stem on the uphill side of the tree. Growth rings were counted by eye to determine age of individual trees.

A Spiegel-Relaskop was used to measure tree height and crown height. Measurements were taken from an observation point 50 feet from the subject tree to the top of the tree, the bottom of the tree and the bottom of the crown. Further crown measurements were made in order to determine an approximate crown area and crown volume. Due to the asymmetric crown shape of whitebark pine and the large variation in crown structure in whitebark pine, calculations of crown area were based on a rectangle and calculations of crown volume were based on a hexahedron. Although there are more accurate methods for measuring crown volume, this method was chosen due to its ease and speed so that it could be easily duplicated in future whitebark pine stand assessments.

The amount of competition for each tree was determined by using a Spiegel-Relaskop to measure basal area in square feet of competing trees surrounding each subject tree. The number of trees (stems) counted was multiplied by the basal area factor (BAF) to determine the basal area in square feet of the competing trees.

# III. Quality Control Methodology.

The first goal in this study is to find an index for determining whether or not a tree is a good cone producer. After several meetings with David Spector, Adam Morill, Katherine Hanson, Courtney Kellum (graduate student of statistics) and John Borkowski (Associate Professor of Statistics M.S.U.), the conclusion was that a good cone production tree is not just one that produces a large quantity of cones but can also the one that produces a certain amount of cones regularly.

If the measure taken as the index for classifying the trees was the total number of cones or the mean number of cones, then the masting trees (those ones with years of large production regardless of the regularity) would be classified as "good" trees even though they do not produce cones regularly. As an example, a tree with this pattern of production along 10 years:

$$0\ 0\ 0\ 0\ 0\ 0\ 200\ 0\ 0\ 0$$

would be equally clasified as the tree with the following pattern:

$$20\ 20\ 20\ 20\ 20\ 20\ 20\ 20\ 20\ 20.$$

For this study the second tree is considered a "better" cone producer than the first one. Our task is to find an index that will penalize the lack of regularity in the first tree.

Another index considered for the classification was:

$$\frac{mean-of-cone-production}{standard-deviation}$$

but since the pattern of cone production is skewed most of the time (as you can observe in *Figure #1*), the mean is not representative of the center of the distribution. Therefore, and as suggested by John Borkowski, we decided to use statistical quality control methods in order to create an index that satisfies the biological aspect of the study as well as the statistical one.

Statistical methods are used in industry to study and improve the performance of a manufacturing process. This performance is related to the variability of the data from the process (the more variability in the data the worse the process' performance) and it is also related to the aim or the target for the process characteristic of interest. In this case, we will apply these methods to natural science data. Specifically, each tree will be considered as a *"production factory of cones"*.

When a production process is studied statistically, the primary interest is *"how well our manufactoring process meets the specifications required by the company"*. In this research, we will study the production data corresponding to each one of the trees in order to rank them with respect to a measurement that indicates the quality of each tree cone production.

In process capability analysis there are indexes (called process capability ratios) that express different aspects of process capability in simple quantitative ways. When it comes to study the capability of a process many other statistical tools are used: charts, histograms, distributions, etc. Process capability ratios are a simple way of determining how good the process is. We will now show how to calculate these process capability ratios.

Most manufacturing processes have *Specification Limits*. These limits are values used to determine whether or not a product meets the specification required for the manufacturer. The largest allowable value for a quality characteristic is the **Upper Specification Limit (USL)** and the smallest allowable value for a quality characteristic is the **Lower Specification Limit (LSL)**. These limits are used to calculate the process capability ratios. For example, the CP ratio is defined as:

$$CP = \frac{USL - LSL}{6\sigma}$$

where $\sigma$ is the standard deviation of the process.

In this case the capability ratio is only a measure of the variability of the data. For example, it does not give any information about the mean of the production process.

There are different process capability ratios. Each ratio provides different information about the process we are studying. In this research the CPL

(Lower process rapability ratio) will be used to order our trees. The reasons we choose this ratio, were:

**(a)** We need a ratio that took into account both production quantity, and variability of production (depending on the year).

**(b)** We wanted a ratio to measure how good the process was with respect to the *lower part of the process.* This means that we are concerned about how well our tree does in producing a minimun amount of cones per year, but we are not going to put any limit in the upper part of the process. That is, the greater production of cones we have, the better. However, we will study whether or not the tree meets the lower specification limit required for the production of cones.

The theoretical form of the capability ratio that we are going use in our analysis is:

$$CPL = \frac{\mu - LSL}{3\sigma}$$

where $\mu$ is the mean of the process and $\sigma$ is the satandard deviation of the process.

Although the formula for the CPL is theoretical, this ratio will be useful if the process follows certain assumptions. The assumption for the process capability ratio is that *the distribution of the data has to be normal or almost normal.*

Figure 1: Histograms corresponding to cone production for trees C1, T3, T4 and U1. Weibull distribution is used to smooth the histograms.

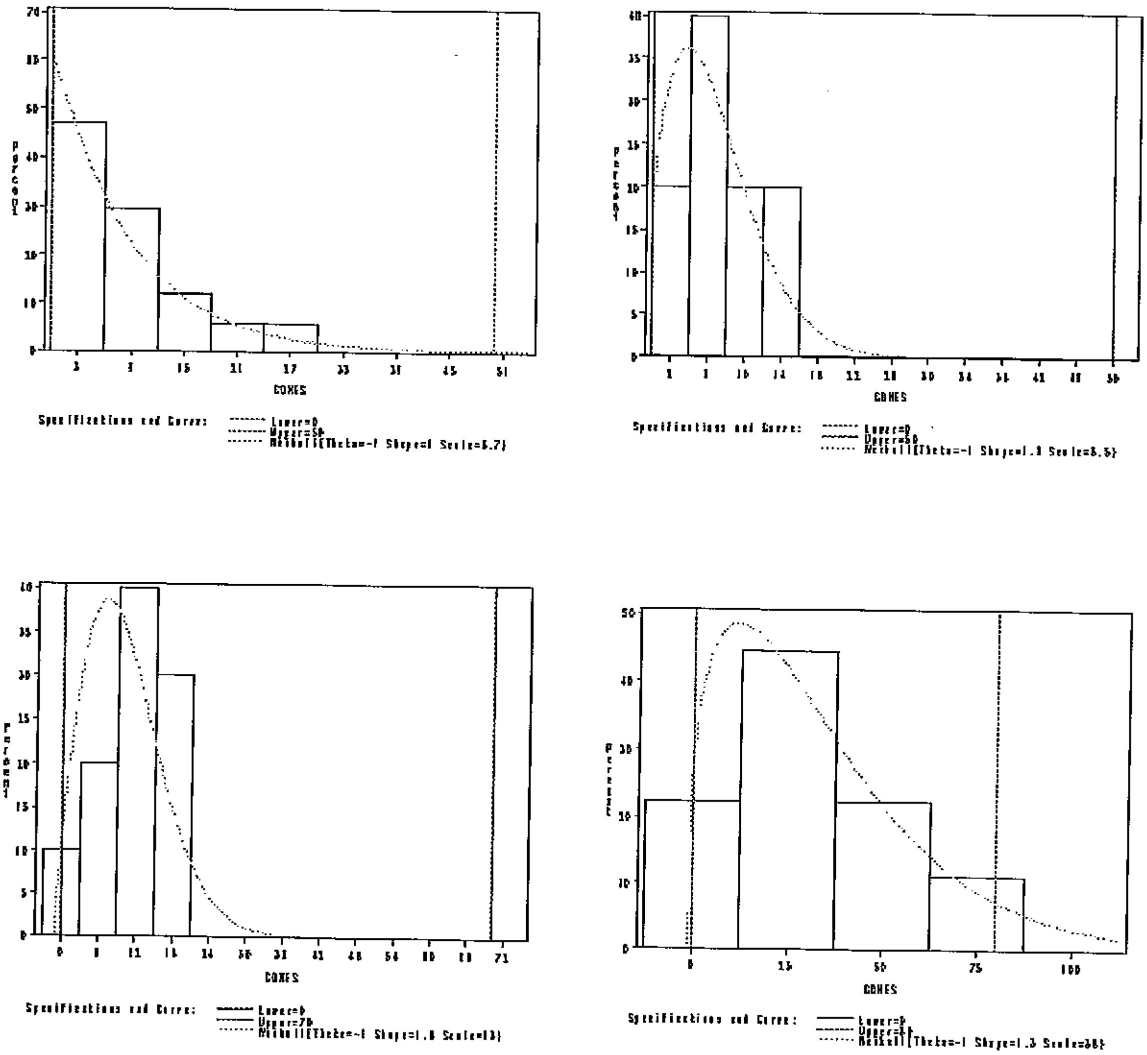


*Figure #1* contains the histograms of the data from four trees. None of these histograms correspond to data coming from a normal distribution. Then, it is natural to ask what do we do if our assumption is not followed. The statistical package *SAS* was used to find the process ratios as well as provide the solution for this problem.

The first thing to do is to choose an appropiate distribution that represents the data. As you can see from the graphs in *Figure #1* the *Gamma*, *Beta* or *Exponential* distribution would be appropiate. However, we wanted a single, but very flexible family of distributions that could represent every single tree, with each tree having different estimated parameter values. The family choosen was the two-parameter Weibull distribution which has the following density function:

$$p(x) = \begin{cases} \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} exp(\text{-}(\frac{x}{\theta})^{\beta}) & x > \theta \\ 0 & x \leq \theta \end{cases}$$

where
$\theta =$ scale parameter ($\theta > 0$).
$\beta =$ shape parameter ($\beta > 0$).

As you can see in *Figure #1*, the Weibull density function is smoothing the histograms and fits reasonably well the data for every single tree. When the software program *SAS* fits the distribution, the scale and shape parameters are estimated for each case (tree) based on the data.

The distribution goodness of fit was checked using the following tests statistics based on the empirical distribution function (EDF): **Anderson-Darling** and **Cramer-von Mises** statistics. For every single case the lack of fit with the Weibull is not significant, implying that the Weibull distribution fits the data reasonably well. This is due to the flexibility of the Weibull distribution. That is, it can represent data arising from distributions with very different shapes.

Once we have the fitted distribution, *SAS* calculates "modified" capability ratios based on the fitted distribution. The ratios are "modified" because the ratios are no longer based on statistics, such as the **mean** or the **standard deviation**, but on ordered statistics, such as the **median** or the **percentiles**.

These modified ratios are:

$$CPL = \frac{P_{0.5} - LSL}{P_{0.5} - P_{0.00135}}$$

where $P_{0.5}$ is the median for the fitted distribution, LSL is the Lower Specification Limit, and $P_{0.00135}$ is the .135 percentile of the fitted distribution. The percentiles are already included in the Quality Control Package of *SAS*.

The modified capability ratios for the fitted distribution are used, and not just simply the percentiles computed from the raw data, because many of the trees have data with a **median of 0**. Therefore, we group the data into histograms after we fit a distribution and calculate the statistics based on that distribution.

In conclusion for this first part of the study , we found an index that considers both quantity of cone production and regularity. This index also satisfies the classification criteria for the trees approved by the biologists-geographers implicated in the study.

# IV Regression Analysis.

As explained in section *II.2*, data corresponding to different variables having an influence on cone production were collected from 1980 to 1997. The second goal in this study is to use variables from this data set to construct a multiple regression model of the response: the **CPL** index, which is the measure of productivity used for ranking the trees. This information will be used the future for predicting cone productivity and managing the stands in order to have better *cone producers.*

There are two other measures of productivity; **average number of cones** and **total number of cones**. Although a regression analysis has been conducted with these response variables, these measures do not take into account the variability in the cone production process. Therefore, we will just mention the results for these analysis at the end of this section. The statistical software packages used for this analysis were *SAS* and *S-Plus.*

## IV.1 Scaling and transforming the data.

The situation where predictor variables are closely linearly related to each other is called **multicollinearity**. In order to improve this situation, the variables were *standarized,* meaning that they were *centered and scaled (the vector of their values has norm one).* Therefore, if the columns of the design matrix, representing vectors of the explanatory variables, show any type of linear relationship, this is not because of the difference in the location of the variables, which could cause a misleading multicollinearity alarm, but because of *'real'* multicollinearity problems.

The goal is to check which variables are correlated and why, so we can decide which ones will be in the model. One way we assess the dimension of the relationship of an independent variable with the rest of independent variables is using *the condition number of the design matrix* $\eta_j$:

$$\eta_j = \sqrt{\lambda_{max}/\lambda_j}$$

where $\lambda_j$ is the eigenvalue of the design matrix corresponding to the standarized explanatory variable $x_j$. Then we can judge the size of the eigenvalue related to the variable $x_j$ in relation with the rest of eigenvalues. In *Table 1* we have an example of what the condition number of a matrix would be for the following model:

**CPL** = 0.6295 (**I**)
+ 0.0001*Crown Area (**CA**)
+ 0.00005*Cross Sectional (**XS**)
+ 0.0036*Total tree height (**TH**)
- 0.0003*Total Basal Area (**BS**)
- 0.001*Ensp. Basal Area (**EB**)

*Table 1*

| EV | CN | I | CA | XS | TH | BS | EB |
|---|---|---|---|---|---|---|---|
| 4.390 | 1.000 | 0.002 | 0.011 | 0.016 | 0.002 | 0.005 | 0.010 |
| 0.795 | 2.350 | 0.001 | 0.010 | 0.000 | 0.000 | 0.001 | 0.909 |
| 0.466 | 3.069 | 0.006 | 0.010 | 0.793 | 0.004 | 0.023 | 0.003 |
| 0.260 | 4.107 | 0.003 | 0.772 | 0.187 | 0.001 | 0.057 | 0.018 |
| 0.068 | 8.050 | 0.188 | 0.163 | 0.001 | 0.059 | 0.872 | 0.001 |
| 0.021 | 14.340 | 0.800 | 0.034 | 0.002 | 0.934 | 0.042 | 0.059 |

**EV** represent the Eigenvalues corresponding to the design matrix where the variables have been standarized. The smallest eigenvalue would be 0.021. **CN** represent the Condition Numbers $\eta_j$ corresponding to each variable, the highest condition number is 14.34, which indicates no collinearity problems among these explanatory variables. The rule of thumb commonly used for considering collinearity problems is the *condition number* to be larger than 30. The rest are the values corresponding to the proportion of the variances for each variable.

The results for the final model were:

**CPL** = 0.695888(**I**)
+ 0.0043560*Sqrt. Crown Area(**SCA**)
- 0.000828*Total Basal Area (**BS**)
- 0.000019555*Int1 (**Int1**)

+ 0.000289*Int2 (**Int2**)
- 0.000728*Int3 (**Int3**)
+ 000033076*Int4 (**Int4**)

| **EV** | **CN** | **I** | **SCA** | **BS** | **Int1** | **Int2** | **Int3** | **Int4** |
|---|---|---|---|---|---|---|---|---|
| 2.50870 | 1.00000 | 0.0000 | 0.0213 | 0.0069 | 0.0247 | 0.0099 | 0.0252 | 0.0146 |
| 1.79470 | 1.18230 | 0.0000 | 0.0180 | 0.0481 | 0.0095 | 0.0326 | 0.0029 | 0.0109 |
| 1.00000 | 1.58389 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.89994 | 1.66962 | 0.0000 | 0.0061 | 0.0142 | 0.7643 | 0.0008 | 0.0205 | 0.0039 |
| 0.56151 | 2.11371 | 0.0000 | 0.2383 | 0.0097 | 0.0959 | 0.0065 | 0.1794 | 0.0001 |
| 0.17409 | 3.79608 | 0.0000 | 0.0682 | 0.4411 | 0.0997 | 0.3225 | 0.1089 | 0.1703 |
| 0.06106 | 6.40971 | 0.0000 | 0.6482 | 0.4800 | 0.0058 | 0.6278 | 0.6630 | 0.8002 |

where

- **Int1** is the interaction between *Sqrt. ensp. basal area* (which reflects the amount of competition between the whitebark pine and the Engelmann Spruce, another tree grown in areas close to whitebark pines) and *Crown Area.*

- **Int2** is the interaction between *Square root of the total basal area* and *Total tree height.*

- **Int3** is the interaction between the *number of stems* and *Total tree height.*

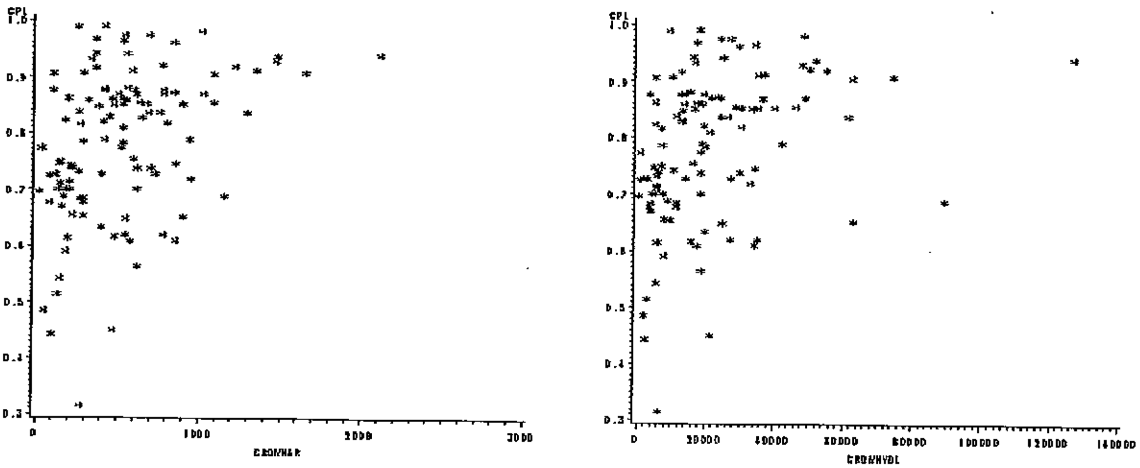
- **Int4** is the interaction between the *number of stems* and *Crown Area.*

The condition number for the design matrix of this model is 6.41. We do not have collinearity problems and the last row shows that the highest proportion of variance corresponds to the variable Int4. Interactions between variables are explained the following way: the result of combining the two variables has a joint effect on the response variable. In this case, the ones selected are significant in the explanation of the variance in the model response.

The procedure was to *standarize* the variables in order to choose the appropiate ones, study the correlation among them and whether or not they have collinearity problems, analyze the model, and find the estimates of the parameters.

In biological data, a transformation of a variable frequently explains even more of the variability in the response than the original variable itself. By plotting each explanatory variable against the response we observe the type of relation between the two variables. Therefore, we can use these plots to decide if we need to transform any of the variables and how they should be transformed. In *Figure 2* we observe how the data suggest a square root transformation for the variables CROWN AREA and CROWN VOLUME.

*Figure 2.* Crown Area and Crown Volume against the response variable CPL.



The variable *ensp. basal area* shows the same pattern suggesting a square root transformation as well. The rest of the variables, however, do not suggest any possible transformation to improve the multiple regression analysis.

## IV.2 Choosing the variables.

David Spector and his master thesis committee decided which variables made more bioligical sense for explaining the response variable. These variables were: Crown Area, Total Tree Height, Total DBH and Total Basal Area, but as the proportion of the variance explained with these variables was not satisfactory we decided to use statistical methods to improve the model.

The procedures used for the selection of the variables are called the *Foward Selection Procedure, Backwards Selection Procedure and Stepwise Selection Procedure.* The *Foward Selection Procedure* starts with no variables in the model and the variables are sequentially added depending on how much remaining variability in the response is explained by an individual variable. The greater the remaining variability explained the sooner they enter in the model. The *Backwards procedure* starts with all the explanatory variables and sequentially eliminates from the model those variables not needed for the explanation of the variance. The *Stepwise Procedure* is a combination of the two of them. It starts with no variables in the model, adds the variables which most influence the response and considers if there is any variable which should be deleted from the model at each step. The variables were chosen depending on the value of $R^2$ or $adjusted - R^2$ (if the value of $adjusted - R^2$ increases the variable to be added is *'needed'* in the model). That is, it is worthy to be in the model even though adding a variable to the model is going to increase the number of parameters to estimate and increases the possibility of multicollinearity among the variables. The influence of the variable with respect to the response is determined by the p-value for the F-tests.

Among the models resulting from the different methods, the final model was chosen based on the variables included (under David Spector's criteria) and on the value of $R^2$ and $adjusted - R^2$. The results are:

$$\begin{aligned}
\textbf{CPL} \quad &= 0.695888(\textbf{I})\\
&+ 0.0043560 * \text{Sqrt. Crown Area}(\textbf{SCA})\\
&- 0.000828 * \text{Total Basal Area } (\textbf{BS})\\
&- 0.000019555 * \text{Int1 } (\textbf{Int1})\\
&+ 0.000289 * \text{Int2 } (\textbf{Int2})\\
&- 0.000728 * \text{Int3 } (\textbf{Int3})\\
&+ 000033076 * \text{Int4 } (\textbf{Int4})
\end{aligned}$$

The summary of this model is the following:

| Coef. Variables | Estimate | Std.Error | t-value | p-value |
|---|---|---|---|---|
| INTERCEP | 0.695888 | 0.05305885 | 13.115 | 0.0001 |
| SCA | 0.004356 | 0.00247137 | 1.763 | 0.0809 |
| BS | -0.000828 | 0.00020879 | -3.963 | 0.0001 |
| Int1 | -0.000019555 | 0.00000558 | -3.505 | 0.0007 |
| Int2 | 0.000289 | 0.00007421 | 3.896 | 0.0002 |
| Int3 | -0.000728 | 0.00032294 | -2.254 | 0.0263 |
| Int4 | 0.000033076 | 0.00002422 | 1.365 | 0.1750 |

Residual standard error: 0.11034 on 105 degrees of freedom
Multiple R-Squared: 0.335
F-statistic: 8.814 on 6 and 105 degrees of freedom, the p-value is <0.0001

*Summary*
The model explains 33.5% of the variability in the CPL response and has p-value less than 0.0001, which measures the *'goodness of the model to fit the data'.* In this case we can accept the model. The variables have a significant influence on the response (small p-values), and the standard errors are reasonably small.

CORRELATION MATRIX

| | **SCA** | **BS** | **Int1** | **Int2** | **Int3** |
|---|---|---|---|---|---|
| **BS** | -0.05682 | | | | |
| **Int1** | 0.22348 | 0.14519 | | | |
| **Int2** | 0.11745 | 0.84003 | 0.31211 | | |
| **Int3** | 0.38256 | 0.17193 | 0.17007 | 0.30223 | |
| **Int4** | 0.74098 | 0.00412 | 0.13733 | 0.07355 | 0.75911 |

The correlation matrix shows how each variable is linearly related to the rest. In this study the correlation matrix has been mainly used to choose the variables (at the same time using the procedures and the recommendations made by David Spector and his committee) to build the model, as well as to avoid collinearity problems.

## IV.3 Diagnostics

Once we have chosen the model for the data:

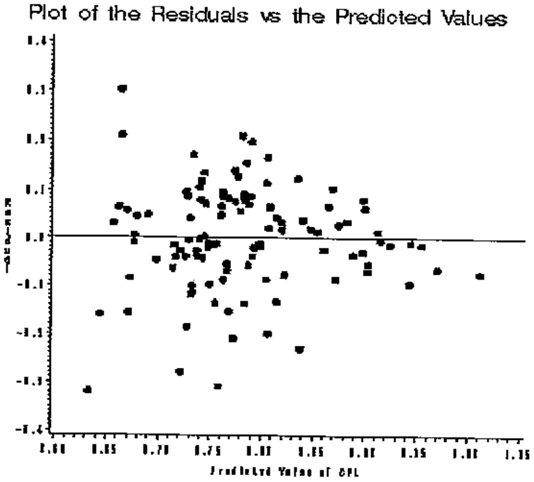
$$
\begin{aligned}
\mathbf{CPL} &= 0.695888(\mathbf{I}) \\
&+ 0.0043560\text{*Sqrt. Crown Area}(\mathbf{SCA}) \\
&- 0.000828\text{*Total Basal Area } (\mathbf{BS}) \\
&- 0.000019555\text{*Int1 } (\mathbf{Int1}) \\
&+ 0.000289\text{*Int2 } (\mathbf{Int2}) \\
&- 0.000728\text{*Int3 } (\mathbf{Int3}) \\
&+ 000033076\text{*Int4 } (\mathbf{Int4})
\end{aligned}
$$

we have to check the diagnostic plots for the residuals in order to determine whether or not any of the Gauss Markov conditions needed to use these methods:

- $\mathrm{E}(e_i) = 0$, (expected value for the residuals $= 0$),
- $\mathrm{E}(e^2{}_i) = \sigma^2$, (homocedasticity),
- $\mathrm{E}(e_i e_j) = 0$ when $i \neq j$, (uncorrelated errors),

have been violated. The first plot to study is the *residuals* of the selected model vs. the *predicted* values given in *Figure 3:*

*Figure 3.* Residuals vs. Predicted values.

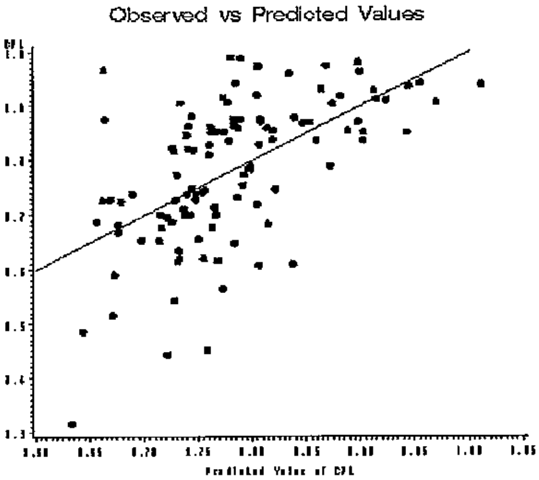


The plot shows residuals randomly scattered about the horizontal line Resid = 0. The interpretation of the plot is that no special pattern is observed in the plot, implying that residuals are uncorrelated, the equality in the vàriance of the different explanatory variables and the mean of the distribution for the residuals being zero.

A particular pattern could indicate several situations such as not including a variable that is needed in the model, some of the variables included should be transformed or weighted or the conditions of homogeneity in the variance or uncorrelation have been violated . The situation for this model is ideal in the sense that we do not observe any strong particular pattern.

The next plot displays *observed values* vs. *predicted values.* This plot is useful to check how well our model predicts the actual values.

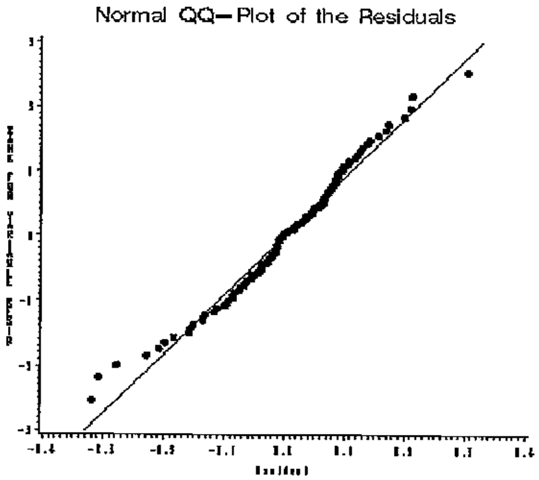
*Figure 4.* Observed values vs. Predicted values.



In *Figure 4* we observe how well the predicted values (straight line) fit the actual data. The desirable plot would show the actual data *(dots)* close to the line, which would indicate a $R^2$*-value* close to one. For biological data is frequent for the model to explain between 30 and 50% of the data, in our case the model explains 33.5% of the data.

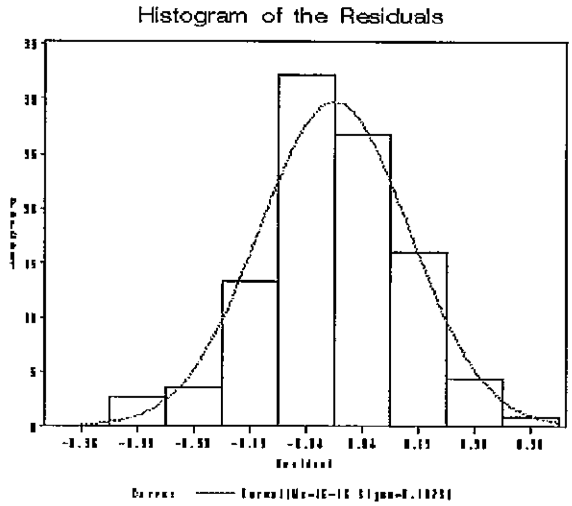
Another assumption we need to check is whether or not the residuals follow a normal distribution by using a *QQplot* of the residuals, as shown in *Figure 5*: .

*Figure 5.* Normal QQplot of the residuals.



The residuals are supposed to follow a Normal distribution $N(0,1)$, which is represented by the straight line. The residuals for this model do not differ significantly from the line, although there is a small discrepancy in the tails (extremes of the plot due to the positive *kurtosis* (the probability is concentrated in a single peak)) and the slight left skewness of the residuals' distribution, as you can observe in *Figure 6*

*Figure 6.* Histogram for the residuals. Normal line smoothing the histogram



Once the assumptions of **normality, homocedasticity** and **uncorrelated errors** are accepted for the distribution of the residuals in this model, we can accept the model. That is, the conditions required for the method used are appropriate.

## IV.4 Other models

Although the model chosen was the best model from the *statistical point of view*, some other models were studied and analyzed. David Spector and his committee decide that the following model was the best one from the *biological point of view*:

**CPL** = 0.599148 (**I**)
+ 0.005886*Sqrt. Crown Area (**SCA**)
- 0014438*Stems (**STE**)
- 0.000288*Total Basal Area (**BS**)

+ 0.003012*Total Tree Height (**TH**)
- 0.014438*Stems (**STE**)
- 0.005314*Int.1 (**I1**)

where **I1** is the interaction between *number of stems* (**STE**) and *sqrt. ensp. basal area.* The summary of parameters' estimates, the corresponding *p-values* and the correlation matrix are:

| Coef. Variables | Estimate | Standard Error | t-value | p-value |
|---|---|---|---|---|
| Intercept | 0.599148 | 0.05642125 | 10.619 | 0.0001 |
| SCA | 0.005886 | 0.00149591 | 3.934 | 0.0001 |
| TH | 0.003012 | 0.00101929 | 2.955 | 0.0038 |
| STE | -0.014438 | 0.01051829 | -1.373 | 0.1728 |
| BS | -0.000288 | 0.00011394 | -2.526 | 0.0130 |
| I1 | -0.005314 | 0.00171506 | -3.099 | 0.0025 |

Residual standard error: 0.11209 on 106 degrees of freedom.
Multiple R-Squared: 0.3071
F-statistic: 9.398 on 5 and 106 degrees of freedom, the p-value $< 0.0001$

CORRELATION MATRIX

| INT1 | SCA | BS | STE | TH |
|---|---|---|---|---|
| **BS** | -0.05682 | | | |
| **STE** | 0.30477 | 0.00317 | | |
| **TH** | 0.28678 | 0.41155 | -0.03597 | |
| **I1** | 0.07920 | 0.15168 | 0.08494 | 0.28422 |

The selection of this model was based on the *biological sense* of the explanatory variables and the possibility of finding a *biological explanation* to the interaction used. Diagnostic plots were conducted for this model and we did not observe any type of violation of the *Gauss-Markov* conditions. Good results are obtained when we analyze the collinearity among the variables.

Before choosing the CPL capability ratio as an index for ranking the trees, the measures taken as indexes were **average number of cones** and **total**

**number of cones**, for both measures a multiple regression analysis was conducted. The methods used to find the *'best'* model were the same as the ones used for the CPL ratio. This is the model for the **average number of cones**:

Average number of cones (**AV**) = - 6.572350 (**I**)
+ 1.143099*Sqrt. Crown Area (**SCA**)
- 0.011645*Cross Sectional (**XS**)
+ 0.016171*Total Basal Area (**BS**)
+ 0.000013011*Int1 (**I1**)
- 0.000061408*Int.2 (**I2**)

where **I1** is the interaction between **CROWN AREA** (without taking the square root) and **XS**, **I2** is the interaction between **CROWN AREA** and **BS**. The corresponding analysis is:

| Coef. Variables | Estimate | Standard Error | t-value | p-value |
|---|---|---|---|---|
| INTERCEPT | -6.572350 | 5.20050560 | -1.264 | 0.2091 |
| SCA | 1.143099 | 0.22441993 | 5.094 | 0.0001 |
| I1 | 0.000013011 | 0.00000274 | 4.740 | 0.0001 |
| BS | 0.016171 | 0.01175320 | 1.376 | 0.1718 |
| I2 | -0.000061408 | 0.00001819 | -3.377 | 0.0010 |
| XS | -0.011645 | 0.00229407 | -5.076 | 0.0001 |

Residual standard error: 7.95695 on 106 degrees of freedom
Multiple R-Squared: 0.4904
F-statistic: 20.402 on 5 and 106 degrees of freedom, the p-value is $< 0.0001$

All the variables appear significant in this model. The proportion of variance explained with the model is 49%, the errors seem stable, and there are no collinearity problems, as we can observe from the following matrices: the **condition number matrix** to check collinearity and the **correlation matrix** to see how correlated the variables are.

CONDITION NUMBER MATRIX

| EV | CN | I | SCA | Int1 | BS | Int2 | XS |
|---|---|---|---|---|---|---|---|
| 3.08138 | 1.00000 | 0.0000 | 0.0138 | 0.0125 | 0.0013 | 0.0107 | 0.0155 |
| 1.16938 | 1.62328 | 0.0000 | 0.0028 | 0.0039 | 0.2516 | 0.0141 | 0.0073 |
| 1.00000 | 1.75539 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.58473 | 2.29560 | 0.0000 | 0.0982 | 0.0175 | 0.0417 | 0.0329 | 0.1415 |
| 0.09367 | 5.73546 | 0.0000 | 0.3638 | 0.6909 | 0.0712 | 0.0774 | 0.5960 |
| 0.07084 | 6.59534 | 0.0000 | 0.5213 | 0.2751 | 0.6343 | 0.8648 | 0.2396 |

The condition number for this model is 6.59, we do not have collinearity problems, although from the last line (as it is expected) we have a strong correlation among Sqrt. Crown Area, Total Basal Area and the interaction between these two variables.

CORRELATION MATRIX

| | SCA | XS | BS | I1 |
|---|---|---|---|---|
| **XS** | 0.55688 | | | |
| **BS** | -0.05682 | 0.03324 | | |
| **I1** | 0.72730 | 0.87620 | 0.01767 | |
| **I2** | 0.78449 | 0.50080 | 0.43879 | 0.66243 |

The analysis for the response **total number of cones** is very similar to the previous one. Therefore we would just show the final model and the results for the correlation matrix:

Total number of cones (**TC**) = - 69.709795 (**I**)
+ 10.602894*Sqrt. Crown Area (**SCA**)
- 0.103586*Cross Sectional (**XS**)
+ 0.163485*Total Basal Area (**BS**)
+ 0.000116*Int.1 (**I1**)
- 0.000569*Int.2 (**I2**)

Below is the correponding analysis:

| Coef. Variables | Estimate | Standard Error | t-value | p-value |
|---:|---:|---:|---:|---:|
| INTERCEPT | -69.709795 | 46.71485704 | -1.492 | 0.1386 |
| SCA | 10.602894 | 2.01590876 | 5.260 | 0.0001 |
| XSE | -0.103586 | 0.02060708 | -5.027 | 0.0001 |
| BS | 0.163485 | 0.10557607 | 1.549 | 0.1245 |
| I1 | 0.000116 | 0.00002466 | 4.702 | 0.0001 |
| I2 | -0.103586 | 0.02060708 | -5.027 | 0.0001 |

And the correlation matrix would be:

| | **SCA** | **XS** | **BS** | **I1** |
|---|---|---|---|---|
| **XS** | 0.55688 | | | |
| **BS** | -0.05682 | 0.03324 | | |
| **I1** | 0.72730 | 0.87620 | 0.01767 | |
| **I2** | 0.78449 | 0.50080 | 0.43879 | 0.66243 |

# V. Conclusions and Final Comments.

After several meetings with David Spector, Adam Morill, Kathy Hanson, Courtney Kellum and John Borkowski, the goal was to find an index to rank the whitebark pine trees depending on the quantity of cone produced and the regularity in the cone production process. In section III we explained how the CPL was chosen as cone production index among the diferent options, but it has not been explained why the lower specification limit in the capability analysis is *'zero'*.

The use of pine seeds by the bears was studied over several years, (Mattson Reinhart, 1990). It was concluded that the heavy use of the tree occurs when the crop's average was at least 13-23 cones per tree. Somehow, we wanted to consider this *boundary* into the index CPL making the lower limit to be 20. This resulted in indices with large negative values since many of the trees have a median of *zero* cones produced. Therefore, because the ranking of the trees did not change, we decided to choose a lower limit of *zero*.

From the regression point of view, the most important conclusion from this study for me has been the difficulty of finding a model that reasonable explains the collected data due to the large variability found in natural/biological data.

The fact that variables involved in the study were highly correlated made the selection of the model more complicated, (*i.e.* the *Crown Volume* data were obtained based on the *Crown Area* values, there were different meaures of Basal Areas, the *Number of Stems* was highly correlated with the *DBH*). Although we knew *apriori* which variables should not be together in the model because they were highly correlated and they would explain the same proportion of the variability, the selection was complicated in the sense that there were many different combinations among those variables equally valid for explaining the variability in the data. One of the decisions was to choose between the *Crown Volume* and the *Crown Area*. For some models, the Volume was a better regression variable than the Area, and, viceversa. Both variables also explained the same proportion of variance. Finally, we decided to use the model with the Crown Area because David Spector thought it was more *biologically* accurate.

Another concern about the chosen models was the lack of significance of the Interaction term in the models. Thus, it makes sense biologically to remove the Interaction term from the model: If the tree has height or area *'zero'*, it is not going to produce cones. Therefore, the CPL (response variable) would be *'zero'*. The problem would be that we could not use $R^2$ as measure for the proportion of the variance explained by the model, and since $R^2$ is the standard accepted and use for biological studies, we decided not to remove the Interaction term from the models.

The fact that the model chosen under the statistical method based on the inclusion of statistical significant variables is not the same as the one chosen by the biologists/geographers involved in the study is mainly due to the difference in explanatory variables used. Even though the model selected by David explains less variability of the data, the interactions in this model can be understood and explained from the biological point of view more easily.

# References

[1] Goossens, M., Mittlebach, F. & Samarin, A. (1994). *The Latex Companion.* Massachusetts: Addison-Wesley Publishing Company.

[2] Montgomery, D. C. . *Introduction to Statistical Quality Control.* New York: John Wiley & Sons.

[3] Sen, A. K. & Srivastava, M. S. (1990). *Regression Analysis: Theory, Methods, and Applications.* New York: Springer-Verlag.

[4] Snow, W. (1992). *Tex for the beginner.* Massachusetts: Addison-Wesley Publishing Company.

[5] Venables, W. N. & Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus.* New York: Springer.

[6] *S-Plus Guide to Statistical and Mathematical Analysis.* Seattle: StatSci Division, MathSoft,Inc.

[7] *SAS/STAT User's Guide.* Cary, N. C. : SAS Institute, 1994.

[8] *SAS/QC Software.* Cary, N. C. : SAS Institute, 1994.