TWO-STAGE AND ZERO-INFLATED MODELLING OF FOREST REGENERATION ON THE PACIFIC NORTHWEST COAST

A Thesis
Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a
Major in Forest Resources
in the
College of Graduate Studies
University of Idaho

by
Robert F. Keefe

April 2004

Major Professor: Andrew P. Robinson, Ph.D.

AUTHORIZATION TO SUBMIT THESIS

This thesis of Robert F. Keefe, submitted for the degree of Master of Science with a major in Forest Resources and titled “Two-Stage and Zero-Inflated Modelling of Forest Regeneration on the Pacific Northwest Coast,” has been reviewed in the final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor

______________________ Date ________________

Andrew P. Robinson

Committee Members

______________________ Date ________________

William R. Wykoff

______________________ Date ________________

Paul Joyce

Department
Administrator

______________________ Date ________________

Jo Ellen Force

Dean of College
of Natural Resources ______________________ Date ________________

Steven Daley Laursen

Final Approval and Acceptance by the College of Graduate Studies

______________________ Date ________________

Katherine G. Aiken

ACKNOWLEDGMENTS

I would like to thank my parents and family for their support throughout my work on this thesis. I would like to thank my advisor, Andrew P. Robinson, for giving me the opportunity to learn about forest biometry, and for letting me explore the topics discussed in this document with a large degree of independence. My committee members, Dr. Paul Joyce and William R. Wykoff have provided a great deal of guidance with the research and have suggested many helpful improvements to the document.

This research was made possible in part by financial and other support from Dr. Tom Spies and Robert J. Pabst as part of the Coastal Landscape Analysis and Modelling Study, a collaboration of the Pacific Northwest Research Station of the US Forest Service and Oregon State University. Additional funding and resources were provided by the Forest Biometrics Laboratory at the University of Idaho.

ABSTRACT

In this thesis, several count models that deal with the problem of excess zeroes (overdispersion) in Long Term Ecological Research (LTER) plots on the Washington and Oregon coast are described and evaluated. The models are to be used within the Pacific Northwest Coast Variant of the Forest Vegetation Simulator (Stage, 1973), and generally follow the two-stage structure of the Regeneration Establishment Model (Ferguson and Carlson, 1993; Ferguson and Crookston, 1991). The probability of any regeneration occurring on a given plot is initially estimated using logistic regression. Then, in the second stage of analysis, the conditional distribution of stocked plots (those with at least one seedling) is assumed to be distributed according to a Weibull density with parameters predicted from plot conditions.

Missing values in abiotic predictors integral to the Regeneration Establishment Model and lack of management history made direct calibration of the existing model implausible, and several new two-stage models were constructed instead. The results were inconclusive. In most cases, there was not a clear linear relationship between the estimated Weibull parameters and plot (or stand) characteristics. Inconsistencies in the fitting data increased the complexity and difficulty of the modelling efforts.

As an alternative approach, a finite mixture regression model was considered. In the finite mixture model, the parameters of the two-stage system are estimated simultaneously in a generalized non-linear model. A simple Zero-Inflated Negative Binomial (Lambert, 1992) model was fit to the same data. The results were similar to those of the two-stage approach. Lastly, several recommendations for future regeneration sampling and modelling efforts are made.

TABLE OF CONTENTS

LIST OF FIGURES

Figure No. Page

LIST OF TABLES

Table No. Page

PART I: GENERAL CONSIDERATIONS

1.0 Forest Regeneration Models in the Literature

Regeneration establishment depends on potentially very complex interactions among many contributing factors (Rogers and Johnson, 1998), which may not be explained by the usual suite of variables measured in conventional, yield-oriented forest inventories (Price et al., 2001). In addition, the process often presents difficulties to modellers due to inconsistent sampling protocols across studies (Shifley et al., 1993).

Several regeneration establishment models have been developed for hardwood forest types in the Eastern and Central United States. Dey (1991) modelled oak regeneration in the Missouri (U.S.A.) Ozarks using individual tree survival models for prediction. Botkin et al. (1970) wrote JABOWA and JABOWA II to simulate oak regeneration using stochastic functions. SIMSEED (Rogers and Johnson, 1993) is also based on stochastic functions, but is unique in that it simulates fluctuations in the *magnitude* of regeneration frequencies over cycles of various lengths (Rogers and Johnson, 1993). Ribbens et al. (1994) calibrated existing seedling recruitment functions in oak and mixed northern hardwood forests by relating species-specific seedling distributions with overstory spatial structure using a likelihood-based approach. A simple overview of these establishment models in the literature was provided by Rogers and Johnson (1998).

Price et al. (2001) provided a much more detailed survey of how regeneration is treated in the gap model literature, and critiqued the suitability of these models (including JABOWA II, SIMSEED, FORET, FORSKA, 4C, and others) for modelling the effects of climate change on recruitment. They observed that “the formulation of regeneration processes [in gap models] has changed very little in the last 15 yr.” (Price et al., 2001, p. 478)

They identified three primary shortcomings of these for simulating processes such as, for example, species migration and species composition changes due to warmer temperatures associated with changes in climate. These were as follows:

1. The treatment of regeneration as a single lumped process, rather than as the combined effect of the component physical and biological processes which comprise it.

2. A common assumption that site conditions are homogenous.

3. A general lack of consideration of the effects of herbivores on regeneration processes.

Of these three items, the authors give most attention to the first. They stress that simple formulations of, for example, regeneration density predictions, ignore the suite of complex ecophysiological processes governing seed production, seed dispersal, seed germination, vegetative reproduction (sprouting), seedling establishment, and seedling growth (Price et al., 2001). Each of these processes may be affected in unique ways by climate change. They also highlighted problems associated with modelling landscape–level effects associated with regeneration (species migration) with distance–independent approaches that tend to be limited to *within–gap* processes. An obvious problem, for example, is that seed dispersal *between* gaps is not well accounted for, but is clearly important for simulating potential changes in species range due to warmer temperatures (Price et al., 2001).

Adding to the difficulty involved in generating landscape-scale recruitment models (due to the biological phenomenon being quite complex) is the problem that definitions of regeneration are highly variable due to variation in sampling designs (Shifley et al., 1993). In particular, the minimum diameter threshold at which a tree may be *counted*, or measured for various characteristics (and assumed to be regeneration) is inconsistent. Databases used in the construction of regeneration models often come from re-measured continuous forest inventory plots, and as a result the amount of time passing between plot visits may also vary. For examples, see Shifley et al. (1993) and Schweiger and Sterba (1997).

Shifley et al. (1993) proposed a method for predicting regeneration from the variable minimum diameter thresholds themselves, which also accounted for variable plot remeasurement times. Their method assumed a maximum possible limit on ingrowth, calculated as difference between the maximum possible crown competition factor (CCF) and the current calculated CCF for a stand. The maximum CCF was a function of minimum diameter threshold.

Ferguson and Carlson (1993) and Ferguson et al. (1986) predicted ingrowth within FVS as a two-stage hurdle model based on the approach of Hamilton (1974). Ferguson and Johnson (1998), Schweiger and Sterba (1997) and many others have modified the two-stage approach in order to develop regional models similar in structure to the North Idaho variant of Version 2 of the Regeneration Establishment Model for new data sets.

Moeur and Stage (1995) introduced *most similar neighbor* (MSN) inference as a general multivariate alternative to traditional regression-based modelling strategies in forestry. Ek et al. (1997) further developed this concept and implemented an imputation-based tabular regeneration model for post-harvest stands in Minnesota[1]. Recently, software has been developed by the U.S. Forest Service to implement MSN imputation for various stand attributes (Crookston, 2002), and this approach has been applied in regeneration prediction within Prognosis^{BC} (Froese et al., 2003).

[1] The tables were constructed from the Forest Inventory and Analysis (FIA) plot database.

2.0 Fitting Data

The data were compiled from four sources. These sources, and their corresponding reference codes, are the Stand Management Cooperative Northwest (SMCNW), the H.J. Andrews Experimental Forest Long Term Ecological Research program (TV010), a separate study at H.J. Andrews Experimental Forest investigating biomass dynamics following logging (TP73), and a small study analyzing the growth and development of mixed coniferous and red alder (*Alnus rubra*) stands (TV038). The number of plots from each source are represented in Table 2.1. Plots were all fixed-area (no variable-radius), but the plot sizes were inconsistent.

Table 2.1 Number of unique plots from each data source

Source	Plots
SMCNW	161
TP73	193
TV010	884
TV038	5

The following sections contain descriptions of each of the four data sources.

2.0.1 TV010

This study is the largest source of data (greater than 63 percent). These data are from the HJ Andrews Long Term Ecological Research plot network (LTER). The LTER permanent plots are situated in several national forests (Willamette, Siuslaw, Deschutes, Mt. Hood, Gifford Pinchot, and Olympic) and national parks (Mt Rainier, and Olympic) in western Washington and Oregon. Most plots were established in one of two general time periods. The first group was established by the Forest Service between 1910 and 1948, with the objective of quantifying growth and yield of commercially important species. A motivating goal was to accurately estimate timber volume lost to mortality over time.

The second wave of plot establishment occurred during the period from 1970 to 1989, and was part of the Coniferous Forest Biome (CFB) project. This was administered by the International Biological Program (IBP), and was funded primarily by the National Science Foundation. The CFB project was short-term, and monitoring of these plots came under the auspices of the

LTER network after the conclusion of sampling through IBP. Motivating interests for the establishment of these plots were broad, and included the study of stand composition and structure, and population and ecosystem dynamics.

Due to the long time period over which plots in the LTER network became established, and to variation in the administration of plot monitoring and study objectives, the sampling designs employed in the laying out of LTER plots are inconsistent. They can generally be classified in one of the following three ways:

> “They are contiguous rectangles subjectively placed within an area of homogeneous forest, circular plots subjectively placed within an area of homogeneous forest, or circular plots systematically located on long transects to cover an entire watershed, ridge or reserve. Rectangular study areas are mostly 1.0 ha or 0.4 ha in size, but range from 0.25 ha to 4.7 ha. Circular plots are 0.1 ha.” (Acker et al., 1998)

Minimum diameter at breast height (DBH) measured in the LTER plots was - apparently in all cases - 5 cm . Plots were measured every 5 or 6 years. In some cases, mortality was recorded annually.

2.0.2 SMCNW

The Stand Management Cooperative Northwest data are from the Levels of Growing Stock (LOGS) study, a collaborative effort between federal, state, and industrial organizations. The LOGS project began in 1962. The plan was developed by Weyerhaeuser Company, with help from the Pacific Northwest Research Station of the Forest Service. The objective was to examine the effect of varying levels of residual growing stock on volume production under eight (repeated) thinning regimes in young Douglas-fir (*Pseudotsuga menziesii*) stands. 27 square 0.08094 ha plots were sampled in each of nine installations at various locations in western Oregon, Washington and British Columbia . Each of the nine installations is a repeated measures completely random split-plot design with three replications of eight different thinnings (treatments), and three untreated control plots.

2.0.3 TP73

These data are from a study of plant biomass dynamics following logging and site preparation (burning) in Watersheds I and II of the HJ Andrews Experimental Forest. Parallel transects were placed along slope contours, with 0.025 ha circular plots every 30.5 m. 132 plots in Watershed I; 61 plots in Watershed II. All trees at least 1.37 m in height (dbh) were measured. There is thus no minimum diameter, per se.

2.0.4 TV038

Data from this source represent a long-term study comparing the growth and development of red alder with that of conifers, primarily Douglas-fir. Beginning in 1941, plots were sampled every 5 years until 1956, and then at irregular intervals through 1996. Stand names include CH11, CH15, CH17, CH22, and CH23. Plot layout: two 0.4047 ha plots and two 0.20235 ha plots. One was placed in each of the following stand types: 1) Alder-conifer mixture, unthinned (stands CH11 and CH15); pure conifer, thinned (CH17); pure alder, thinned (CH22); and pure alder, unthinned (CH23). Diameters were measured in 2.54 cm classes, with the minimum being the 5.08 cm to 7.62 cm class.

The data were cleaned and compiled as a single tree file by Robert J. Pabst, Research Assistant in the Department of Forest Resources at Oregon State University. Repeated measurements were then grouped by plot, and trees with status=2 (ingrowth) were counted by plot and year. Counts of regeneration at time $t + 1$ were paired with plot measurements from the previous (temporal) sampling (t) for prediction. Two variables related to mortality counts contain information from observations occurring over the *preceding* measurement period (time $t - 1$).

2.1 Inconsistencies

Due to the inconsistency in the four sampling designs described in the previous chapter, and to changes in sampling designs *within* studies over time, a great deal of exploratory analysis was required in order to ascertain whether variables were defined consistently throughout. Some of the problems encountered are identified in the following sections.

2.1.1 Definition of Response Variable

The common definition of forest regeneration is *the number of trees of a known diameter at breast height (or in some cases, root-collar) entering a plot since the previous sampling time (Shifley et al., 1993).*

This definition assumes a consistent minimum diameter threshold[1] throughout the data. In the fitting data, MDT varies between 0 cm and 5 cm. The number of observations represented in each MDT class are shown in Table 2.2.

In addition to MDT, plot remeasurement time was also inconsistent, both between and within studies. Although the median number of years between plot visits was 5, and the mean

[1] It is possible for the MDT to be zero.

Table 2.2 Number of observations at each threshold diameter

MDT (cm)	Observations
0	768
0.254	40
1.27	12
3.81	854
4.064	96
5	3232

5.46, there was a non-ignorable amount of variation in the plot remeasurement time. Figure 2.1 shows box-and-whiskers plots for each study, with box widths proportional to the number of observations represented by each. The long horizontal line shows the complete data median (5 years).

Although the medians of *all* studies are contained within the overall inter-quartile range (IQR) of [5,. . . ,6] years, it is important to note that the medians of two of the four largest studies, DFGY and TP73, are 4 and 6 years, respectively. Because of the high rates of change of early-age regeneration counts over time, the effects of one and two years' inconsistency in remeasurement time should not be underestimated. The overall mean regeneration rates for plots with 4, 5, and 6 years between plot visits are 6.4, 3.9, and 2.2 respectively.

The potential difficulty presented by the combined variation in remeasurement time and MDT is depicted in Figure 2.2, which shows the growth curve of a theoretical (slow-growing) tree. Vertical lines on the plot represent years in which the plot was sampled, and horizontal lines correspond to the observed MDT's in the fitting data. Any *intersection* of horizontal and vertical lines represents a slightly different definition of the response variable.

2.1.2 Mortality Counts

Initial construction of the summary regeneration data set was done by distinguishing each unique sampling year in the tree file for a given plot, and then (separately) summing each tree defined as *regeneration* by the variable `status`. These two subsets were then merged, resulting in an n by k matrix[2], in which $n =$ the sum of the total number of sampling years for each plot, and $k =$ the number of predictor variables. The count of trees with $status = 2$ (trees defined as regeneration) was the response variable of interest.

[2] A data frame, in the statistical computing package, `R` (R Development Core Team, 2004).

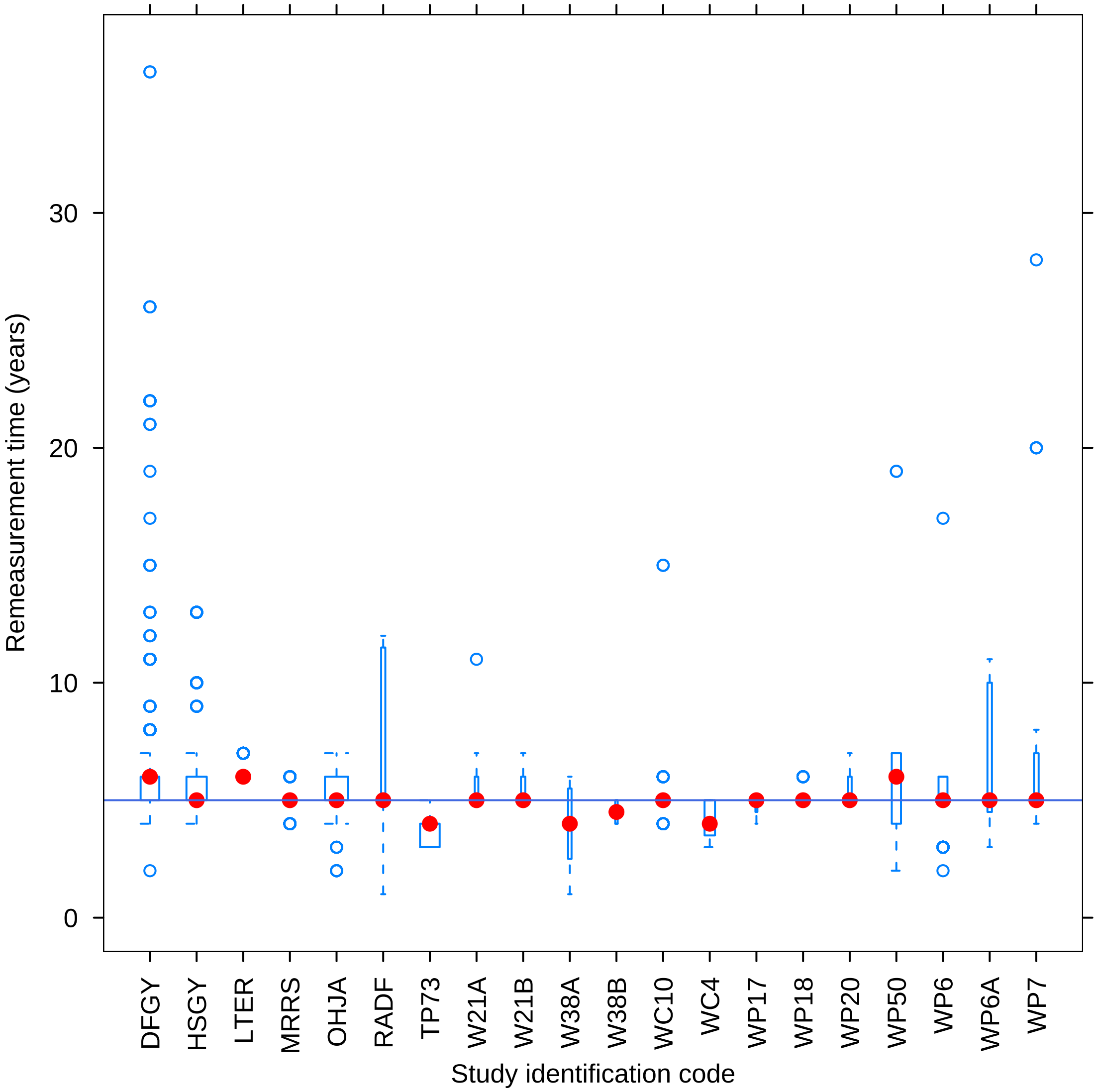

Figure 2.1 Variation in plot remeasurement time

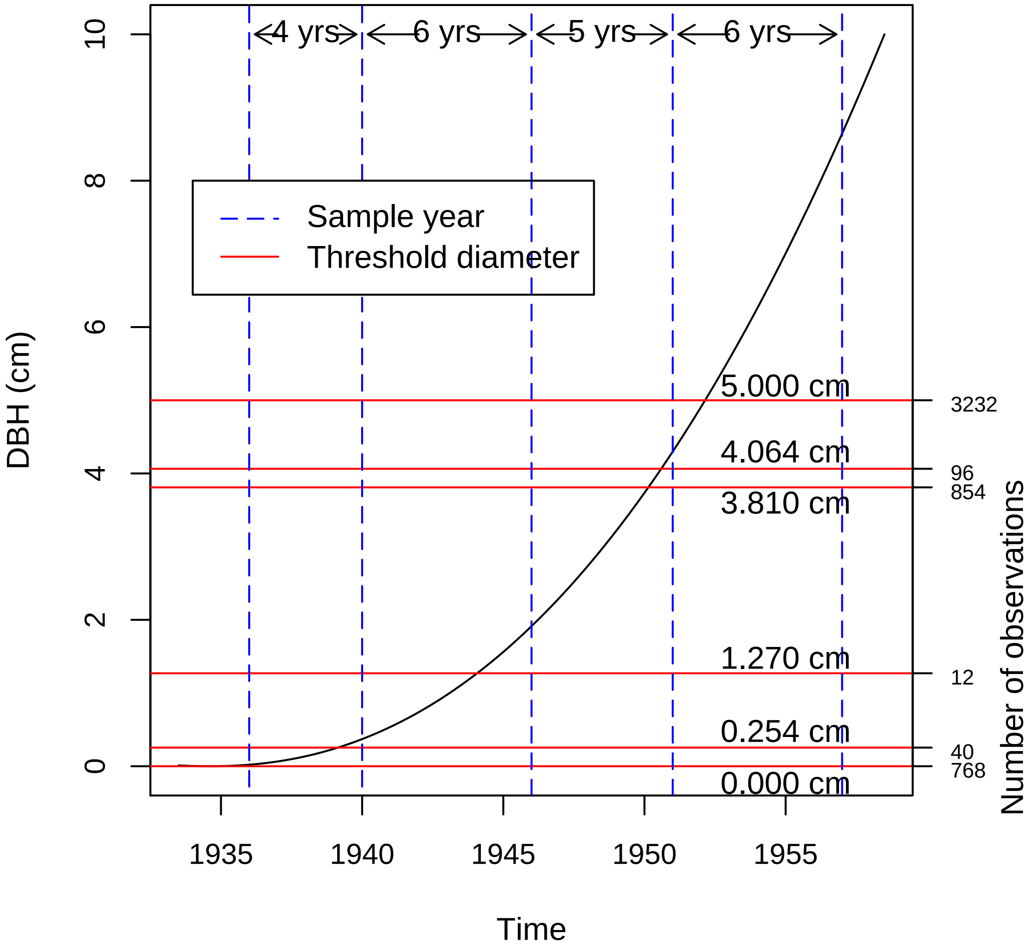

Figure 2.2 Variation in minimum diameter threshold and remeasurement time

Initial exploratory data analysis and model fitting revealed that a pattern of inconsistency existed with four variables: `tph`, `qmd`, `totbamha` and `rd`. There was no observed regeneration (all counts were 0) in years with missing values for these variables. *Mortality checks* conducted annually in the HJ Andrews Experimental Forest (OHJA) included in that component of the data had been sorted and treated as years in which *all* variables on a given plot had been measured[3].

2.1.3 Missing Values in Key Predictors

The PNW coast fitting data have large numbers of missing values in three abiotic variables that have traditionally been treated as integral predictors of regeneration. These are slope, aspect and elevation. Table 2.3 shows the total number of observations for each study, and the total number of values missing for each of these variables.

Table 2.3 Missing values in abiotic predictors

Study	Slope	Aspect	Elevation	Observations
DFGY	7	7	4	694
HSGY	0	0	0	824
LTER	0	326	0	326
MRRS	0	64	0	240
OHJA	0	39	0	1084
RADF	0	0	0	35
TP73	0	0	768	768
W21A	0	0	0	26
W21B	0	0	0	38
W38A	0	0	0	24
W38B	0	0	0	12
WC10	168	168	0	168
WC4	216	216	216	216
WP17	8	8	0	8
WP18	96	0	0	96
WP20	0	0	0	30
WP50	170	170	170	170
WP6	159	0	0	159
WP6A	40	40	0	40
WP7	0	0	0	44

[3] This coding error was identified with help from Rob Pabst.

2.2 Possible Solutions

In the previous section, several problems associated with the fitting data were presented. These can be split into two general groups, as follows:

- Those having to do with the response variable.
- Those having to do with predictors.

The first group contains variable minimum diameter threshold and remeasurement time, and is problematic primarily because it results in violation of the assumption of independent error terms. Solutions to these problems have to do with model form, and will be dealt with in subsequent chapters.

The second group results from some predictors simply not having been measured in one study or another, and it is this group that is the focus of this chapter. Three potential methods for dealing with missing predictors are considered, and the best method is chosen.

2.2.1 Excluding Observations

One solution to the problem of observations missing values in one or more variables is simply to exclude these rows from model fitting. This method has been employed extensively in forest modelling, such that it might well be considered the status quo. However, excluding partial observations eliminates otherwise useful data and can introduce bias.

Were it possible to assume that `slope`, `aspect`, and `elevation` values were missing completely at random (MCAR) or even missing at random (Harrell, 2001), then excluding observations in which one or more of these were not available would have little effect on overall model bias.

The MAR assumption would be a poor one, however. It is evident from Table 2.3 that the missingness of these variables consistently resulted from the sampling design and protocol of studies representing large, distinct components of the overall data. Thus, excluding observations with missing values would likely result in model bias as variation in other, non-missing, predictors present in the excluded studies would be misrepresented.

Like mortality, forest regeneration is a rare phenomenon. Counts of ingrowth on plots occur sporadically over large temporal scales, and these counts tend to be small. Zero counts occur frequently. For this reason, losing data due to missing values in one or two predictors is particularly undesirable, and the costs and benefits of any decision to do so should be considered in detail.

It was decided early on in this project that simple exclusion of observations with missing values should be employed only when absolutely necessary. An example would be a plot or stand for which there was irreconcilable measurement, transcription or conversion error for the minimum diameter threshold of the response variable. Preliminary modelling efforts conducted with observations containing missing values in `slope`, `avAspect` or `elev` suggested that the loss of observations due to missing abiotic predictors was wasteful as these predictors were not important in initial predictive models constructed.

2.3 Practical Constraints on Model Form

The model building process was largely guided by the intended application. The model must be able to make predictions based on data input into the Pacific Northwest Coast variant of the Forest Vegetation Simulator, and it must return predictions on temporal and spatial scales appropriate to that application.

The general, ideological approach to model building was as follows:

1. Fit the best predictive model given the data, regardless of final application.
2. Consider practical constraints of having that model function within final application framework.
    - Does FVS include important predictors?
    - How can the model predict on a ten year scale?
    - What threshold indicators (in FVS) will initiate the regeneration model?
    - What upper boundary indicators can be used to constrain prediction to biologically realistic levels of regeneration?
3. Assess costs of making these changes to original model, and present them as cautionary information to user.

2.4 Initial Framework

Ferguson's Regeneration Establishment Model (Ferguson and Carlson, 1993; Ferguson and Crookston, 1991) predicted regeneration as a *two-stage* or *hurdle* model, following the methods described by Hamilton (1974). The form was as follows:

1. Use logistic regression to predict the probability of a plot being stocked with at least one seedling.
2. Conditional on the plot being stocked with at least one seedling, model the distribution of counts on stocked plots assuming a two-parameter Weibull distribution.

Ferguson's two-stage approach is sensible given the inherent nature of the response variable. Ingrowth occurs infrequently, so it is common for forest regeneration data to contain large numbers of plots with zero counts. These extra zeroes result in situations in which the data are substantially over-dispersed relative to a given distribution that might otherwise be a logical candidate for the non-zero data.

The two-stage approach is sound, and has useful inferential benefits for the biological phenomenon of interest. In particular, it allows for distinction between inference related to why, when, how, and where regeneration takes place (the first stage, or hurdle), and how the distribution of regeneration varies with other predictors (second stage).

2.5 Detection of Potential Outliers

The methods employed in the detection of potential outliers for a given regression component of the final (complete) model followed those suggested by Neter et al. (1996). Neter et al. classify extreme observations as *Y outliers*, *X outliers* or *X and Y outliers.*

2.5.1 Outliers in the Y direction

In order to identify `Y` outliers, plots of studentized residuals and studentized deleted residuals were examined. Studentized residuals are error terms divided by their standard deviations, or

$$r_i = \frac{e_i}{s\{e_i\}} \tag{2-1}$$

Deleted residuals are the differences between observed values of y, and the values of y predicted by a model that is fit without each point included , or

$$d_i = Y_i - \hat{Y}_i \tag{2-2}$$

where $\hat{Y}_i$ is predicted by a model fit *without* Y_i included in estimation. The *studentized deleted* residual, denoted t_i, is then

$$t_i = \frac{d_i}{s\{d_i\}} \tag{2-3}$$

2.5.2 Outliers in the X direction

In order to identify potential X outliers, leverage values were employed. The diagonal elements, h_{ii} of a model fit are "a measure of the distance between the X values for the ith case and the means of the X values for all n cases." (Neter et al., 1996). Neter et al. (1996) recommend examining in detail any observations for which the leverage is greater than twice the *mean* leverage, $\bar{h}$. Because the sum of all individual h_i's is equal to the number of model parameters ($0 \leq h_{ii} \leq 1$), the mean leverage statistic, $\bar{h}$, is equal to p/n:

$$\bar{h}_{ii} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n} \tag{2-4}$$

and leverage values greater than $2p/n$ indicate suspect X values.

2.5.3 Influence

The preceding paragraphs were concerned only with *identifying* potential X and Y outliers. What is truly of interest is determining the magnitude of the effect that these observations have on a regression model (Neter et al., 1996). Cook's Distance was used to determine the influence of observations initially screened as outliers.

Cook's Distance, D_i takes into account the effect of ignoring a given observation on *all* predicted values:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_i - \hat{Y}_{j(i)})^2}{pMSE} \tag{2-5}$$

By default, `R` calculates Cook's Distance for most component families of models (e.g., linear, Generalized Linear, etc.) utilized in this study, and this was the primary statistic used to determine the influence of outliers. D_i was assumed to approximate an F distribution with n and $n - p$ degrees of freedom, and observations for which D_i came close to the 50th percentile of $F_{n,\ n-p}$ were considered influential (Neter et al., 1996).

3.0 Predicting the Probability of Stocking

Ferguson used simple logistic regression to predict the probability of a given plot being stocked with at least one seedling.

3.1 Model Form and Fitting Techniques

Initial prediction of the probability of a plot being stocked with at least one seedling generally followed this example. Probabilities were predicted in a generalized linear model (GLM) context, using iteratively weighted least-squares (IWLS).

The simple binary logistic model with only fixed-effects and no interaction terms is shown in Equation 3-1.

$$Pr(Y = 1|X) = \frac{1}{1 + e^{-X\beta}} \tag{3-1}$$

where X is a vector of predictor variables and β is a vector of coefficients.

The exponential family member was the binomial distribution, and the link function was the canonical link (logit). The logit link function is shown in Equation 3-2:

$$\eta_i = log\left[\frac{\pi_i}{1 - \pi_i}\right] = log\left[\frac{\mu_i}{1 - \mu_i}\right] \tag{3-2}$$

The logistic model retains the assumption of independent error terms. However, given that the canonical link function was used, they are assumed to be binomially distributed.

The inclusion or exclusion of model terms was based on the following assumptions:

1. A conceptual model of the biological and physical processes leading to forest regeneration should preclude and guide the initial inclusion of terms.

2. Given that the large volume of fitting data tends to result in models with proportionally large numbers of degrees of freedom, statistical significance alone is not reason enough to include (or retain) predictors.

3. Taking **1** and **2** into account, inclusion of terms was based on a demonstrated reduction of model deviance relative to the null deviance (and taking into account the variation in the number of degrees of freedom for largely different model forms).

4. Term-wise transformations of each potential predictor were checked, in order to isolate potentially parabolic, logistic, or otherwise non-linear relationships between response and predictor.

5. In all cases, graphical methods were employed in order to seek out otherwise non-obvious (but important) relationships between the response and predictor, and within and between predictors (Harrell, 2001).

6. In the event that there is reason to believe error terms may not meet the assumption of independence, correctly specifying their structure must preclude inference.

7. All else being equal, a simple model is assumed to be preferable over a complex one.

All models were fit using the statistical computer program `R` (R Development Core Team, 2004). Analysis of Deviance was used to test the significance of terms within a given model. The deviance, D, of two models with p and t parameters, respectively, assuming the latter is nested within the former, is approximately χ^2 distributed with $p-t$ degrees of freedom (Lindsey, 1997). Competing model comparison was thus performed in `R` with a χ^2 test argument included in the call to the `anova.glm` function. All terms of `fit.0` were contained in `fit.1` if step-wise comparison was proceeding forward, and the opposite if comparison was conducted backwards. As a similar, likelihood-based metric of within-model term contribution, Akaike's Information Criterion (AIC) was considered. The `stepAIC` function in `R`'s `MASS` library (Venables and Ripley, 1994) was used to compute the AIC for all generalized linear models.

4.0 Modelling the Conditional Distribution of Stocked Plots

Ferguson and Carlson modelled the conditional distribution of stocked plots assuming a two-parameter Weibull density. They asserted that the distribution of trees (per stocked plot) is of greater utility than simple prediction of the *mean.* This is because a simple arithmetic mean would largely underpredict response classes with the highest probabilities of occurrence (plots with only one or a few ingrowth trees), and these are of primary interest.

The Weibull probability density function, shown in Equation 4-1, is a simplified form of the three-parameter Weibull with the location parameter assumed to be equal to 0.

$$f(X) = \frac{\beta}{\eta}\left(\frac{X}{\eta}\right)^{\beta-1} e^{-(\frac{X}{\eta})^{\beta}} \tag{4-1}$$

where η is the scale parameter, and β is the shape (or slope) parameter.

Ferguson and Carlson split the data into 96 groupings, and – assuming there were at least 25 plots within each group – estimated unique values of η and β for each. They also calculated the mean values of potential plot-level predictors within each grouping. Then, simple linear models were used to predict the Weibull parameters from mean site characteristics within each grouping. In this way, a nice framework by which to predict stocking *within FVS* was developed. The 96 categories resulted from clustering all combinations of the following divisions:

- 4 habitat type series: *Pseudotsuga menziesii*, *Abies grandis*, *Thuja and Tsuga*, and *Abies lasiocarpa*.
- 4 Aspect classes: north $+/-$ 45 degrees, east, south, and west.
- 3 Years-since-disturbance classes: 2-7 years, 8-12 years, and 13-20.
- 2 Temporal spruce budworm (*Choristoneura fumifera* (Clemens)) outbreak classes: 0-2 years, and 8-12 years.

4.1 Necessary Adaptations for Pacific Northwest Coast Variant

In keeping with the status quo, the Weibull was regarded as a good starting point for predicting the distribution of stocked plots. Given the large sample size and the discrete response variable, other potential model forms are the Poisson and Negative Binomial. Based on the large variance to mean ratio evident in the data, the Negative Binomial was considered to be the most appropriate model.

The utility of these probability distributions – the Poisson and the Negative Binomial, in particular – was initially explored, though the Weibull provided the best fit to the complete data. Early on, it became evident that it was necessary to make several simplifications of the existing (e.g., North Idaho) ingrowth model. These included the following:

- Alternative grouping of data by forest characteristics (within which parameter prediction took place).
- Elimination of the treatment of ingrowth as *advanced*, *subsequent*, or *excess* regeneration.

Each of these changes is discussed in further detail in the following sections.

4.1.1 Grouping of Data

Due to the very different structure of the fitting data from that used by Ferguson and Carlson in building the North Idaho regeneration model, it was obviously necessary to develop alternative criteria for subsetting plot characteristics into groups. For example, there was no record of spruce budworm defoliation in the PNW fitting data, nor was there a well-represented record of disturbance history. Moreover, as mentioned in Chapter 1, a noticeably large proportion (20.8 %) of the `avAspect` variable were missing.

4.2 Fitting Techniques

Initial model choice was based on Chi-squared goodness-of-fit tests for several reasonable candidate models. The counts of each unique number of ingrowth trees occurring on (non-zero count) plots were tabulated. The expected response of each count's occurrence under the assumptions of Poisson, negative binomial, and Weibull distributions was calculated, and the sum of squared residuals (*observed* - *expected*) were divided by the *expected* value for each cell (or level). The sum of these deviances was compared with the critical value for a χ^2 distribution with $k-1$ degrees of freedom, where k was the number of unique regeneration counts.

Parameters for each distribution were estimated using maximum likelihood. In the case of the Poisson distribution, the mle is the sample mean. The negative binomial and Weibull likelihoods were maximized using Ripley's `fitdistr` function (Venables and Ripley, 1994), which in turn calls `optim`. The maximum likelihood estimates for the size (n) and μ (mean) parameters of negative binomial approximation to the complete, non-stocked data were 0.76 (s.d. = 0.02) and 8.47, (s.d. = 0.22) respectively. The maximum likelihood estimates for the shape and scale parameters of the Weibull distribution were 0.77 (s.d.= 0.01) and 6.95 (s.d.=0.21), respectively.

The number of counts having expected frequencies less than 5 for the Poisson model was 74. These were grouped (separately) for the Poisson, negative binomial, and Weibull models in order to meet that criterion of the χ^2 test. The fit of the Poisson was quite poor, and that of the Weibull was better than the negative binomial.

The expected numbers of regeneration trees occurring at each observed level of stocking were predicted, conditional on the assumption of Poisson, negative binomial, and Weibull distributed response. Figure 4.1 shows the observed counts at each level, and the predicted counts for each model. Although the Poisson and negative binomial are discrete distributions, lines connecting predicted (discrete) values have been added to aid ease of distinction in areas where values predicted by the models overlap severely in the upper tails.

It is reasonable to assume that the same clustering of the response variable within data source, study, stand and plot that was addressed in predicting the probability of stocking should also be accounted for in modelling the distribution of stocked plots. The assumption of independent counts across these various grouping levels would be weak.

In order to account for the variability in the sampling designs, the methods of Ferguson and Carlson (1993) were altered slightly. Where they split the stocked plot data into smaller components according to habitat type, aspect and years-since-disturbance classes, and spruce budworm defoliation history, we estimated unique Weibull parameters for each *stand* in the data.

Initially, the *plot* was considered the most desirable candidate splitting variable, in that it would capture the highest degree of variation in potential covariates. An effort was made to estimate unique Weibull parameters for the regeneration counts occurring on individual plots *in time*. However, this approach was found to be undesirable because the number of elements (plot-years) in each grouping tended to be very small, such that the quality of parameter estimation was poor. The standard errors of the ML estimates were approximately as large as the estimates themselves.

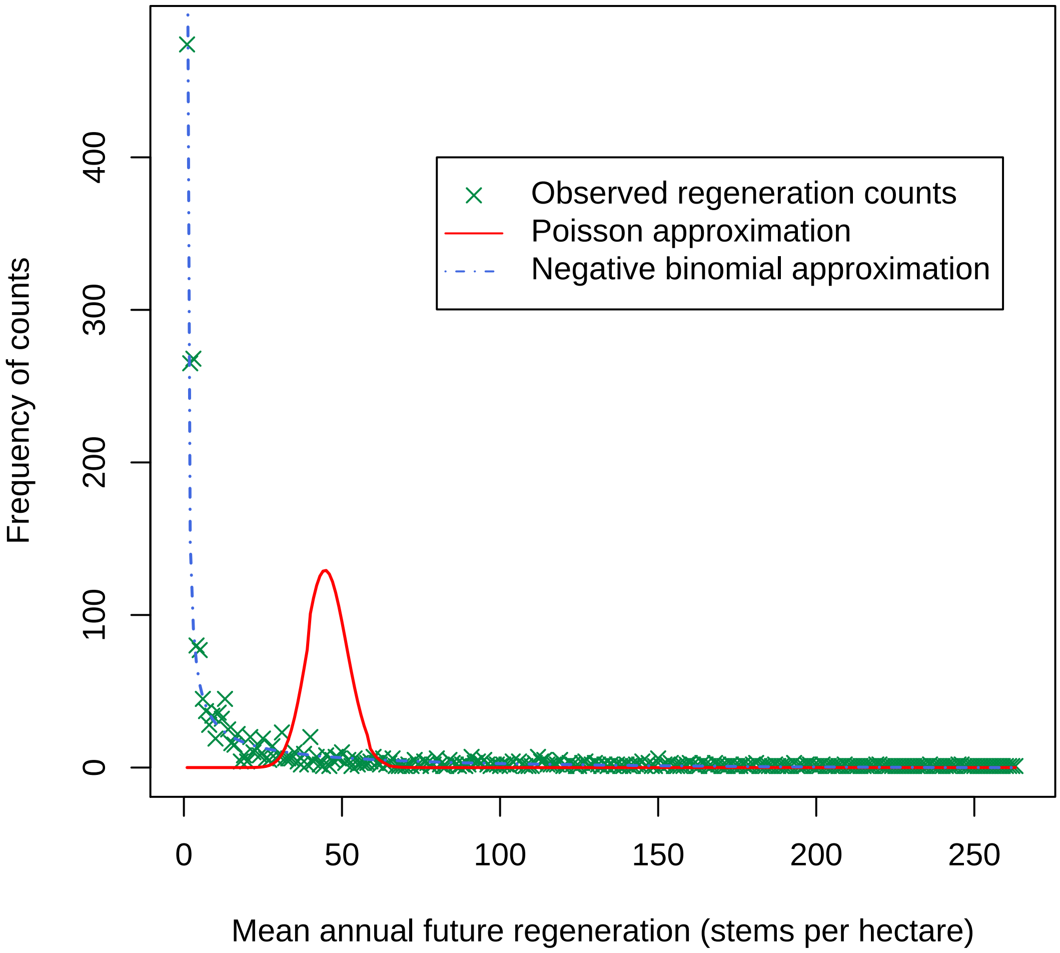

Figure 4.1 Observed and expected counts of regeneration assuming Poisson and negative binomial distributed response

PART II: TWO-STAGE MODELS

The following 3 chapters describe alternative two-stage approaches that were tried for model building. Within each section, the important model assumptions are described, the best fitting model is presented - though in some cases there may be more than one - and the quality of the best fitting model is assessed through numerical metrics and graphical diagnostics. In order to limit confusion regarding the comparison of different *models* (e.g., $y \sim \beta_0 + \beta_1 x_1$) and different ways of organizing the data used in fitting these, I will use the convention that the former are referred to as *models* and the latter as *structures*.

As with several other areas in the analysis, a few alternative solutions were examined here in order to resolve violations of model assumptions stemming from heterogeneity in the survey designs from which the fitting data were compiled.

In particular, two potential solutions to the problem of variable minimum diameters were examined, as was an alternative definition of the response variable. The fitting data in each chapter is referred to as a Structure, and are enumerated as follows:

- Exclude all trees[1] with less than a constant minimum diameter threshold, and fit the best two-stage model given the criterion listed in the previous sections.
- Leave trees of all diameters in the analysis. Include the term `minEst` (estimated minimum diameter threshold) as a covariate, and fit the best model based on the criterion listed in the previous section.
- Redefine which trees are considered *regeneration* based on a new set of criteria, and model the standardized mean response for a given plot, rather than the simple *count*.

There was a somewhat chronological, evolutionary basis for the sequence of the structures[2]. Although Structure 3 is, in principle, the most highly developed approach, the earlier attempts are presented as distinct, credible efforts. The analyses have been re-visited several times in the writing of this thesis. As would be expected, the different data structures resulted in different amounts of variation being explained by individual covariates.

[1] Not plots.

[2] An evolutionary basis that was not independent of the author's increasing familiarity with the complexity of the data, nor of his education in basic statistical concepts.

Table 5.1 Analysis of Deviance, Structure 1

	DF	Deviance	Resid.DF	Resid.Dev	F	Pr(>F)
NULL			5001	6409		
log.qmd	1	21	5000	6389	21	5.4e-06
hab	3	544	4997	5845	181	< 2e-16
log.year	1	135	4996	5710	135	< 2e-16

5.0 Structure 1

5.1 Description and Assumptions

Initial analysis was conducted under the assumption that trees initially defined as forest regeneration in the complete tree file were classified correctly. Trees with the variable `status` equal to 2 were used to build predictive count models. The number of trees defined in this way were tabulated for each year of measurement on each individual plot. All trees with diameters less than the largest minimum diameter threshold (MDT) were *excluded* from analysis.

The plots were treated as independent sampling units selected from a population of possible plots, each of which had an equal and independent probability of being selected.

The best two-stage model of the general form described in the previous chapter was constructed.

For the logistic regression predicting the probability of stocking with only trees greater than 5.0 cm in diameter included in the analysis the important predictors were the log of quadratic mean diameter, habitat type, and the log of the measurement year. The results are shown in the analysis of deviance in Table 5.1.

5.2 Probability of Stocking

Important predictors of the probability of any regeneration stocking for Structure 1 were the log of quadratic mean diameter (`qmd`), habitat type `hab` and the log of the mean year (`year`) of sampling. The coefficients for this model are shown in Table 5.2, and their relative importance in reducing model deviance is shown in Table 5.1.

Table 5.2 Coefficients for logistic model, Structure 1

	Estimate	SE	z	Pr(> |z|)
(Intercept)	-384	35	-11	< 2e-16
log.qmd	-0.095	0.038	-2.5	0.012
habPISI	2.6	0.14	19	< 2e-16
habPSME	1.6	0.22	7.6	4.2e-14
habTSHE	1.9	0.12	16	< 2e-16
log.year	50	4.6	11	< 2e-16

5.3 Conditional Distribution of Stocked Plots

In the second stage of analysis, Weibull distribution parameters (2) were estimated for each of 91 stands (of the 120 total possible) in the data. Then, linear regression models were constructed to predict the Weibull parameters (separately) from the means of potential covariates. Stands excluded from this analysis, either due to not being *stocked* or because of insufficient data, are shown in Table 5.3.

5.3.1 Shape Parameter

Several models predicting the shape parameter were evaluated. No Multiple-R^2 value was greater than 0.13. This model had only the intercept and the *Picea sitchensis* habitat type included. No other terms were significant. A scatterplot matrix of the shape parameter and the log of the shape parameter against all possible predictors revealed no obvious relationships, although clear correlations exist between covariates. The scatterplot matrix of a few possible surrogates for competition effects (basal area, relative density, quadratic mean diameter, and total number of trees) is shown in Figure 5.2.

Scatterplot matrices were used to explore various bivariate relationships. They were used as a tool to visually identify potential relationships (e.g., linear, quadratic, logarithmic) between the response variable – in this case the Weibull shape parameter – and potential predictors. They were also used to visually identify relationships *between* covariates. For example, Figure 5.2 shows that the total basal area per hectare represented by a given plot (**totbamha**) appears to share a fairly positive linear relationship with Curtis' relative density index (**rd**), as we would expect. This suggest a multi-collinearity problem if both variables were to be included in a model. A complete matrix of all predictors was too large to include in this document, but revealed no stronger relationships between `B` (or `logB`) and any covariate.

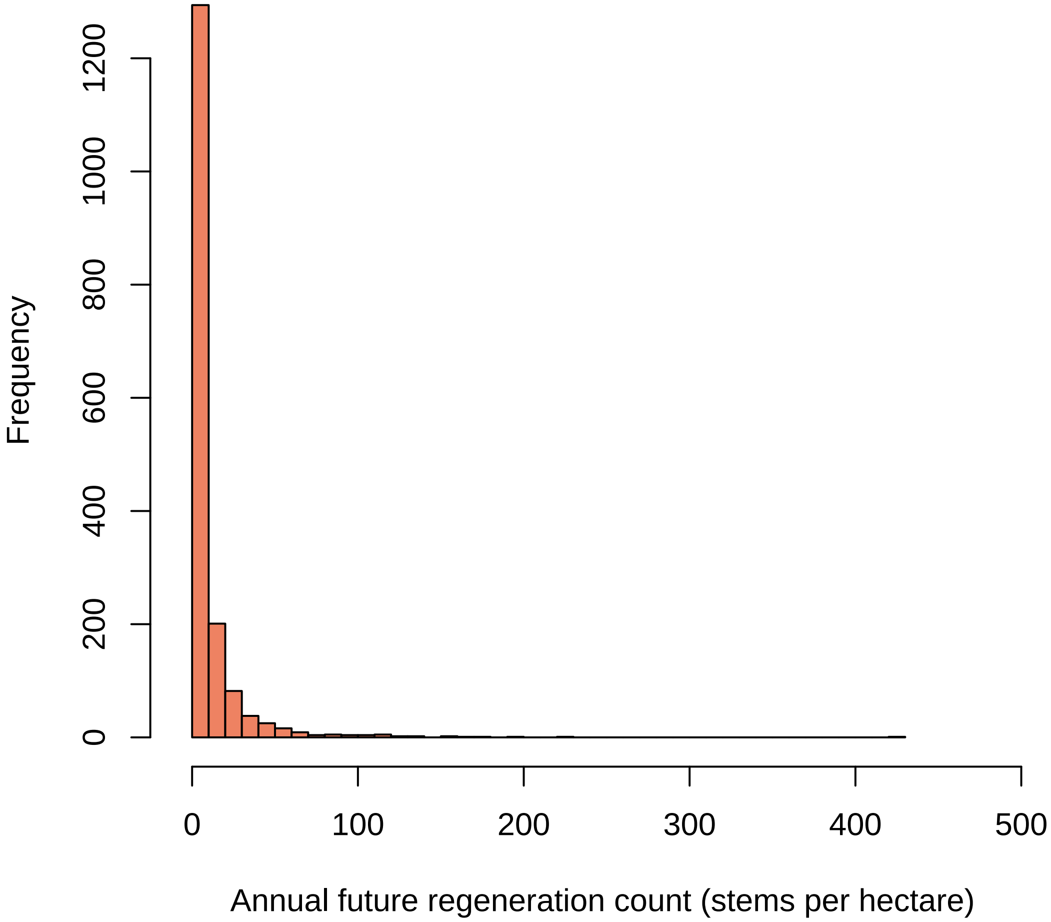

Figure 5.1 Histogram of mean annual regeneration counts per hectare (stocked plots), Structure 1

Table 5.3 Stands excluded from second stage of analysis, Structure 1

No regeneration	Insufficient data
CH06	
CH09	
	CH41
	CH42
GP02	
	RS28S
	RS32L
RS32S	
SI04	
SI05	
	SI06
SI07	
SI08	
SI09	
SI10	
SI20	
	TA01
TB13L	
	TB13S
	TO11L
W21A	
W21B	
W38B	
WC10	
WC4	
WI05	
WP17	
WP20	
WP6A	

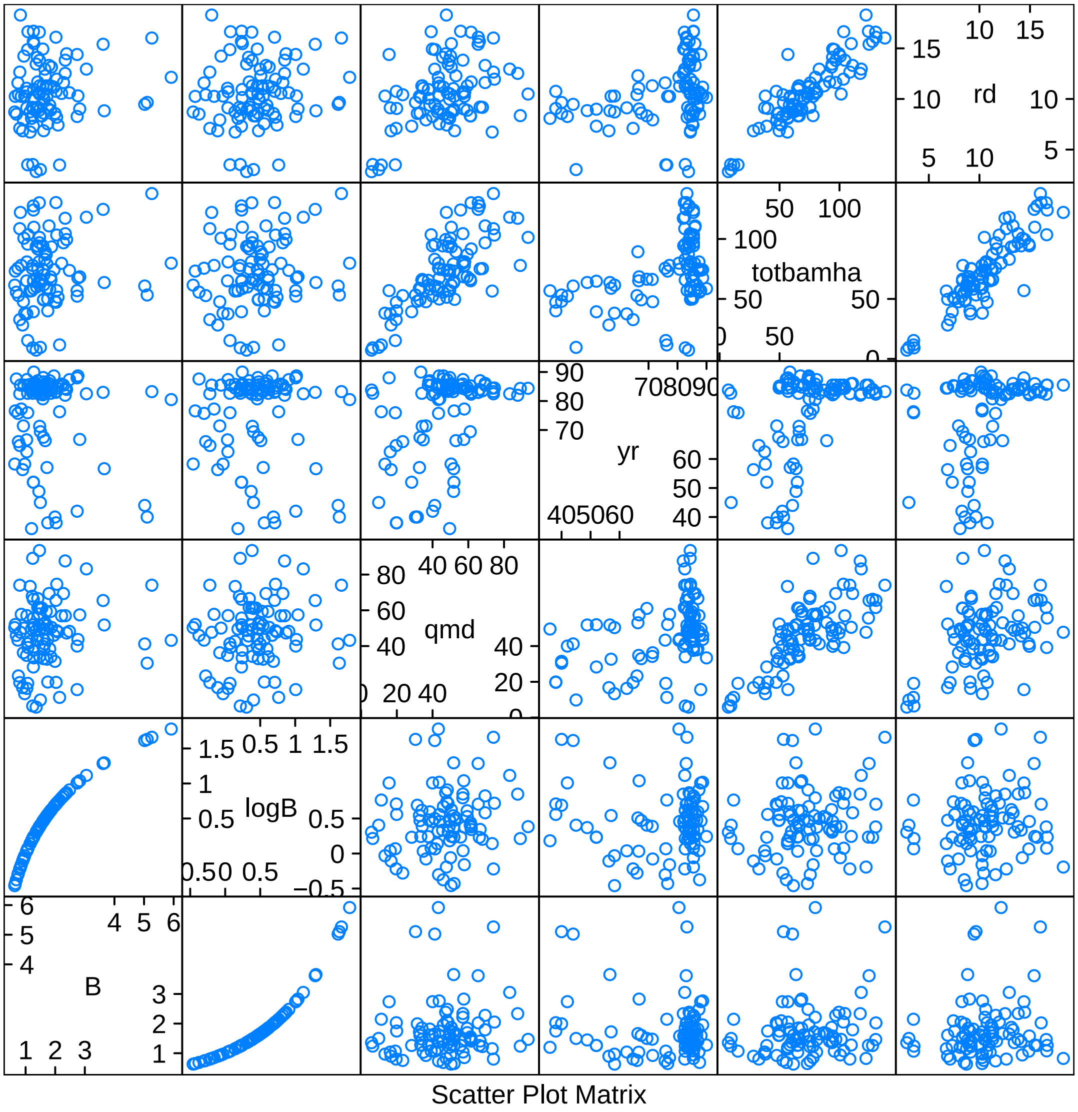

Figure 5.2 Example scatterplot matrix used in shape parameter prediction, Structure 1

An ANOVA table for this model is shown in Table 5.4.

Table 5.4 Analysis of Variance table for model predicting shape parameter, Structure 1

	DF	SS	MS	F	Pr(> F)
habPISI	1	11	11	13	0.00047
Residuals	89	76	0.85		

The coefficients for this model are shown in Table 5.5. Inclusion of terms other than the intercept and the *Picea sitchensis* habitat type would be arbitrary.

Table 5.5 Coefficients for model predicting shape parameter, Structure 1

	Estimate	SE	t	Pr(>|t|)
(Intercept)	2	0.11	18	< 2e-16
habPISI	-0.81	0.22	-3.6	0.00047

5.3.2 Scale Parameter

Prediction of the Weibull scale parameter was initially performed via a Box-Cox transformation of the response variable due to extreme non-normality (based on a Normal QQ-plot of residuals) and heteroscedasticity of the residuals (plotted against fitted values). The value of $\hat{\lambda}$ minimizing the model (negative) log-likelihood was -0.01. Because 0 was contained in the confidence interval on $\hat{\lambda}$, the log transform was used instead. The final model chosen is shown in Equation 5-1. This model, with the log of $\hat{C}$ predicted from the log of quadratic mean diameter, had a multiple R^2 of 0.24 and a RMSE[1] of 8.21. The total number of trees in each plot (variable `ntrees`) was highly significant when included as a covariate and increased the amount of variation explained by the model by almost 100 %. This term was unacceptable because it is not standardized to the per-hectare level. The corresponding per-hectare covariate, trees per hectare (`tph`), was not significant.

$$log(\hat{c}) = \beta_0 + \beta_1 \log \texttt{qmd} \tag{5-1}$$

An ANOVA table for this model is shown in Table 5.6.

[1] Based on the untransformed predicted values.

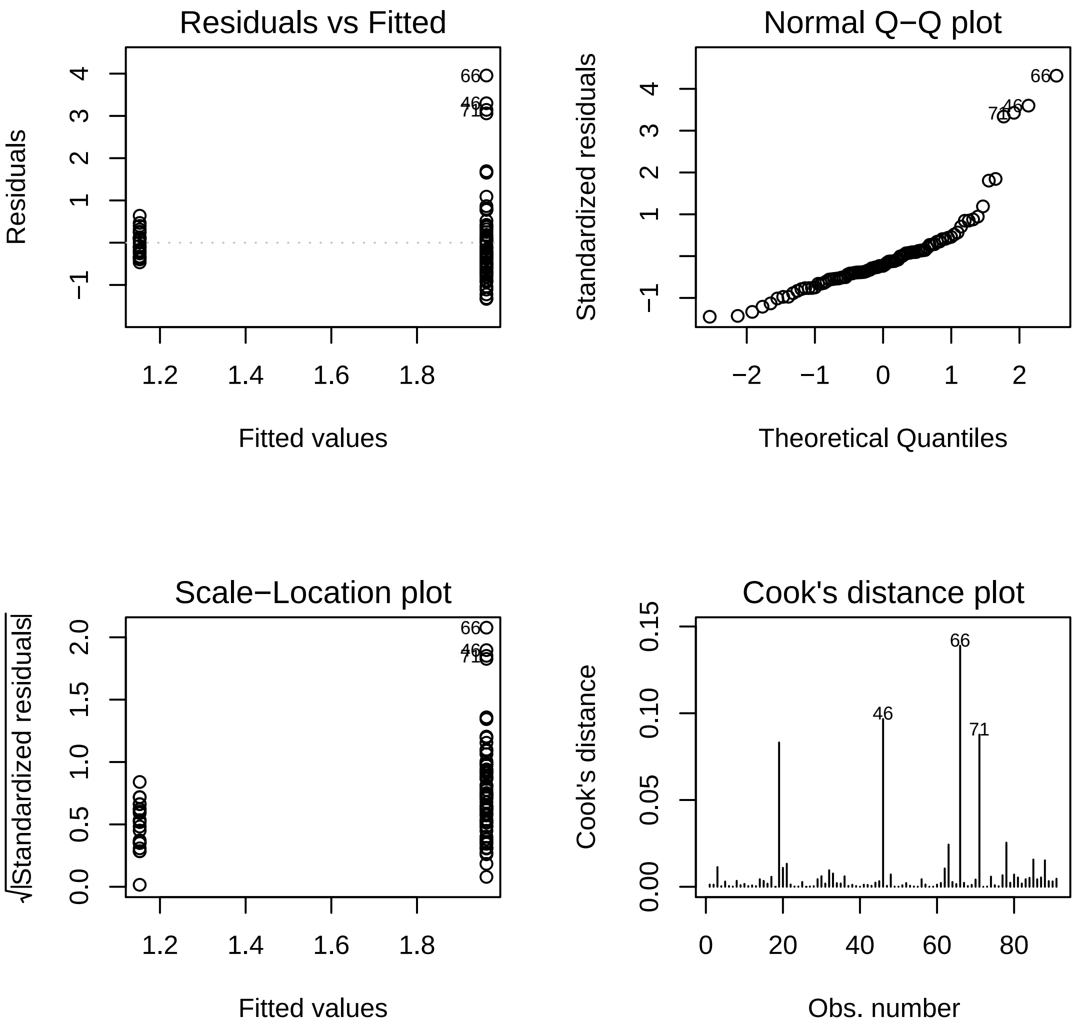

Figure 5.3 Residual plots for model predicting shape parameter, Structure 1

Table 5.6 Analysis of Variance table for model predicting scale parameter, Structure 1

	DF	SS	MS	F	Pr(> F)
log.qmd	1	37	37	28	9.8e-07
Residuals	89	120	1.3		

The coefficients for this model are shown in Table 5.7.

Table 5.7 Coefficients for model predicting scale parameter, Structure 1

	Estimate	SE	t	Pr(>|t|)
(Intercept)	2.4	0.33	7.3	9.4e-11
log.qmd	-0.035	0.0066	-5.3	9.8e-07

Residual plots for this model are shown in Figure 5.4.

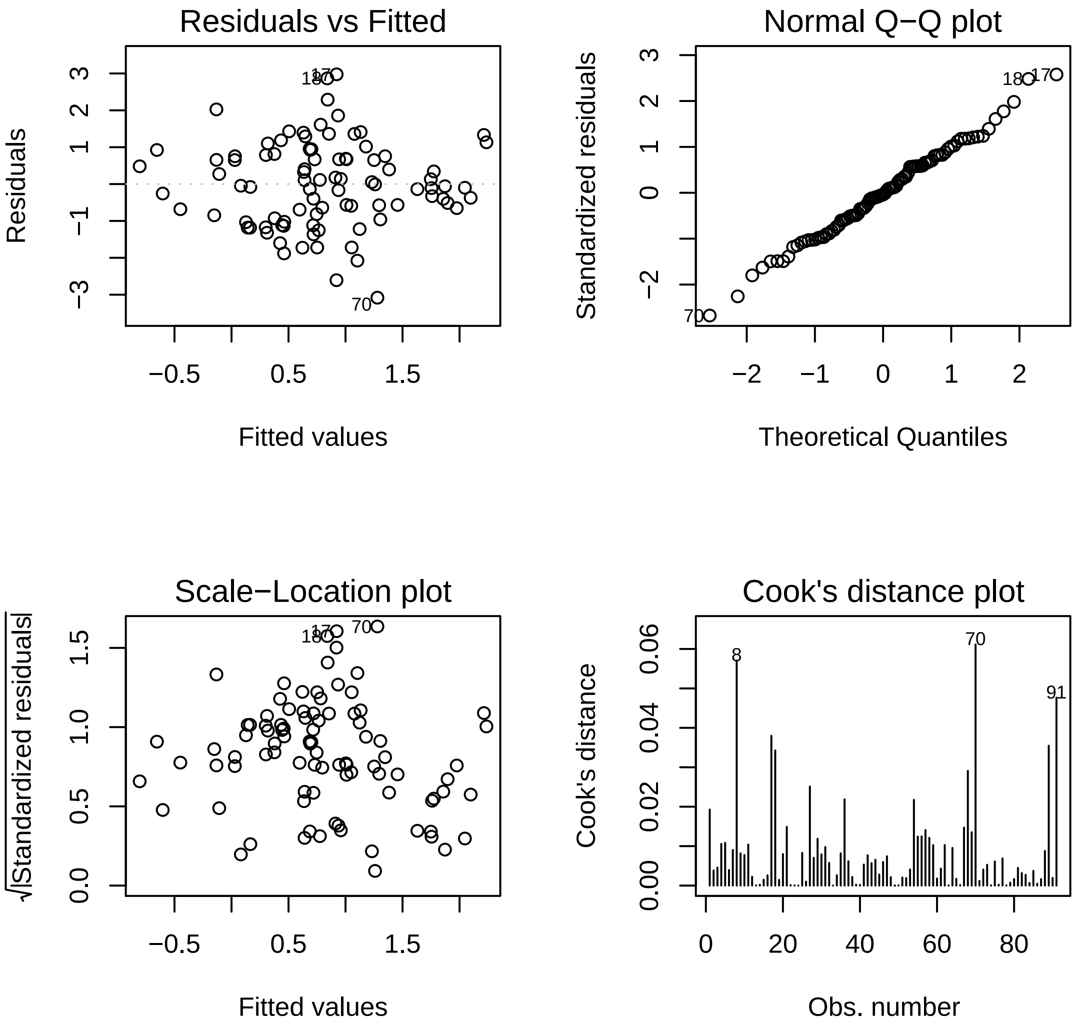

Figure 5.4 Residual plots for model predicting scale parameter, Structure 1

6.0 Structure 2

6.1 Description and Assumptions

In this structure, `all` trees initially defined as regeneration were included in analysis, regardless of minimum diameter threshold. The minimum diameter thresholds for each `stand` (and in some cases, each `plot`) were estimated using primarily visual methods. Regeneration trees[1] were extracted from the large tree file, and their diameters at breast height (`dbh`) were plotted separately by stand or plot[2].

For each of these plots, estimates of the minimum diameter at which trees were measured were made by plotting a simple horizontal line at the minimum observed value of `dbh`, and visually judging that this line was not the result of possible measurement, recording, or tabulation error. These plots were generally quite easy to interpret due the large volume of data. For example, a large dense band of points at 5.0–5.5 cm in many cases (and no points below this band) suggested beyond reasonable doubt that the MDT for the stand in question was 5.0 cm. Each of the 5 levels of the `minEst` variable estimated by this method correspond to common whole number MDT values in either metric or American Standard units.

6.2 Probability of Stocking

For the binomial GLM with all trees included regardless of minimum diameter threshold, the log of quadratic mean diameter, habitat type, and the log of measurement year were the key predictors. All were significant with p-values less than $2.2e - 11$. The results are shown in the Analysis of Deviance table in Figure 6.1. The minimum diameter threshold term (`minEst`) was significant.

Table 6.1 Analysis of Deviance, Structure 2

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			5001	6850		
log.qmd	1	292	5000	6558	292	< 2e-16
hab	3	709	4997	5849	236	< 2e-16
log.year	1	161	4996	5688	161	< 2e-16
minEst	1	45	4995	5643	45	2.3e-11

[1] With variable `status`= 2.

[2] For stands in which earlier information made clear that MDT *within* the stand was known to be inconsistent (Pabst, Pers. Comm.)

The coefficients for this model are shown in Table 6.2.

Table 6.2 Coefficients for logistic model, Structure 2

	Estimate	SE	z	Pr(> |z|)
(Intercept)	-476	36	-13	< 2e-16
log.qmd	-1.7	0.1	-16	< 2e-16
habPISI	2.9	0.14	21	< 2e-16
habPSME	1.4	0.2	6.7	1.8e-11
habTSHE	2.2	0.11	20	< 2e-16
log.year	63	4.7	13	< 2e-16
minEst	0.29	0.044	6.6	3.6e-11

A count histogram of plots having at least one regeneration tree per hectare before the subsequent plot remeasurement is shown in Figure 6.1.

6.3 Conditional Distribution of Stocked Plots

25 stands were not included in the analysis. 16 of these had *no* observed regeneration counts at any time. 9 stands – CH41, CH42, RS28S, RS28L, SI06, TA01, TB13S, TO11L and WC10 – were excluded due to insufficient data necessary for maximization of the log-likelihood function in parameter estimation. Additionally, the standard errors of estimated parameters were compared with the estimates themselves for high coefficient of variation. It was not necessary to exclude any stands based on this criteria.

Scatterplot matrices of the shape and scale parameters (and the log transformations of these) with 14 candidate predictors were examined for potential linear relationships. Although obvious correlations were evident among predictors, relationships between Weibull parameters and covariates were weak to non-existent. An example scatterplot of the Weibull shape parameter, the log of the shape parameter, and 5 possible predictors is shown in Figure 6.2.

Several linear models predicting the parameters, the log of the parameters, and Box-Cox transformations of the parameters were explored.

6.3.1 Shape Parameter

The best model predicting the Weibull shape parameter was a simple linear regression of the Box-Cox transformed shape parameter predicted from quadratic mean diameter and the *Picea sitchensis* habitat type, with no interaction term:

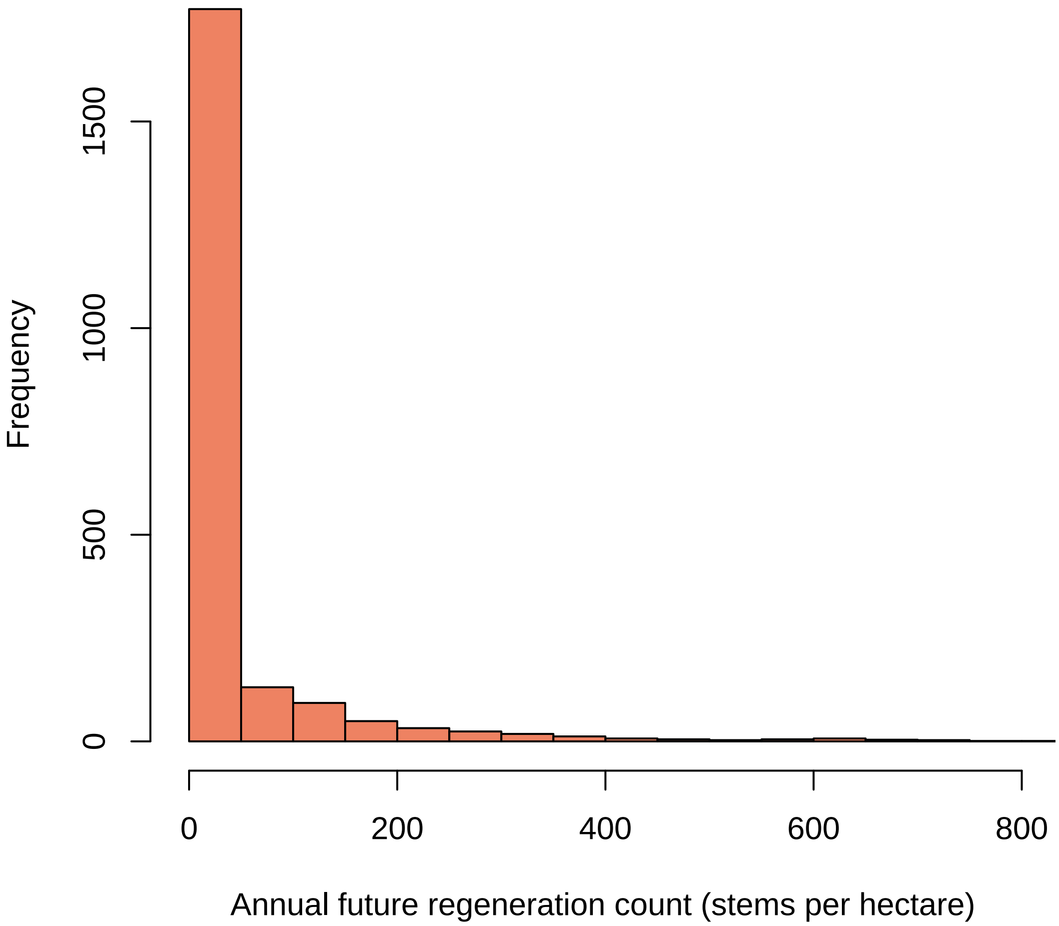

Figure 6.1 Histogram of mean annual regeneration counts per hectare (stocked plots), Structure 2

Table 6.3 Stands excluded from second stage of analysis, Structure 2

No regeneration	Insufficient data
CH06	
CH09	
	CH41
	CH42
GP02	
	RS28S
RS32S	
	RS32L
SI04	
SI05	
	SI06
SI07	
SI08	
SI09	
SI10	
SI20	
	TA01
TB13L	
	TB13S
	TO11L
W21A	
	WC10
WI05	
WP17	
WP20	

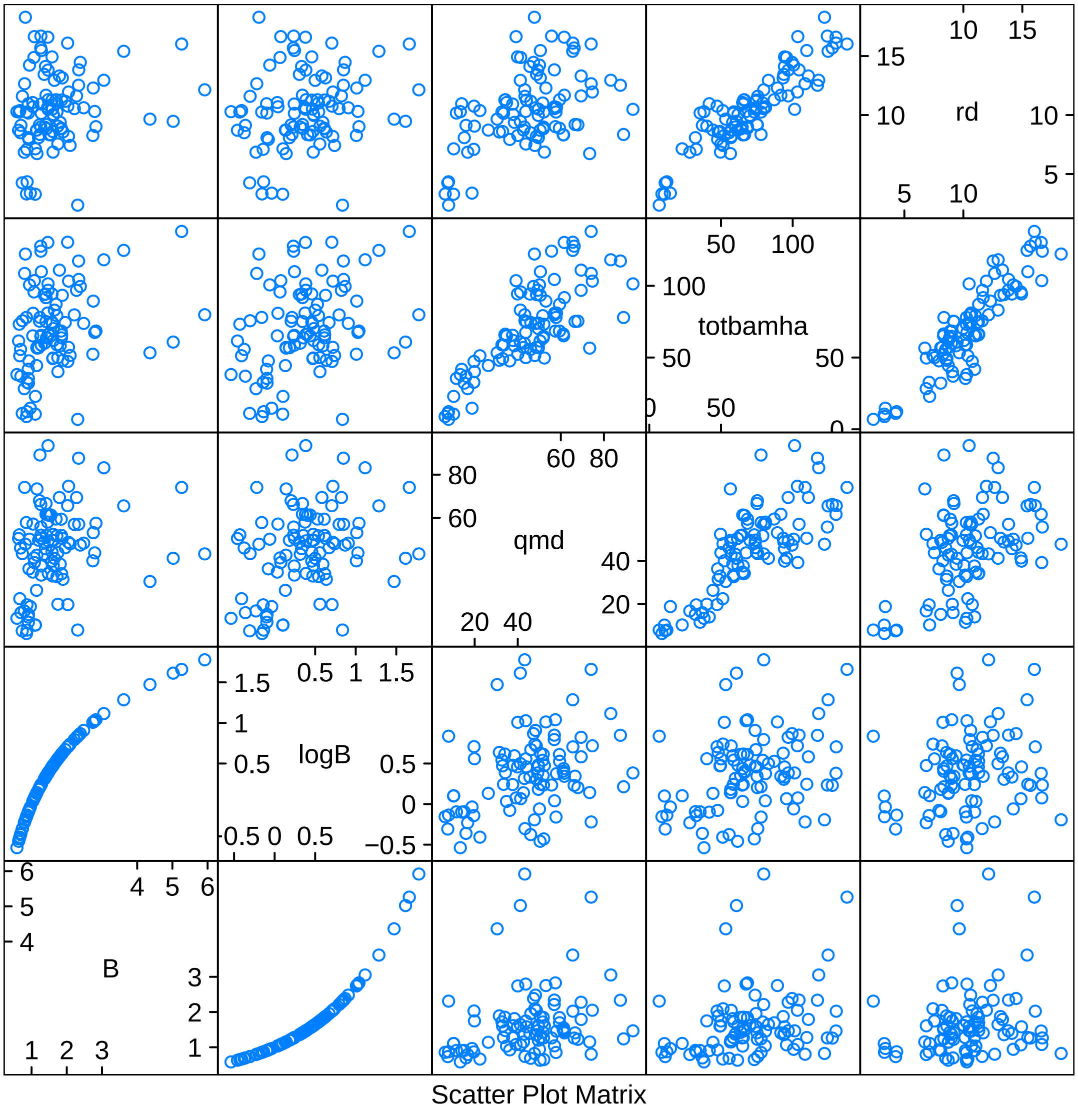

Figure 6.2 Example scatterplot matrix used in shape parameter prediction, Structure 2

$$\hat{b}^{\lambda} = \beta_{0+}\beta_1\texttt{qmd} + \beta_2\texttt{habPISI} \tag{6-1}$$

where $\hat{b}$ is the mle estimate of the Weibull shape parameter, λ is the optimized value of the Box-Cox transformation of b_{ij}, β_0, β_1 and β_2 are coefficients, and `qmd` and `habPISI` are quadratic mean diameter and habitat type (either PISI series or not), respectively.

The coefficients for this model (and their standard errors) are shown in Table 6.4.

Table 6.4 Coefficients for model predicting shape parameter, Structure 2

	Estimate	SE	t	Pr(>|t|)
(Intercept)	0.99	0.037	27	< 2e-16
qmd	-0.0037	0.00075	-5	2.5e-06
habPISI	0.18	0.034	5.4	6.3e-07

The Box-Cox transformation of the shape parameter was used because plots of the residuals for this model against the fitted values revealed a high degree of heteroscedasticity, thereby violating the linear regression assumption of constant residual variance. The Box-Cox transformation substantially improved the residual against fitted plot, as is evident in Figure 6.3.

An ANOVA showing the relative importance of predictors in reducing the mean-squared error for this model is shown in Table 6.5.

Table 6.5 Analysis of Variance table for model predicting shape parameter, Structure 2

	DF	SS	MS	F	Pr(> F)
qmd	1	0.36	0.36	18	5.1e-05
habPISI	1	0.57	0.57	29	6.3e-07
Residuals	92	1.8	0.02		

6.3.2 Scale parameter, Model 1

Models predicting the Weibull scale parameter were generally better than models predicting the shape parameter. However, determining the relative importance of predictors – and thus the best model – was difficult due to their being highly correlated with one another.

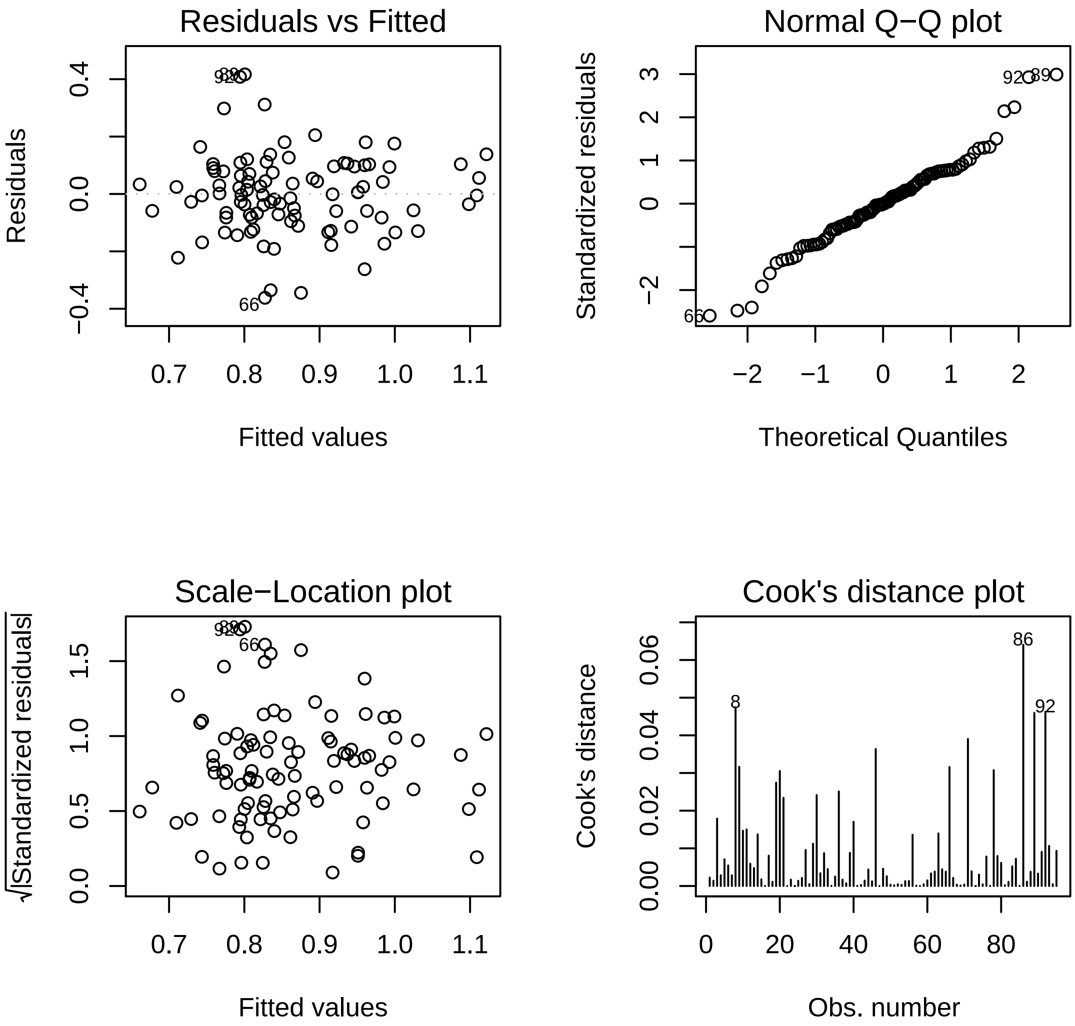

Figure 6.3 Residual plots for model predicting the shape parameter, Structure 2

An initial model was constructed with the log of quadratic mean diameter, the year of sampling, the total number of trees in the stand or plot, and the *Picea sitchensis* and *Tsuga heterophylla* habitat type series:

$$\hat{c} = \beta_0 + \beta_1 * log(\texttt{qmd}) + \beta_2 * \texttt{mean.year} + \beta_3 * \texttt{ntrees} \quad (6\text{-}2)$$

where $\hat{c}$ is the mle estimate of the Weibull scale parameter, β_0, β_1, β_2 and β_3 are coefficients or vectors of coefficients, and all other variables are as previously defined. This model had a Multiple-R^2 of 0.6458. The RMSE for the model was 12.73, indicating high expected average variation in scale parameter predictions.

More importantly, residual plots for the model showed substantial violations of linear regression assumptions. The residual vs. fitted values plot in Figure 6.4 shows extreme heteroscedasticity, signifying failure of the constant residual variance assumption. The Normal QQ plot is not straight, providing evidence that the assumption of Gaussian-distributed error terms is not justified. Plots of individual predictors against `C` (Not shown here) suggested that the fundamental assumption of a linear relationship between X and Y was weak, at best.

The coefficients for this model (and their standard errors) are shown in Table 6.6.

Table 6.6 Coefficients for Model 1 predicting scale parameter, Structure 2

	Estimate	SE	t	Pr(>|t|)
(Intercept)	-1146	190	-6	3.4e-08
log.qmd	-34	2.8	-12	< 2e-16
mean.year	0.64	0.098	6.6	3.1e-09
habPISI	14	4.8	2.8	0.0058
habTSHE	19	4.4	4.5	2.4e-05
ntrees	-0.03	0.0064	-4.6	1.4e-05

An ANOVA showing the relative importance of predictors in reducing the mean-squared error for this model is shown in Table 6.7

6.3.3 Scale Parameter, Model 2

In an effort to resolve violations of the assumed Normality and constant variance of error terms, an alternative model predicting the Weibull scale parameter for Structure 2 was considered. A Box-Cox transformation of the response variable was performed, and the same model discussed in the previous section was fit, with the response (`C`) raised to the optimal value of $\hat{\lambda}$.

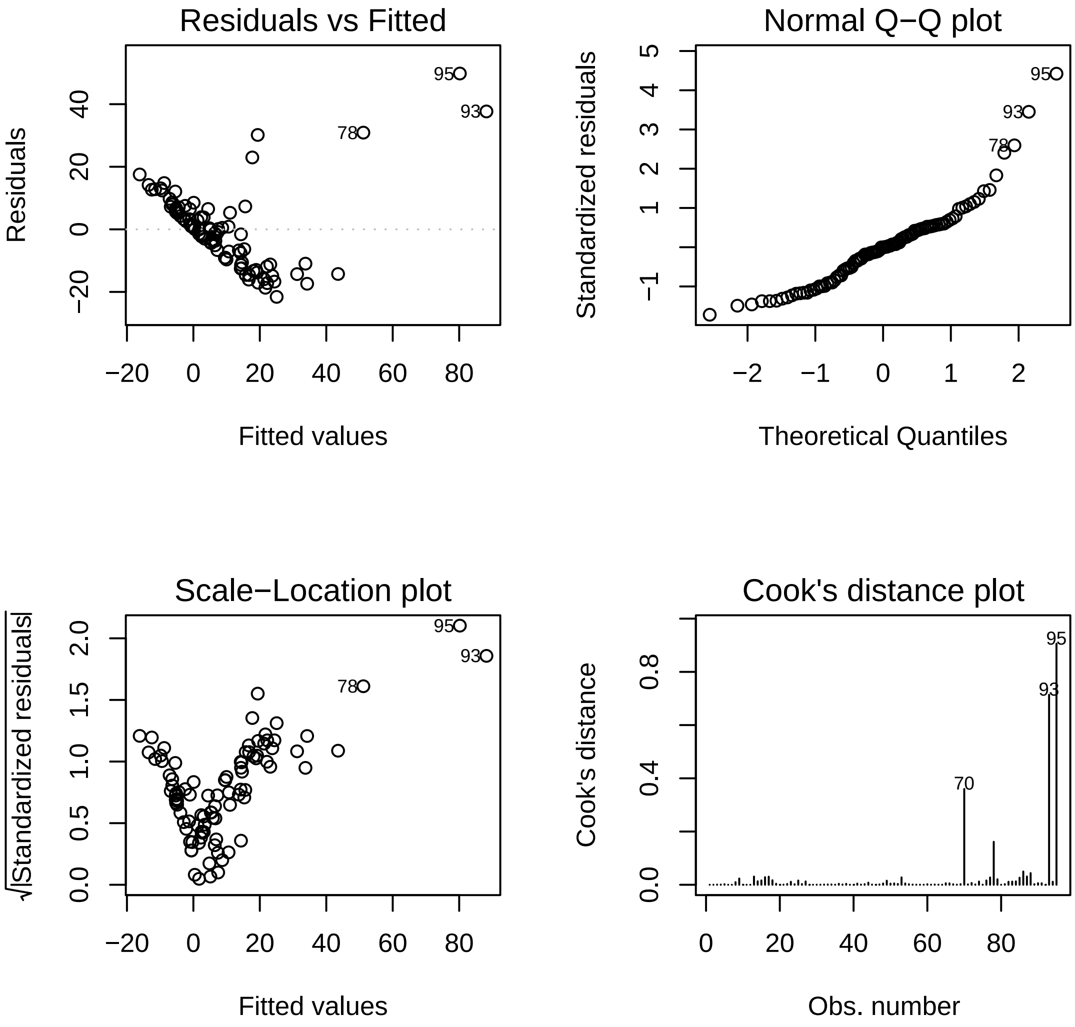

Figure 6.4 Residual plots for Model 1 predicting scale parameter, Structure 2

Table 6.7 Analysis of Variance table for Model 1 predicting scale parameter, Structure 2

	DF	SS	MS	F	Pr(> F)
log.qmd	1	14689	14689	91	3.2e-15
mean.year	1	5254	5254	32	1.6e-07
habPISI	1	222	222	1.4	0.24
habTSHE	1	2731	2731	17	9.0e-05
ntrees	1	3417	3417	21	1.4e-05
Residuals	89	14431	162		

The resulting model had substantially more homoskedastic error terms [3], as is shown in Figure 6.5. However, the quantiles of the errors were still noticeably non-Normal. The relative importance of the predictors was altered. Because the estimate of $\hat{\lambda}$ was close to zero (0.2), the log transform of the response was used:

$$log(c) = \beta_0 + \beta_1 * log(\texttt{qmd}) + \beta_2 * \texttt{habPISI} + \beta_3 * \texttt{ntrees} \tag{6-3}$$

The average year of sampling (`mean.year`), the *Tsuga heterophylla* habitat type (`habTSHE`) and the *Picea sitchensis* habitat type were no longer significant. Although `ntrees` was significant, this variable depends very strongly on plot size, which is inconsistent throughout the data. Because the significance of the term was most likely an artifact of the data and not the biological process, and because the estimated coefficient for this term was very small (less than 0.01), the term was not included.

Table 6.8 Analysis of Variance table for Model 2 predicting scale parameter, Structure 2

	DF	SS	MS	F	Pr(> F)
log.qmd	1	85	85	63	4.1e-12
Residuals	93	125	1.3		

The RMSE for this model – calculated by taking the exponent of the predicted values – was slightly lower than that for the model without the transformed response, and the model is preferable due to the improved residual distribution. The RMSE for this model was 15.47, as compared with 12.73 (on 89 degrees of freedom) for the non-transformed response model. Model coefficients and their standard errors are shown in Table 6.9.

Table 6.9 Coefficients for Model 2 predicting scale parameter, Structure 2

	Estimate	SE	t	Pr(>|t|)
(Intercept)	6.9	0.75	9.2	9.9e-15
log.qmd	-1.6	0.2	-8	4.1e-12

[3] Plotted against fitted values.

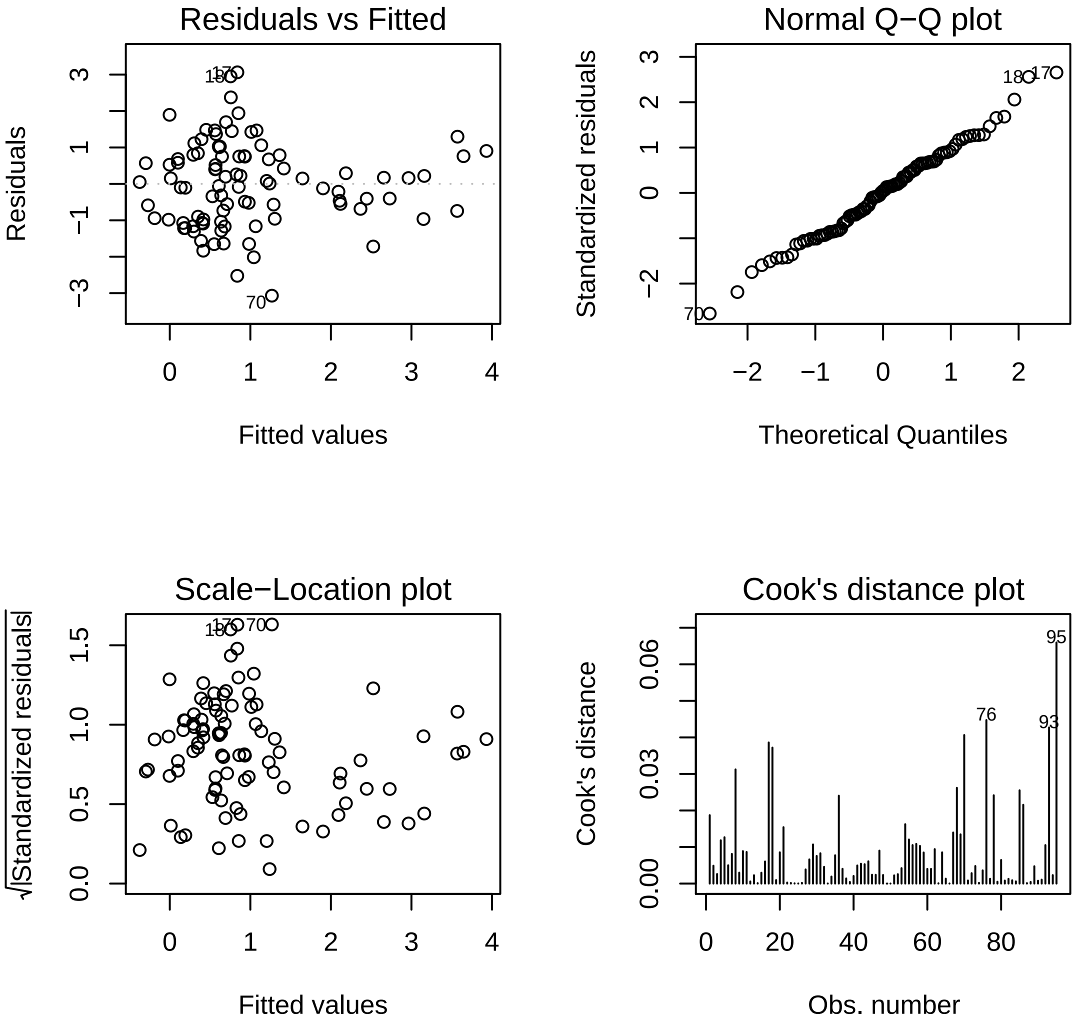

Figure 6.5 Residual plots for Model 2 predicting scale parameter, Structure 2

7.0 Structure 3

7.1 Description and Assumptions

In the analysis discussed in this chapter, changes to the fitting data *structure* and *model form* were made. The rules by which trees were determined to be regeneration were changed and the first stage of the prediction model was treated in a mixed-effects modelling paradigm.

7.1.1 Change in data structure: a newly defined and standardized response

In this structure, it was assumed that the grouping of plots of variable size and remeasurement time as independent samples violated fundamental axioms of probability and, stemming from this, a basic assumption of linear regression. A plot 1 hectare in size is more likely to have regeneration than a 0.10 hectare plot, and a plot visited 20 years after the previous sampling has a higher probability of regeneration occurring than a plot visited only 3 years later.

The data were again arranged such that the total number of trees defined as forest regeneration at the subsequent measurement (variable name num.regen) were aligned horizontally (in rows) with potential predictors measured at the previous plot visit. Only one predictor, the number of years between sampling visits (variable time), contained information from the subsequent visit. Observed covariates for the final year of measurement of a given plot were not included in analysis, because these *years* had no corresponding response variable. (Future regeneration not having occurred yet).

As an alternative to grouping plots of variable size and remeasurement time as independent samples, it was assumed that standardizing the predicted future regeneration, denoted R, for all plots would reduce the likely correlations that existed between adjacent responses (y_i). To account for the variations in plot size and remeasurement time, the mean annual count of future annual regeneration per hectare for an individual visit to a given plot was modelled:

$$R = \frac{10}{na} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \qquad (7\text{-}1)$$

where R is the annual ingrowth per hectare for a plot visit, a is the plot area in hectares, m is the total number of ingrowth trees observed, n is the number of years between plot samplings, and y_{ij} is the jth regeneration tree counted in the ith year.

The 10 year interval was chosen over annual predicted regeneration as a programming convenience because the Forest Vegetation Simulator functions primarily in 10 and 20 year loops.

In order to extract the maximum possible information from the large tree file, a program was written to extract all trees greater than or equal to 5.0 cm dbh, regardless of how large they were when they were *first* recorded.

This is different from simply excluding all diameters below 5.0 cm (as in Structure 1) because some proportion of these trees surpass the 5.0 cm threshold at a later time that is contained in the overall sampling period for the plot recorded in the large tree file. For this reason, simply deleting records based on their *initial* inclusion criterion is wasteful (in the sense that it results in a loss of useable information).

For this reason, the initial definition of regeneration (variable `status` = 2) was ignored, and a new set of characteristics by which to define regeneration was constructed. Trees considered regeneration had all of the following characteristics:

- Trees not measured in the first year of sampling for any plot.
- Trees whose dbh was greater than or equal to 5.0 cm.
- Trees whose dbh at the previous plot visit was less than 5.0 cm.

7.1.2 Change in Model Form: a Mixed-Effects Logistic Regression

Although the preceding definition of the response variable is a large step in the direction of standardization, defining regeneration counts in this way does not resolve the tendency for multiple measurements made on a given plot to be *more* similar than measurements made on separate plots. Similarly, measurements on plots made within a given stand or study tend to have more in common than measurements made on plots from separate stands or studies.

For this reason, treatment of the sampling unit - the variable `plot` - as a random effect, seemed desirable. Pinheiro and Bates (2000), Shabenberger and Pierce (2002), Robinson and Ek (2000) and others have suggested mixed effects models for situations in which there is a necessary hierarchical structure to the data.

7.1.3 Model Description

For the first stage model predicting the probability of stocking, a mixed effects logistic regression model was fit using penalized quasi-likelihood (PQL). PQL is an approximate inferential method used in generalized linear mixed models (GLMM) to simplify what are often very intensive computational requirements. PQL assumes a Normally distributed random effect in the linear predictors. The algorithm used to construct PQL estimates is similar to that used for iteratively re-weighted least-squares (IWLS), and was conceived of as an *approximation* to true maximum likelihood estimation through numerical integration. (Breslow, 2003; Venables and Ripley, 1994).

The response variable was indicator of any future annual future regeneration per hectare for a given plot (`mean.regen.acre.year`). The variable `plot` was treated as a random effect, and a separate intercept was estimated for the logistic regression lines corresponding to each observed level of this predictor. The hierarchical structure of the data was assumed to have 3 levels of nesting, with `stand` nested in `studyID` and `studyID` nested within `dataCode`. An autocorrelation model of the same nesting structure was added. The decision to include this model component was based on the following criteria:

- An initial conceptual justification for the possibility that autocorrelated error terms still presented a potential violation of (all fixed effects) model assumptions.
- A significant test of the ratios of the likelihoods of the data given equations excluding and including (respectively) the autocorrelation model.

7.2 Probability of Stocking

The most important predictor in the first stage model was the log of quadratic mean diameter (`log.qmd`). Remeasurement time (`time`), trees per hectare `tph` and estimated minimum diameter threshold (`minEst`) were initially included in the model but were later removed either due to the statistical or practical insignificance of their effects. The estimated intercept and `log.qmd` coefficient are shown in Table 7.1.

Table 7.1 Fixed-effects coefficients for logistic model predicting probability of stocking, Structure 3

Value	SE	DF	t	Pr....t.
5.9	0.61	4881	9.7	< 2.22e-16
-1.9	0.15	4881	-13	< 2.22e-16

A count histogram of plots having at least one regeneration tree per hectare before the subsequent plot remeasurement is shown in Figure 7.1.

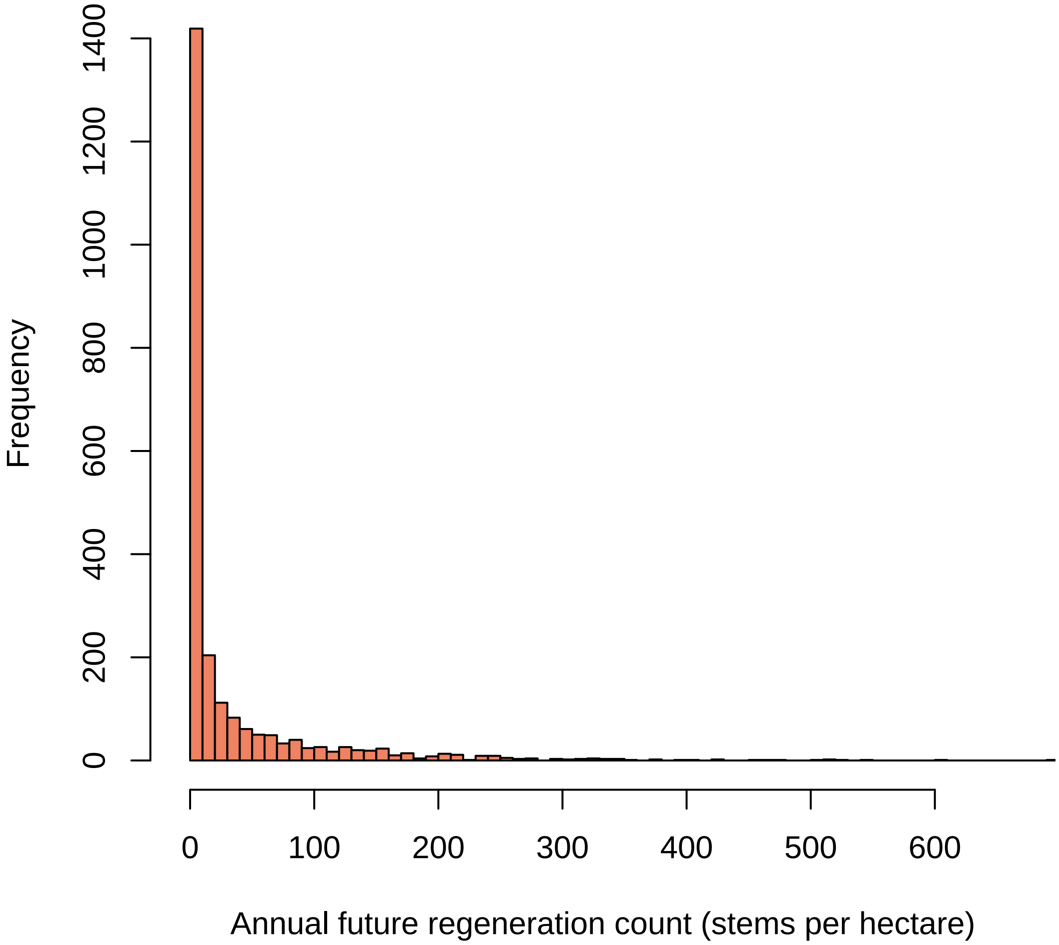

Figure 7.1 Count histogram of mean annual regeneration per hectare (stocked plots), Structure 3

7.3 Conditional Distribution of Stocked Plots

In the second stage, the conditional distribution of stocked plots were again modelled under the distributional assumption of a two-parameter Weibull. Plots with 0 counts of regeneration were excluded from analysis. Weibull scale and shape parameters were estimated for each of 95 stands (variable `stand`) using maximum likelihood estimation, as were the means of all possible observed predictors.

7.3.1 Shape parameter

Important predictors of the Weibull shape parameter in Structure 3 were the log of quadratic mean diameter (`qmd`) and the *Picea sitchensis* habitat type (`habPISI`). A Box-Cox transformation of the response was used, with $\hat{\lambda} = -0.64$. The final model contained an intercept and coefficients for main effects only (no interaction term). This model is shown in Equation 7-2.

$$\hat{b}^{\lambda} = \beta_0 + \beta_1 \texttt{qmd} + \beta_2 \texttt{habPISI} \tag{7-2}$$

The coefficients for this model are shown in Table 7.2.

Table 7.2 Coefficients for model predicting shape parameter, Structure 3

	Estimate	SE	t	Pr(>|t|)
(Intercept)	1.6	0.13	13	< 2e-16
log.qmd	-0.24	0.035	-6.8	8.2e-10
habPISI	0.29	0.05	5.9	5.7e-08

Residual plots for this model are shown in Figure 7.2, The assumption of constant variance of the error terms is not well justified. There are several points on the range $x = [0.5, \ldots, 0.9]$ that are noticeably closer to the fit line than exist in the higher range of $x = [0.9, \ldots, 1.1]$ (a fan shape). The Normal QQ-plot shows that the error terms are heavier-tailed than would be expected were they normally distributed.

The relative importance of predictors in explaining the variation in `B` is shown in the ANOVA in Table 7.3.

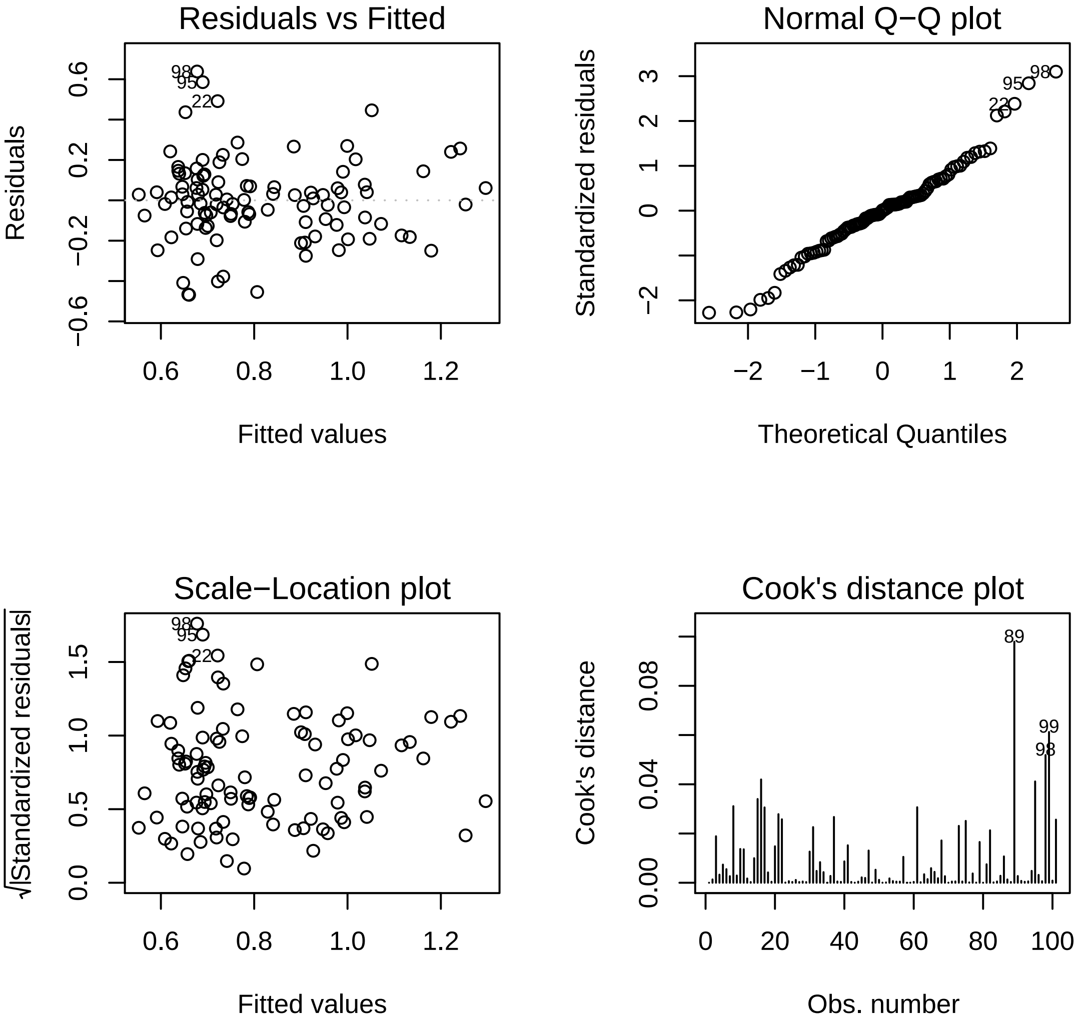

Figure 7.2 Residual plots for model predicting shape parameter, Structure 3

Table 7.3 Analysis of Variance table for model predicting shape parameter, Structure 3

	DF	SS	MS	F	Pr(> F)
log.qmd	1	1.6	1.6	38	1.8e-08
habPISI	1	1.5	1.5	35	5.7e-08
Residuals	98	4.2	0.043		

7.3.2 Scale parameter

Several models predicting the scale parameter were evaluated. Initially, `C` was predicted directly, and a fairly complex model with 6 parameters (intercept, and coefficients for `qmd`, `mean.year`, `ntrees`, `origin`, and `qmd:totbamha`) was constructed. This model is shown in Equation 7-3.

$$\hat{c} = \beta_0 + \beta_1\texttt{qmd} + \beta_2\texttt{mean.year} + \beta_3\texttt{ntrees} + \beta_4\texttt{origin} + \beta_5\texttt{qmd*totbamha} \tag{7-3}$$

Although all terms in this model were significant at least the 0.05 level, residual plots for the model were very poor. Moreover, plots of individual predictors against `C` suggested that the basic assumption of a linear relationship between predictor and response was ill-founded.

In order to normalize and reduce heteroscedasticity of the residuals, a Box-Cox transformation of the response variable was implemented. The straightness of the Normal QQ-plot was improved noticeably, and the plot of residuals against fitted values was acceptably homoskedastic. However, after transforming the response variable, the relative importance of several terms (`totbamha`, `mean.year`, `origin`, `qmd:totbamha`) in explaining the variation in `C` became negligibly small. In addition, the optimal value of the Box-Cox transformation, $\hat{\lambda}$, minimizing the negative log-likelihood of y was extremely close to zero ($\hat{\lambda} = -0.1$). For these reasons, the simple log transformation of y was considered to be sufficient.

The final, simplified model chosen is shown in Equation 7-4. Important predictors of `C` in the final model chosen for Structure 3 were the log of quadratic mean diameter (`qmd`) and the *Picea sitchensis* habitat type (`habPISI`). The model form was as follows:

$$\log(\hat{c}) = \beta_0 + \beta_1\texttt{qmd} + \beta_2\texttt{habPISI} \tag{7-4}$$

This model had an RMSE[1] of 107.32. Residual plots for the final model are shown in Figure 7.3.

The coefficients for this model predicting the Weibull scale parameter are shown in Table 7.4.

The relative importance of predictors in explaining the variation in `C` is shown in the ANOVA in Table 7.5.

[1] Calculated after untransforming (taking the exponent) of the predicted values.

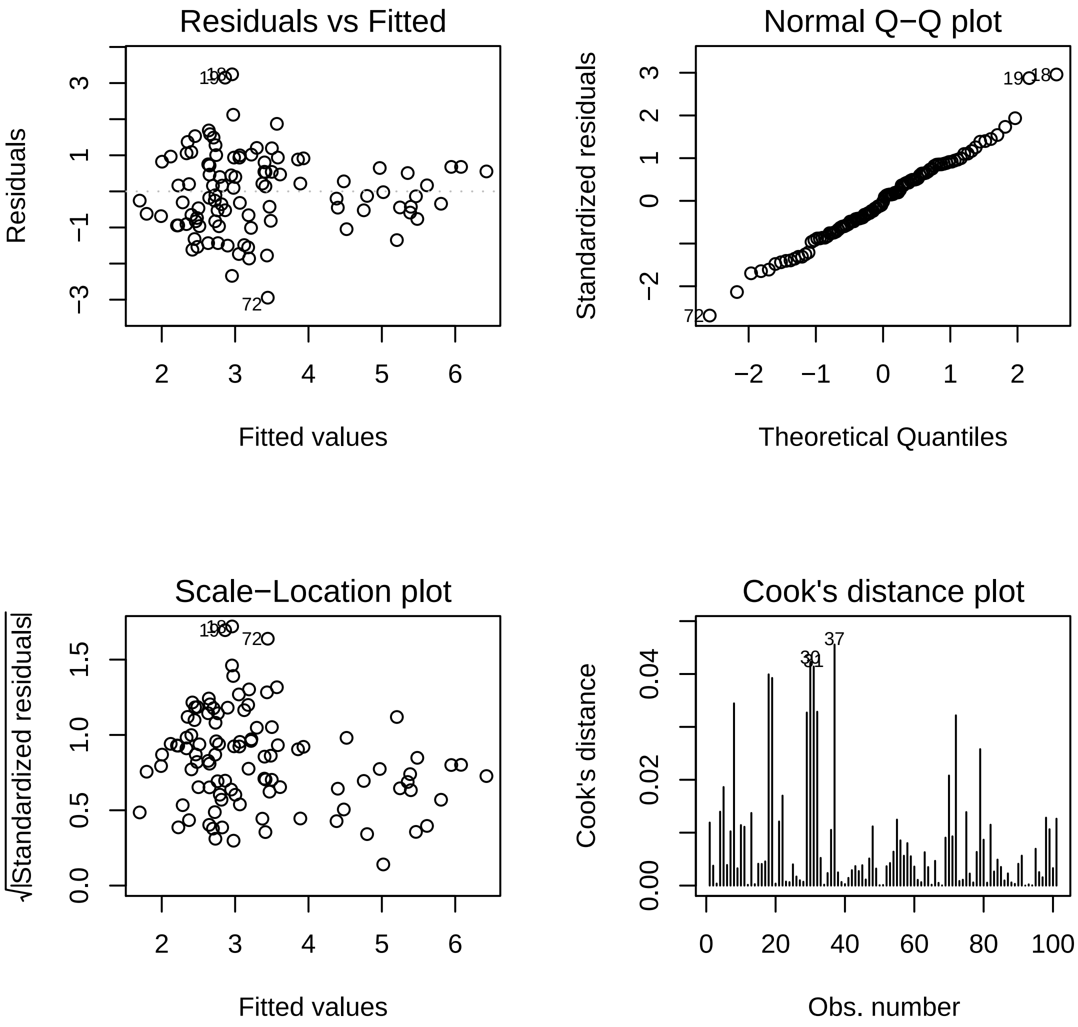

Figure 7.3 Residual plots for model predicting scale parameter, Structure 3

Table 7.4 Coefficients for model predicting scale parameter, Structure 3

	Estimate	SE	t	Pr(>|t|)
(Intercept)	9.8	0.69	14	< 2e-16
log.qmd	-1.8	0.19	-9.7	6.9e-16
habPISI	0.7	0.26	2.7	0.0091

Table 7.5 Analysis of Variance table for model predicting scale parameter, Structure 3

	DF	SS	MS	F	Pr(> F)
log.qmd	1	108	108	88	2.4e-15
habPISI	1	8.6	8.6	7.1	0.0091
Residuals	98	119	1.2		

7.4 Anova Model

In order to determine whether any of the relatively complex models constructed in this and the preceding two chapters had noticeably improved predictive power over a simple, tabular, mean response model, a simple analysis of variance (ANOVA) regeneration model was considered. In this framework, all predictors were nominal (factor) variables, and the usual linear regression assumption restrictions were relaxed. Rather, the response was assumed to be the simple arithmetic mean for a given factor - or interaction of factors - level.

Continuous predictors were made into class variables by establishing 10 divisions based on observed deciles. In the case of quadratic mean diameter (qmd), for example, the observed deciles are shown in Table 7.6.

Table 7.6 Observed deciles used to establish quadratic mean diameter classes

	Decile	Value
1	1	7.98991
2	2	13.61509
3	3	22.77030
4	4	34.37591
5	5	42.90312
6	6	48.78177
7	7	54.77401
8	8	59.82723
9	9	67.16616
10	10	113.11009

Any observations with values of quadratic mean diameter less than the first decile were in qmd Class 1, any values between the first and second decile were in Class 2, and so forth. The same procedure was used with trees per hectare, total basal area, total number of stems, and total number of mortality in the preceding measurement interval. Because they have consistently been the most important predictors, habitat type and quadratic mean diameter class were included in all models. Then, each of the other constructed class variables was included, as was its interaction. Two other factors, origin (plant or natural), and coastal proximity[2] were also included.

[2] A definition of this variable is in the appendix.

The best model, based on a relevant reduction in root mean-squared error, is the simple ANOVA including habitat type and quadratic mean diameter class, with no interaction term:

$$R = \beta_0 + \beta_1 hab + \beta_2 qmdClass \tag{7-5}$$

Each of the four habitat types and each of ten quadratic mean diameter classes were highly significant with p-values less than 2e-16.

Quadratic mean diameter was substantially more important in explaining the variation in the observed mean regeneration per hectare. The relative explanatory power of the two are shown in Table 7.7.

Table 7.7 Analysis of Variance model, Structure 3

	DF	SS	MS	F	Pr(> F)
hab	3	377573	125858	105	<2e-16
qmdClass	9	5896491	655166	545	<2e-16
Residuals	4989	5999732	1203		

The lower 5 qmd classes explain the majority of model variation, suggesting that a model with quadratic mean diameter divided into less classes would result in similar predictions.

One positive characteristic of the ANOVA model as compared with the regression-based approaches discussed in earlier sections, is that the assumption of a linear relationship between predictor and response is relaxed. The cell means for the observed levels of X and Y are of interest, and not the relationship *between* these levels.

PART III: SIMULTANEOUS FITTING

In recent years, the popularity of Zero-Inflated Poisson regression has increased as an alternative to the two-stage approach to modelling count data that was utilized in the previous chapters.

7.5 Finite Mixture Regression

The zero-inflated model is a special case of what are more generally called Finite Mixture Regression (FMR) count models. This family assumes that a random variable, y_i, was drawn from a mixture of C populations. Each population exists in some proportion, π_j, in a *super*population. Cameron and Trivedi (1998) explain that the superpopulation is "an additive mixture of C distinct populations in proportions $\pi_1, \ldots, \pi_C$," where $\sum_{j=1}^{C} \pi_j = 1$ and $\pi_j \geq 0$. They provide the following equation for the mixture density:

$$f(y_i \mid \Theta) = \sum_{j=1}^{C-1} \pi_j f_j(y_i \mid \theta_j) + \pi_C f_C(y_i \mid \theta_C) \tag{7-6}$$

The first term on the right-hand side of Equation 7-6 is the individual mixing probability, π_j times the density of the component population, $f_j(y_i \mid \theta_j)$ (Cameron and Trivedi, 1998). The mixing probability in the second term, π_C, is equal to $1 - \pi_j$ and $f_C(y_i \mid \theta_C)$ is the alternative density function. Because the π_j's (and π_C's) are unknown, these must be estimated along with other unknown parameters. It is conventional to parameterize the model via the logit function, such that $\pi_j = \exp(\lambda_j)/(1 + \exp(\lambda_j))$. λ_j is then modelled as a function of observed predictors (Cameron and Trivedi, 1998).

Although it is possible to assume component populations with different distributions, it is most common to assume they are the same.

7.6 Zero-Inflation

Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models are a special case of Finite Mixture Regression, in which two component densities are *not* assumed to be the same. Rather, some of the component populations are presumed to come from a *point mass* or *degenerate* distribution defined only at zero.

Lambert (1992) described the Zero-Inflated Poisson distribution, which assumes a mixture of the Poisson and zero categories of random variables. Because the steps involved in deriving the ZIP and ZINB likelihood functions are quite similar, only the ZIP will be considered in the following explanation. In the simplest ZIP parameterization, the mixture and Poisson parameters depend on different covariate vectors. The random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ have the properties that

$$\begin{aligned} Y_i &\sim 0 && \text{with probability } p_i \\ Y_i &\sim \text{Poisson}(\lambda_i) && \text{with probability } 1 - p_i \end{aligned}$$

and

$$\begin{aligned} Y_i &= 0 && \text{with probability } p_i + (1-p_i)e^{-\lambda_i} && \text{(7-7)} \\ Y_i &= x && \text{with probability } (1-p_i)e^{-\lambda_i}\lambda_i^x/x!, && \text{(7-8)} \end{aligned}$$

for $x = 1, \ldots, n$.

The mixture (binomial) parameter, $\mathbf{p} = p_1, \ldots, p_n$, and the Poisson parameter, $\lambda = \lambda_1, \ldots, \lambda_n$ are not modelled directly. Rather, the *logit* and *log* links of the parameters (respectively) are employed. These are the canonical link functions for binomial and Poisson generalized linear models (GLM's) (Lindsey, 1997). The parameters are related to distinct covariate matrices **B** and **G** and their corresponding coefficient vectors, β and γ in the following way:

$$\begin{aligned} \text{logit}(\mathbf{p}) &= \log\left(\frac{\mathbf{p}}{1-\mathbf{p}}\right) = \mathbf{G}\gamma \\ \log(\lambda) &= \mathbf{B}\beta \end{aligned}$$

The likelihood function for the ZIP model is constructed from equations 7-7 and 7-8. First, the link functions are solved for $\mathbf{p}$ and λ:

$$\begin{aligned} \log\left(\frac{\mathbf{p}}{1-\mathbf{p}}\right) &= \mathbf{G}\gamma \\ \mathbf{p}/(1-\mathbf{p}) &= e^{\mathbf{G}\gamma} \\ \mathbf{p} &= e^{\mathbf{G}\gamma} - \mathbf{p}e^{\mathbf{G}\gamma} \\ \mathbf{p} + \mathbf{p}e^{\mathbf{G}\gamma} &= e^{\mathbf{G}\gamma} \\ \mathbf{p} &= \frac{e^{\mathbf{G}\gamma}}{1 + \mathbf{e}^{\mathbf{G}\gamma}} \end{aligned}$$

and

$$\begin{aligned}\log(\lambda) &= \mathbf{B}\beta \\ \lambda &= e^{\mathbf{B}\beta}\end{aligned}$$

Then, substituting these into the combined density function, we get the likelihood (Lambert, 1992; Qin et al., 2003):

$$\begin{aligned}
l(\beta,\gamma;y) &= \prod_{y_i=0}\{p_i + (1+p_i)\exp(-\lambda_i)\}\prod_{y_i>0}\{(1-p_i)\exp(-\lambda_i)\lambda_i^{y_i}/y_i!\} \\
&= \prod_{y_i=0}\{\frac{\exp(\mathbf{G_i}\gamma)}{1+\exp(\mathbf{G_i}\gamma)} + \frac{1}{1+\exp(\mathbf{G_i}\gamma)}\exp(-\exp(\mathbf{B_i}\beta))\} \\
&\quad \prod_{y_i>0}\{\frac{1}{1+\exp(\mathbf{G_i}\gamma)}\exp(-\exp(\mathbf{B_i}\beta))[\exp(\mathbf{B_i}\beta)]^{y_i}/y_i!\} && \text{(7-9)} \\
&= \prod_{y_i=0}\frac{1}{1+\exp(\mathbf{G_i}\gamma)}\{\exp(\mathbf{G}\gamma) + \exp(-\exp(\mathbf{B_i}\beta))\} \\
&\quad \prod_{y_i>0}\frac{1}{1+\exp(\mathbf{G_i}\gamma)}\{\exp(-\exp(\mathbf{B_i}\beta))[\exp(\mathbf{B_i}\beta)]^{y_i}/y_i!\} && \text{(7-10)} \\
&= \{\prod_{i=1}^{n}\frac{1}{1+\exp(\mathbf{G_i}\gamma)}\}\prod_{y_i=0}\{\exp(\mathbf{G_i}\gamma) + \exp(-\exp(\mathbf{B_i}\beta))\} \\
&\quad \prod_{y_i>0}\{\exp(-\exp(\mathbf{B_i}\beta))[\exp(\mathbf{B_i}\beta)]^{y_i}/y_i!\} && \text{(7-11)}
\end{aligned}$$

The log of this function simplifies to the log-likelihood shown in Equation 7-12.

$$\begin{aligned}
L((\beta,\gamma;y) = \sum_{y_i=0}\log(\exp(\mathbf{G_i}\gamma)) + \exp(-\exp(\mathbf{B_i}\beta)) + \sum_{y_i>0}(y_i\mathbf{B_i}\beta - \exp(\mathbf{B_i}\beta)) \\
-\sum_{i=1}^{n}\log(1+\exp(\mathbf{G_i}\gamma)) - \sum_{y_i>0}\log(y_i!) \qquad \text{(7-12)}
\end{aligned}$$

7.7 Zero-Inflated Models in the Literature

There are many examples of Zero-Inflated Poisson models in the literature. Their lineage can be traced back to the Finite Mixture modelling literature, of which the earliest reference by most accounts is Pearson (1894). Pearson's treatment of measurements made on crabs as a mixture of two normally distributed random variables (representing two species) was early evidence in support of evolutionary theory.

The ZIP acronym is a catch-all that has been employed to mean different things in different situations. This can become confusing when a fundamental description of the fitting structure is not described early on in a paper, but the term *ZIP* is used as though it could refer to only one model parameterization. This is in part due to the fact that several parameterizations and several *special cases* are available in ZIP regression. For example, a simplified likelihood and slightly different fitting methods are possible when both the mixture parameter (p) and the Poisson parameter (λ) are predicted from *the same* covariate matrices (Hall, 2000; Lambert, 1992). On the other hand, p and λ may depend on entirely different predictors, in which case the likelihood is slightly more complex.

Lambert's parameterization of the ZIP was fit using the EM algorithm (Lambert, 1992). Hall (2000) developed Zero-Inflated Binomial (ZIB) regression models and ZIP and ZIB models incorporating both fixed and random effects in order to model the control of silverleaf whiteflies on poinsettia plants via subirrigated pesticide applications. In the same paper, Hall also suggested a hierarchical structure in which to view Lambert's printed wiring board data, such that mixed-effects were appropriate for analysis. He demonstrated that his mixed-effects ZIP model provided a better fit than Lambert's all fixed-effects approach.

Cheung (2002) wrote an elegant paper on the utility of one of Lambert's original model constructions in the analysis of an infant growth and development study. The count response was the number of blocks in block towers built by toddlers, and the mixture parameter represented the probability of stacking a critical number of blocks used as a traditional test of developed abilities.

The models described by Hall were fit in a fashion very similar to those developed by Lambert, using the EM algorithm. Yau and Lee (2001), however, suggested an alternative parameterization of the mixed-effects ZIP model that could be fit using the Newton-Raphson algorithm. Their example was a study analyzing the success of Workplace Risk Assessment Teams (WRATS) in reducing injuries among hospital workers.

In describing attempts at modelling the abundance of Leadbetter's Possum, Welsh et al. (1996) provided an excellent summary of the (often confusing) differences between some of the standard *conditional* and *mixture* approaches to modelling over-dispersed data.

7.8 A Zero-Inflated Negative Binomial Regeneration Model

Several models involving either the simultaneously fit ZIP or ZINB were examined. These were fit to the complete mean-response greater than 5.0 cm regeneration data defined and described as Structure 3. In all cases, the ZINB models were better. The non-linear algorithms used to estimate mixture and location parameters for these failed to converge for the majority of the ZIP variants.

Standardized regeneration rates were *rounded* in order to fit these discrete models.

Convergence was successful for several of the ZINB models, and the quality of the fit of one of the simplest of these was quite good. In order to clarify description of this model, I will refer back to the earlier section (9.2) of this chapter titled *Zero-Inflation.* In that section, a parameterization of the ZIP model was described in which the Poisson parameter, λ, and mixture parameter, $\mathbf{p}$, depend on unique vectors of covariates. However, a few other possible parameterizations of the model exist as well. It is possible, for example, for both $\mathbf{p}$ and λ to depend on *the same* vector of covariates. It is also possible for either λ *or* $\mathbf{p}$ to depend on a vector of covariates, and for the other parameter to be estimated empirically as an unknown.

The model discussed in the following paragraph is of the latter form; only the location parameter varies as a function of covariates. Moreover, where there were only two parameters in the ZIP model previously discussed (λ and $\mathbf{p}$), there is one additional parameter in the ZINB. This is because the Poisson parameter, λ, represents - by definition - both the mean and variance of the random variable. The Negative Binomial distribution has *two* parameters, one estimating the location of the center of the data, and the other the shape (or variance).

This simple ZINB regression model with covariates predicting the location parameter (but *not* the mixture parameter) as a linear function of the log of quadratic mean diameter and habitat type had an AIC of 9994.4. This was a very minor improvement over the ZINB with the first term (`log.qmd`) included (AIC=10249.59)

Several more complex models with separate covariate vectors predicting the mixture parameter were attempted, as well as models with *both* the mixture and location parameter estimated as a function of the same covariates. Some of these converged, though most did not. For those that did, the estimated standard error of the mixture parameter was very large. The parameter estimates (and their standard errors) for the simple model with `log.qmd` as the sole predictor are shown in Table 7.8.

Some simple relationships in a correlation matrix of the parameter estimates from the ZINB model are quite informative. It is reassuring, for example, that the same strongly negative relationship between the log of quadratic mean diameter and mean annual regeneration per

Table 7.8 ZINB parameter estimates and standard errors

	Parameter	Estimate	SE
1	Location	7.12	0.1090
2	log.qmd	−1.61	0.0391
3	Mixture	0.73	0.1860
4	Shape	−1.09	0.1100

hectare that was observed in the two-stage models is evident in a simultaneously-fit version. It appears in the matrix shown in Table 7.9 that the shape and mixture parameters *themselves* are very highly correlated.

Table 7.9 ZINB correlation matrix

	Location	log.qmd	Mixture	Shape
Location	1	-0.85	0.15	-0.13
log.qmd	-0.85	1	-0.58	0.55
Mixture	0.15	-0.58	1	-0.94
Shape	-0.13	0.55	-0.94	1

7.8.1 Problems and discussion

Although the model predicts *better*, in some respects, than several of the two-stage attempts, there are fundamental assumption violations preventing acceptance of it on this basis. Most importantly among these is the violation of independence of the error terms. Variation in these data is highly dependent on grouped structures (e.g. `dataCode`, `studyID`, etc), as was discussed in the justification for Structure 3. Although the ZINB is attractive and elegant in its simplicity, the same strong theoretical basis for the incorporation of both fixed and random effects exists in the latter two-stage models also exists for their simultaneously-fit counterparts.

As discussed briefly in the literature review at the beginning of this section, Yau and Lee (2001) and Hall (2000) have formulated Zero-Inflated model variants incorporating a random intercept[3] estimated within groups. Although the potential for this model to be the optimal solution to this modelling project exists, it is above and beyond the current abilities of this author to derive and program that model.

[3] But not, to my knowledge, a random slope.

One potential reason for the failure of the models with covariates predicting the mixture parameter is the apparently high correlation between the mixture and shape parameters. Lindsey (In Pers. Comm) has stated that the ZIP and ZINB likelihood surfaces become quite flat when redundant parameters are included, and convergence of the optimization routing therefore becomes difficult. Given that the shape and mixture parameters appear highly correlated, using overlapping covariates as separate predictors of these would assumably result in redundant over-parameterization.

It should be noted that literature pertaining to the appropriate residual diagnostics to be used in ZIP- and ZINB-regression is virtually non-existent (Lindsey, Pers. Comm.).

8.0 Discussion

Three general data Structures, 2 approaches to fitting bimodal models, 3 possible response variables and a suite of *families* of regression models (linear, linear mixed-effects, non-linear, generalized linear, generalized linear mixed effects, generalized non-linear) have been utilized for the basic analyses comprising the body of work completed for this thesis. I have made a concerted effort to be careful about letting the relative importance of any predictor under one combination result in the inclusion of that predictor in another combination. I have tried, rather, to use what I have learned from the earlier modelling efforts, for example the simple logistic regressions predicting the probability of stocking for Structures 1 and 2, merely as reconnaissance data guiding the initial choice of potential terms to be included in later models. In no instance was a term included in any model based solely on its relative importance in another[1].

Both in models discussed in this thesis and in additional attempts that were not formally presented here, no predictor was consistently as important in explaining the variation in regeneration counts (by whichever definition) as quadratic mean diameter (`qmd`). In Part 1, the relationship between `qmd` and the probability of stocking was negative, as was the relationship between `qmd` and both the shape and scale parameters in second-stage models. In Part 2, `qmd` was an important predictor of the mixture parameter in the ZINB model evaluated, and the same negative relationship was observed.

This negative relationship is in contrast to the results of Schweiger and Sterba (1997), who found that, with all species grouped together, "An increasing quadratic mean diameter distinctly increases the probability of regeneration occurrence." (Schweiger and Sterba, 1997, p. 113) in primarily *Picea abies* stands in Austria. In the same paper, however, the authors presented a plot of the probability of *Picea* regeneration against quadratic mean diameter for 3 levels of crown competition factor (ccf) (Schweiger and Sterba, 1997, p. 114) which shows the probability of stocking reaching a maxima at approximately QMD = 46 cm. After this point, the probability of *Picea* stocking begins to decrease with qmd.

Because the distribution of qmd is spread further into the higher quantiles in the fitting data for the analysis presented in this thesis (due to the age of the LSOG coastal stands), it may be that the negative relationship observed here is the weighted sum of the effects, i.e. above and below, some unknown apex, and the former is taking precedent. Although in this analysis

[1] Although, of course, there are very good arguments for including terms in models based on biological or theoretical importance regardless of their performance in significance tests

I have essentially assumed (negative) linearity between the probability of stocking and qmd, it is likely that the relationship varies locally.

Conceptually, it is reasonable to assume that there is some level of qmd below which regeneration counts increase with stand density, and beyond which competition among mature trees reduce light, moisture and nutrient availability for ingrowth[2].

8.1 Further Exploration of an Important Linearity Assumption

An important assumption of linear regression that may often be overlooked is that a *linear* relationship actually exists between X and Y. For this reason, plots of all 4 predictor/response combinations of the relationship between the (standardized) regeneration rates, the log of these rates, quadratic mean diameter, and the log of quadratic mean diameter were examined. These are shown in Figure 8.1, and indicate that the relationship between the log of R (R is the mean future regeneration count per hectare) and the log of `qmd` comes closest to approximating linearity. In further modelling efforts, the log-transformed data (in both the X and Y plane) should be considered in parameter prediction. It is important to note that the log-log model has the functional form shown in Equation 8-1, and that error terms in this structure are therefore assumed to be multiplicative. However, because the model itself is fit in the log-scale, errors are additive on that scale and the Central Limit Theorem justifies the normally distributed residual assumption.

$$y_i = \beta_0 x_i^{\beta_1 e_i} \tag{8-1}$$

The effect of the large number of excess zeroes on the log data relationship is evident in Figure 8.2, which shows fits of the log of `R` regressed on `log.qmd` for both the complete and conditionally stocked (non-zero) response. The complete data regression line is somewhat theoretical, as the log of 0 is $-\infty$. As a weak approximation, the log of (R + .0001) was substituted as the response.

After quadratic mean diameter, the most important predictor in several modelling attempts was habitat type (`hab`).

The predicted probabilities of stocking varied substantially between habitat types for Structures 1, 2 and 3, as is shown in Figure 8.3. The median estimated probability is highest for the *Picea sitchensis* series and is followed by *Tsuga heterophylla* in the best models for all 3 data structures.

[2] The effects of large disturbances and gap creation not-withstanding

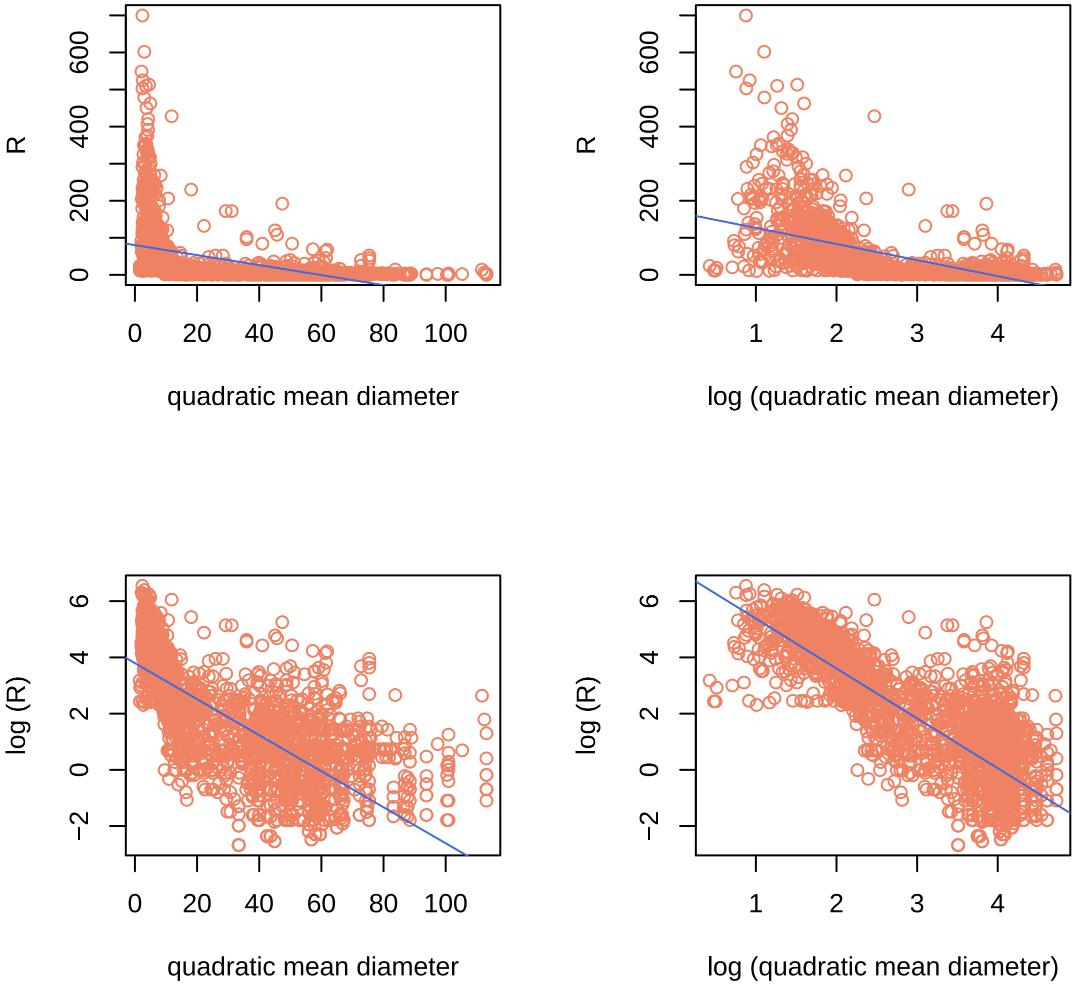

Figure 8.1 Transformations of `R` and `qmd`

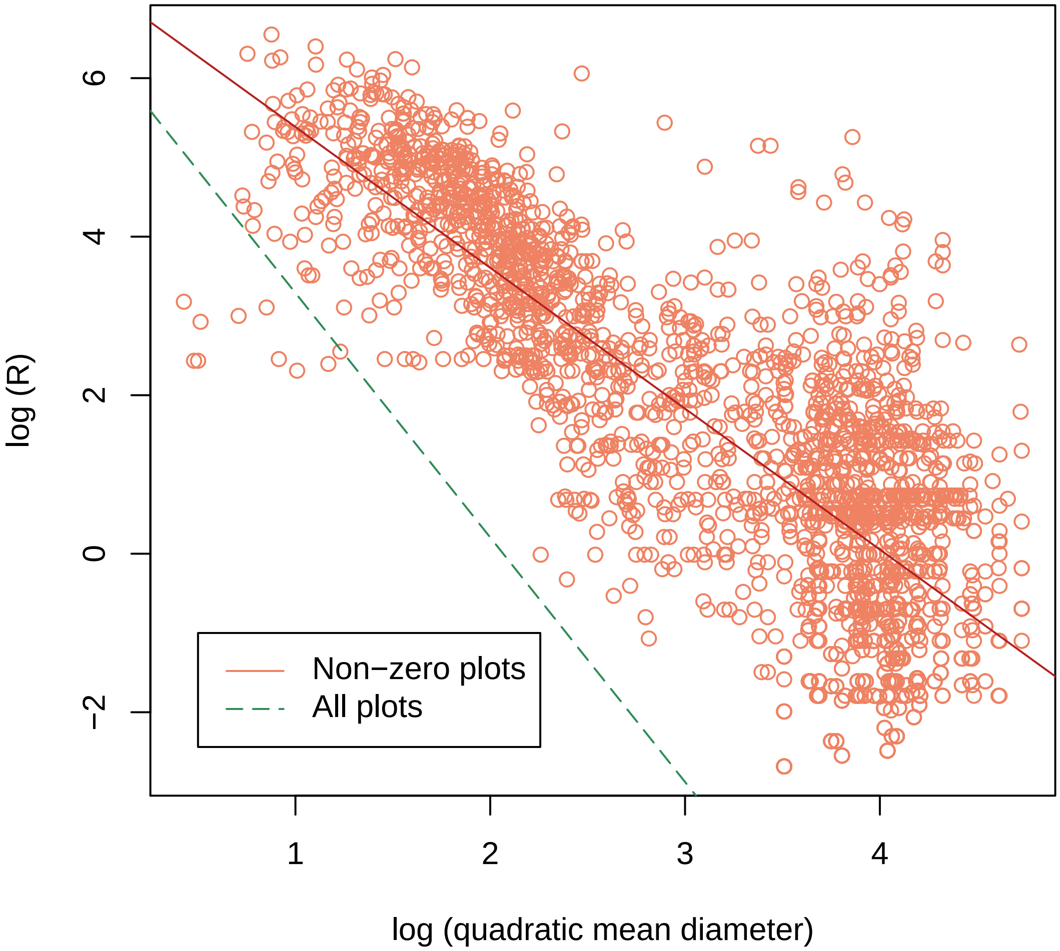

Figure 8.2 Linear relationship of log(R) and log(qmd)

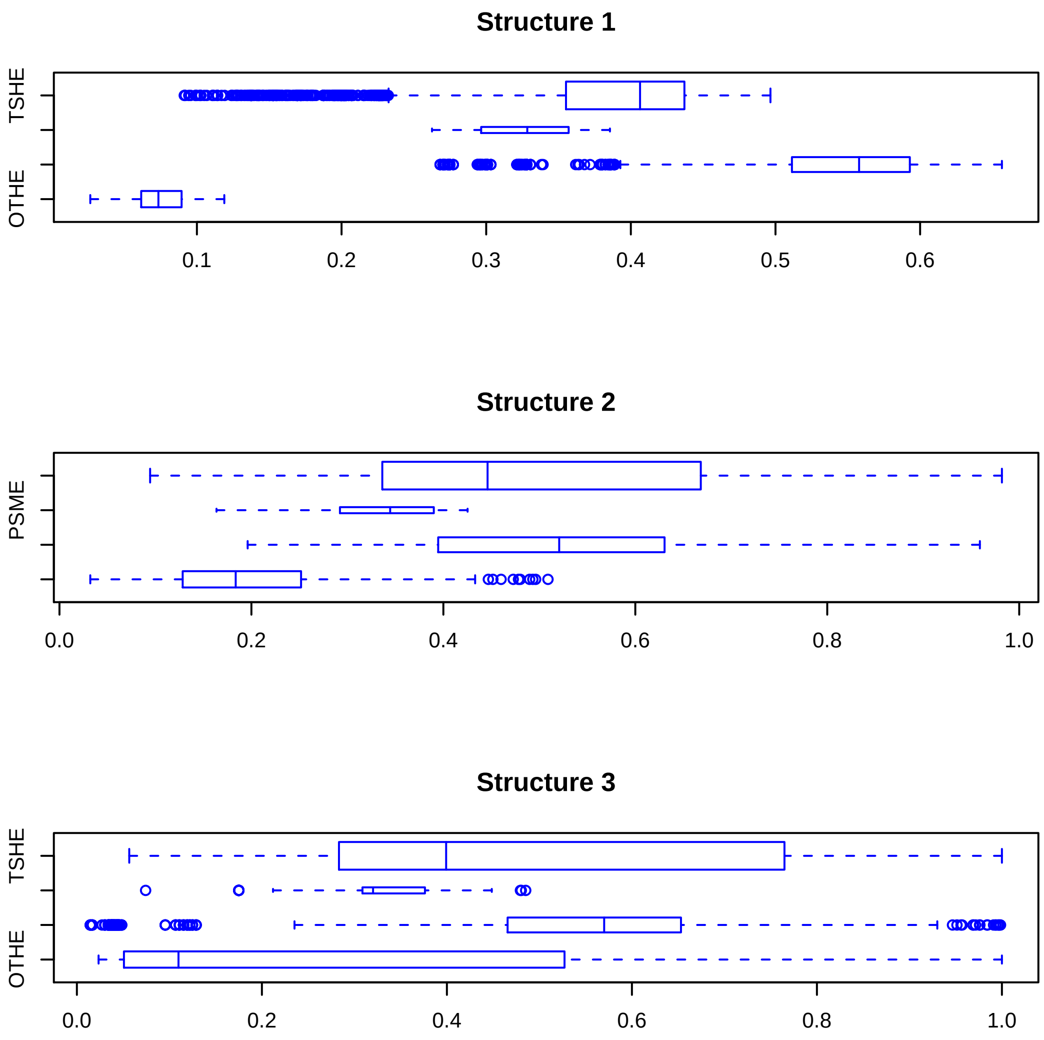

Figure 8.3 Fitted probabilities of stocking by habitat type

8.2 Relationship between tree size and stand density

A good model is one that abstracts the biological process that we are interested in enough to avoid over-fitting, but not so much as to be too general. The balance varies with the process or relationship of interest, with spatial and temporal scale, and as a function of the model's intended use.

Several of the models discussed in this thesis tend toward being overly abstract, due largely in part to excessive attention having been paid to the modelling techniques being employed rather than to the biological processes themselves and to their expected behavior. That said, it appears that at least one allometric relationship has made itself evident: there is a reasonably strong linear relationship between the log of the average tree basal area (QMD) and the log of average future regeneration.

A negative relationship between the size of trees and the size of regeneration counts is intuitive in LSOG stands. However, a model based solely on this relationship does not take into account the potentially inverse effects of density on competition for growing space. For this reason, it is useful to look at the relationship between the size of trees and their relative density in further detail.

The coefficients for the qmd term in the logistic stage of the two-stage models were small, but slightly negative. It is important to note that this corresponds to an *increase* in the log-odds of the probability of stocking. In other words, the probability of *any* regeneration occurring increases with increasing quadratic mean diameter. This component of the two-state system agrees with the model presented by Schweiger and Sterba (1997). It is the the conditional distribution of regeneration stocking that is negatively related to average basal area (qmd).

Due to the age of the majority of stands in the data, the relationship between tree size and density is positively related. This relationship is evident in the plot of total basal area per hectare (m^2/ha) against quadratic mean diameter (cm) for the complete data shown in Figure 8.4. In least-squares regression models utilized for exploratory purposes, the log of qmd explained more variation in regeneration than the log of total basal area, and both were negatively sloped.

Although it was initially believed to be the opposite, the range of diameters represented in the data used for this study is considerably larger than that used in at least the validation of the model presented by Schweiger and Sterba (1997). The ranges of qmd that they report for their three validation stands are $qmd_1 \subseteq [21.8, \ldots, 44.6]$, $qmd_1 \subseteq [29.3, \ldots, 48.7]$, and $qmd_3 \subseteq [8.9, \ldots, 38.1]$. The range of qmd represented in the fitting data used in the present study is $qmd_{PNW} \subseteq [1.53, \ldots, 113.1]$, with a mean of 40.19 and third quartile of 57.83. The proportion

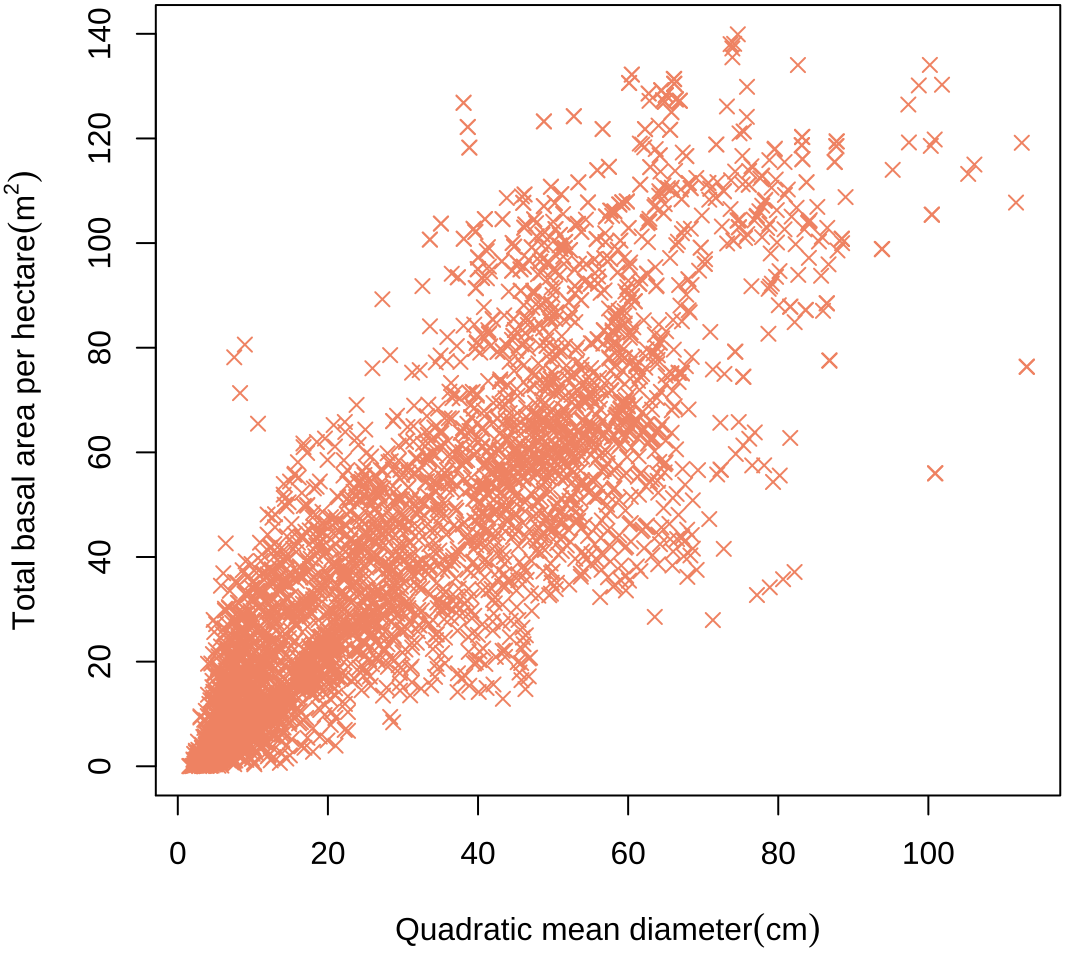

Figure 8.4 Relationship of Total Basal Area Per Hectare and Quadratic Mean Diameter

of the x-axis in the unscaled plot of regeneration against qmd (upper left hand corner) in Figure 8.1 represented by the range of data in the Austrian study is quite small, and it is clear that analysis of only that component would lead to very different conclusions. The authors did not report summary statistics of the covariate as used in fitting their model. Personal communications with the corresponding author (Dr. Huber Sterba) further clarified that the conditional distribution of stocked plots had not been modelled[3].

8.3 Importance of missing values

In order to reconsider the potential importance of abiotic predictors that were excluded early on in the analysis, missing elevation values were obtained from digital elevation maps for all but 12 observations in the fitting data[4]. Elevation and squared elevation were included as covariates in the final two-stage models described in Structure 3. Neither term was significant in the mixed-effects logistic regression model predicting the probability of stocking, nor in the linear regression models predicting the conditional distribution of regeneration established on stocked plots.

Imputation is a term generally used to describe the replacement of missing values in data by one of several means. In *simple* imputation, one may replace missing values with a given (constant) statistic of interest, such as the sample mean, or with nearest-neighbor or moving average algorithms. In more advanced Multiple Imputation (MI) models, Bayesian methods are used to predict missing values from other predictors and from the response variable as well (Harrell, 2001). Then, the uncertainty associated with these predictions is included in the overall error estimated for the original regression model of interest, fit to the observed and imputed complete data.

Harrell (2001) outlined several general guidelines for the use of MI in a variety of missing data scenarios. These were based on situations in which the missing portion of the data under consideration is classified as either *missing completely at random* (MCAR), *missing at random* (MAR), or *informative missing* (IM). MCAR values are missing elements that occur according to random chance. Harrell cites, for example, a missing response in a laboratory experiment “Because of a dropped test tube (if it was not dropped because of knowledge of any measurements).” (Harrell, 2001)

[3] Only the conditional *probability* of any *Picea* regeneration was modelled.

[4] It was not possible to obtain DEM data for these observations because the geographic coordinates (latitude and longitude) were also missing.

MAR data are elements similar to those previously described as MCAR, but the description is qualified, as follows: MAR elements of a given predictor must occur for levels of another predictor - or levels of other predictors - that are *present* in the data (Harrell, 2001).

It may be possible to argue that the remaining missing values in slope and aspect variables are MAR, and thus utilize MI methods in order to include these variables as covariates in predictive models. However, the terms utilized in an MI model used to predict missing values should be a *superset* of the final regression model of interest (Harrell, personal comm.). Harrell's `R` software for MI is not presently equipped to handle interaction terms in which both terms contain missing values. The MI models run into problems when the interaction term - for example, the interaction of the cosine of slope with aspect - with missing values must be predicted from the component variables which have a similarly structured missingness. The Regeneration Establishment Model contains several interactions and transformations of either slope, aspect, or both, which further complicates the problem.

A more practical solution to imputing the remaining missing slope and aspect values would be to estimate them as was done with elevation, using digital elevation maps. However, the low resolution of the geographic coordinates available for plots - often only latitude and longitude at the stand level - makes it unlikely that the added information would improve the models considered.

8.4 Outliers

Frequently in the two-stage analysis, the choice of whether or not to exclude outliers was difficult due to the generally poor quality of the linearity assumed to exist between predictor and response. A common phenomenon was that in considering the removal of a small number of potential outliers (e.g. 2) identified via Cook's Distance, the resulting effect would be to alter the slope of the regression line (indicating high leverage values), which would in turn result in other observations or small clusters being identified in the new model under consideration. The same procedure was repeated several times in some cases. This general lack of robustness in the models was - at the risk of observing the obvious - a result of the simple lack of a strong linear relationship. An example is the second linear regression model predicting the Weibull scale parameter in Structure 2, which has 3 observations (rows 76, 93 and 95) identified as outliers by Cook's distance statistic. The removal of these 3 observations and subsequent re-fitting of the model results in 3 new outliers being identified by the Cook statistic (rows 17, 70 and 78).

8.5 Other Models Considered

The difficulties arising from the inconsistent structure of the data and the two-stage (hurdle) modelling approach used throughout Part 2 suggest a wide array of possible combinations of component models that might justifiably be used to satisfy one or another group of assumptions. Although they have not been explicitly discussed in this document, the following approaches to explaining the variation in either the *count* or *mean* regeneration response in at least 1 of the 3 structures discussed were at one time or another considered and attempted in limited detail:

- Building predictive models from the Structure 1 data with all observations having missing values in abiotic predictors excluded from analysis.
- Building predictive models with data in the second-stage analysis grouped by important predictors[5], rather than by structural variables (`stand`, `plot`, etc.).
- Building predictive models from design matrices (theoretical Structure 4) constructed by randomly selecting individual observations from each plot.
- Building predictive models from (theoretical Structure 5) design matrices constructed by fitting and predicting reverse diameter-growth models to all regeneration trees (original definition, Structure 1) in order to tabulate their expected year of plot-entry as the response variable of interest.
- Using linear mixed-effects models to predict the conditional distribution of stocked plots.
- Building excess-zero *and* excess-1 models (three-stages, rather than two) in the form of two sequentially fit logistic regressions, followed by a conditional distribution of non-zero and non-1 regeneration counts.

8.6 Models That Could Be Considered

Stemming from the list of model frameworks mentioned in the previous section, it is perhaps worthwhile to note a small group of additional possible combinations that were not attempted due to lack of time, but would certainly be interesting to explore in the future. These are presented in the following list:

- Construct tables of the mean regeneration response conditional on the relevant predictors of interest. Fit probability distributions to the present and future mean response, R, and predict the magnitude of *change* in these parameters.

[5] Primarily quadratic mean diameter and habitat type.

- Construct diameter-distribution tables for all trees (whether regeneration or not) in the fitting data. Predict regeneration in an imputation context, where the predicted (e.g., 10 year) future diameter-distribution of trees is compared against the stand-level results of the next FVS cycle of interest.

- Following from the previous possibility, it might be useful to consider a more simplified - or simply different - basic interaction structure for component sub-routines in FVS. In survival regression, it is assumed that the probability of survival is equal to $1-$ the probability of mortality. In FVS, ingrowth and mortality function more or less independently. What we are *really* interested in are the complex seedling survival processes. Begin with the maximum number that could be added to a tree file, and then build better mortality models.

8.7 Recommendations For Future Regeneration Modelling Studies

The largest source of data for this thesis was the Long Term Ecological Research (LTER) component.

Although this portion of the data represents the greatest range of total measurement time over which sampling took place, it also presents some of the greatest problems in terms of analysis. This is because, despite the value implicit in having data collected over a large temporal scale, the original study designs, based upon the information available in the literature, do not hold up well against the basic standards of contemporary sampling design.

An inherent assumption in most experimental design theory is an element of randomness in selection of the sampling unit. As far as can be determined from the available literature, a large portion of the plots in the LTER network were located “subjectively”, suggesting strong potential for bias of some sort on the part of the individual responsible for choosing their locations. Design is inseparable from analysis in statistical modelling, and this is an instance where the latter is being severely compromised by the former.

I believe that there is no substitute in science for sound experimental design. I recommend that a more modern, more efficient, and more consistently implemented sampling design for the LTER network be conceived. Conclusions drawn from studies associated with inappropriate designs are - and should be - subject to severe scrutiny. The design of the network has made the analysis for this project quite difficult.

If an objective of future sampling efforts in the LTER network is to better understand forest regeneration characteristics, then the following recommendations are of critical import:

1. Incorporate regeneration plots into the sampling protocol, measuring root-collar, diameter and species of all seedlings.

2. Assure that the aforementioned protocol measures trees less than 4.5 feet (dbh) in height.

3. If it is not possible to achieve recommendation number 2, then try to assure that there is *no* minimum diameter measured at breast height. Rather, measure all small-diameter trees, regardless of *how* small, and define simple guidelines to deal with problems such as how to determine diameter at breast height when the terminal leader or needles stemming from the terminal leader occur at that position.

4. Sample coarse woody debris (CWD).

5. Measure crown length and width.

It has been shown that predicting diameter from height is easier than predicting height from diameter for small trees. Because height and diameter are primary predictors driving individual tree models like FVS, it follows that having sampling protocols leading to the best estimates of these fundamental random variables is of interest.

The efforts of Shifley et al. (1993) and others to account for variation in sampling design *post facto* not withstanding, the best solution is to alter the sampling design itself. If it is thought that height-to-width prediction is more accurate than width-to-height prediction for small trees, then why persist in the measurement of width in an effort to estimate heights?

The shrub layer in LSOG forests is extremely dense. Santiago (2000) and Harmon and Franklin (1989) have found that nurse logs are important sites for regeneration on montane cloud and LSOG forests (respectively).

Gove et al. (2002) and Ståhl et al. (2002) have suggested methods for relaskope sampling of coarse woody material (CWM), and Ducey et al. (2002) presented methods for efficient sampling of snags. It is highly recommended that CWM sampling be added to the LTER protocol, in order to account for the important effects of nurse logs as regeneration microsites in LSOG stands.

8.8 Penalized Quasi-Likelihood

The Penalized Quasi-Likelihood (PQL) method used to estimate the parameters of the Generalized Linear (binomial) Mixed Model (GLMM) used to predict the probability of stocking in Structure 3 has been the subject of much interest in statistics over the last decade. Breslow (2003) provided an excellent summary of the technique, and compared the performance of the estimation methods used in the PQL approximation[6] in a few different statistical packages (GLIMMIX, MLwiN) with alternative methods such as Laplacian approximation and Gauss-Hermite quadrature (used in SAS PROC NLMIXED). All of these refer to how the marginal likelihood function in the lme component of GLMM models is integrated. In simulations, Breslow (2003) found that in general the PQL methods performed well in the estimation of coefficients, but underestimated their variance components (their within-cluster variance components, in particular) severely. This was especially true for small cluster sizes, e.g. < 40.

The method for fitting GLMM's that generally appears to be most widely accepted in the statistics community is the Gauss-Hermite Quadrature approximation to the marginal likelihood used in SAS PROC NLMIXED. Interestingly, Breslow (2003) found that a 6th-order Laplace approximation to the likelihood function was "at least as accurate" (Breslow, 2003, p. 14) as the Gauss-Hermite Quadrature approach, even when 20 quadrature points were used.

In order to further examine the quality of parameter estimates calculated using Venable and Ripley's approach to PQL (`glmmPQL()`), I compared coefficient and standard error estimates for a simple GLMM as fit using `glmmPQL` `GLMM` in `R`. Although parameter estimates were generally quite similar for the two methods, estimates of their standard errors initially appeared to vary substantially between PQL estimates returned by `glmmPQL` and `GLMM`. However, the large variation initially observed appears to have been the result of an error in the previous version of the `GLMM` function. Current versions of the function result in reasonably close estimates of coefficient standard errors. The coefficients and estimated standard errors for a simple model with a random intercept estimated for each *studyID* nested in `dataCode` and one fixed effect (log.qmd) using the fitting data from Structure 3 is shown in Table 8.1. It appears that there is little variation in the estimates and standard errors returned by the methods, other than the obiously substantial bug in the earlier version of `GLMM`.

Table 8.1 Comparison of generalized linear mixed-effects model coefficients using different estimation methods

Function	Value	SE	DF	z	Pr (>|z|)
glmmPQL	4.3	0.51	4981	8.5	$2.000e-16$
GLMM(old Laplace)	4.3	8.1	4981	0.53	$5.943e-01$
GLMM (new PQL)	4.3	0.52	4981	8.3	$2.000e-16$
GLMM (new Laplace)	4.3	0.52	4981	8.3	$2.000e-17$

[6] There are several different estimation techniques referred to as PQL. These are not the same.

8.9 Conclusions

1. The data are not Zero-Inflated Poisson distributed.

2. The Zero-Inflated Poisson regression model is not more useful than the two-stage approach.

3. Both the two-stage Weibull and the Zero-Inflated Negative Binomial regression models are useful in understanding the process.

4. The probability of any future regeneration occurring varies negatively with quadratic mean diameter, and negatively with total basal area. However, the effect is small.

5. Quadratic mean diameter is the most important predictor of the conditional distribution of stocked plots.

6. The conditional distribution of stocked plots varies negatively with increasing quadratic mean diameter, and the relationship is approximately linear if the log transformation of both predictor and response is used.

7. Not taking into account the observed allometric relationship between regeneration and quadratic mean diameter, the simple ANOVA model violated model assumptions to the least degree – becuase it does not assume any linear relation between predictor and response – and was thought to be the best model.

APPENDIX I: Variable names and definitions (tree file)

1. `DataCode` Data set code

2. `StudyID` Study identification code

3. `Stand` Stand identification code

4. `Plot` Plot number

5. `Tag` Tree tag number

6. `Year` Year of measurement

7. `Xcoord` X-coordinate location of tree

8. `Ycoord` Y-coordinate location of tree

9. `Species` Tree species

10. `DBH` Tree dbh (cm)

11. `Bam` Tree basal area (m^2/ha)

12. `Cclass` Tree canopy class

13. `Vigor` Tree vigor code

14. `Crratio` Tree crown ratio

15. `Status` Tree status code

16. `Age` Tree age

17. `Height` Tree height (m)

18. `Mcrbase` Measured height (m) to base of live crown

19. `Ccrbase` Calculated height (m) to base of live crown (from height, crown ratio)

20. `Mclngth` Measured crown length (m)

21. `Cclngth` Calculated crown length (m)

22. `Expf` Plot expansion factor for tree in stand or plot

23. `Bamha` Tree basal area per hectare (m^2/ha)

24. `Totbamha` Total basal area per hectare (m^2) in stand or plot
25. `Qmd` Quadratic mean diameter (cm) in stand or plot
26. `Rd` Curtis' relative density in stand or plot
27. `Slope` Slope (degrees) of stand or plot
28. `Stndarea` Stand area (ha)
29. `Plotarea` Plot area (ha)
30. `Cnstmin` Constant minimum diameter in stand or plot (Y/N)
31. `Elev` Elevation (m) of stand or plot
32. `AvAspect` Average aspect (degrees) of stand or plot
33. `Lat` Latitude (degrees) of stand or plot
34. `Long` Longitude (degrees) of stand or plot
35. `Habtype` Primary habitat type of stand or plot
36. `Sitespp` Species used for site index
37. `Si` Site index (not available)
38. `Establyr` Year of stand (not study) establishment
39. `Origin` Stand origin (N=natural, P=planted)
40. `Trmt` Silvicultural treatment (0=unthinned, 1=PCT, 2=CT, 3=PCT+CT)
41. `Thinyr1` Year of first thinning
42. `Thinyr2` Year of second thinning
43. `Thinyr3` Year of third thinning
44. `Thinyr4` Year of fourth thinning
45. `Thinyr5` Year of fifth thinning
46. `Thinyr6` Year of sixth thinning

APPENDIX II: Additional variable names and definitions (fitting data)

1. `any.regen` Any trees with status=2 in stand or plot at following sampling (1=Y, 0=N)
2. `num.regen` Number of trees with status = 2 (or as defined in Section 7.1) in stand or plot at subsequent sampling
3. `numSpreg` Number of species of trees with status = 2 (or as defined in Section 7.1) in stand or plot at subsequent sampling
4. `mean.year` Average year (between sampling)
5. `all.regen.plot` Total number of regeneration trees observed on a plot, all sampling years
6. `mean.regen.ha.year` Mean annual future regeneration per hectare before subsequent sampling
7. `log.mean` Natural log of `mean.regen.ha.year`
8. `any.mort` Any trees with status = 6 in stand or plot at current sampling (1=Y, 0=N)
9. `num.mort` Number of trees with status = 6 in stand or plot at current sampling
10. `coast` Coastal proximity (1/0). The following stands have value=1: OL01, OL02, OL03, OL04, HR01, HR02, HR03, HR04, HS01, HS02, HS03, WP17, CH01, CH03, CH04, CH06, CH07, CH08, CH09, CH10, CH13, CH14, CH15, CH16, CH17, CH22, CH23, CH41, CH42, NCNA, SI04, SI05, SI06, SI07, SI08, SI09, SI10, WP7, WP50
11. `time` Number of years between present and subsequent plot sampling
12. `ysd` Number of years since most recent thinning, or since stand establishment
13. `minEst` Estimated mininum diameter threshold (MDT) measured
14. `log.qmd` Natural log of quadratic mean diameter in stand or plot
15. `qmdClass` Value (1:10) assigned to plot based on observed deciles of `qmd` as described in Section 7.4
16. `log.ba` Natural log of total basal area per hectare (m^2/ha) in stand or plot
17. `ba2` Square of total basal area per hectare (m^2) in stand or plot

BIBLIOGRAPHY

Steven A. Acker, W. Arthur McKee, Mark E. Harmon, and Jerry F. Franklin. *Long–Term Research on Forest Dynamics in the Pacific Northwest: a Network of Permanent Forest Plots*, volume 21 of *Man and the Biosphere Series; Forest Biodiversity in North, Central and South America and the Caribbean.* The Parthenon Publishing Group, Pearl River, New York, 1998.

D. B. Botkin, J. F. Janak, and J. R. Wallis. A simulator for northeastern forest growth: A contribution of the hubbard brook ecosystem study and ibm research. Research Report 3140, IBM, 1970.

Norman Breslow. Whither pql. UW Biostatistics Working Paper Series 192, University of Washington, 2003.

A. Colin Cameron and Pravin K. Trivedi. *Regression analysis of count data.* Cambridge University Press, 1998.

Yin Bin Cheung. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*, 21:1461–1469, 2002.

Nicholas L. Crookston. User's guide to the most similar neighbor imputation program: Version 2. General Technical Report 96, USDA Forest Service Rocky Mountain Research Station, September 2002.

D. C. Dey. *A comprehensive Ozark regenerator.* PhD thesis, University of Missouri, Columbia, 1991.

Mark J. Ducey, Greg J. Jordan, Jeffrey H. Gove, and Harry T. Valentine. A practical modification of horizontal line sampling for snag and cavity tree inventory. *Canadian Journal of Forest Research*, 32:1217–1224, 2002.

Alan R. Ek, Andrew P. Robinson, Phillip J. Radtke, and David W. Walters. Development and testing of regeneration imputation models for forests in minnesota. *Forest Ecology and Management*, 94:129–140, 1997.

Dennis E. Ferguson and Clinton E. Carlson. Predicting regeneration establishment with the prognosis model. Research Paper INT–467, USDA Forest Service, August 1993.

Dennis E. Ferguson and Nicholas L. Crookston. User's guide to version 2 of the regeneration establishment model: Part of the prognosis model. General Technical Report INT–279, USDA Forest Service, May 1991.

Dennis E. Ferguson and Ralph R. Johnson. Developing variants for the regeneration establishment model. In Alan R. Ek, Stephen R. Shifley, and Thomas E. Burk, editors, *Forest growth modelling and prediction: Proceedings of the IUFRO conference*. USDA Forest Service, 1998.

Dennis E. Ferguson, Albert R. Stage, and Raymond J. Boyd. Predicting regeneration in the grand fir-hemlock ecosystem of the northern Rocky Mountains. 32(1), 1986.

K.L. Froese, V. Lemay, P. Marshall, and A.A. Zumrawi. Regeneration imputation models for prognosis bc, idfdm2 subzone variant, invermere forest district. Technical Report R02–07, Forestry Innovation Investment, 2003.

Jeffrey H. Gove, Mark J. Ducey, and Harry T. Valentine. Multistage point relascope and randomized branch sampling for downed coarse woody debris estimation. *Forest Ecology and Management*, 155:153–162, 2002.

Daniel B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56:1030–1039, 2000.

D. A. Hamilton. Event probabilities estimated by regression. Research Paper INT–152, USDA Forest Service, 1974.

M. E. Harmon and J. F. Franklin. Tree seedlings on logs in *Picea–Tsuga* forests of oregon and washington. *Ecology*, 70:48–59, 1989.

Frank E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag New York, Inc., 2001.

Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

James K. Lindsey. *Applying Generalized Linear Models*. Springer-Verlag New York, Inc., 1997.

Melinda Moeur and Albert R. Stage. Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science Monographs*, 41(2):337–359, 1995.

John Neter, Michael H. Kutner, Christopher J. Nachesteim, and William Wasserman. *Applied Linear Statistical Models*. The McGraw-Hill Companies, Inc., 3 edition, 1996.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, 185:71–110, 1894.

Jose C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag New York, Inc., 2000.

David T. Price, Niklaus E Zimmermann, Peter J Van Der Meer, Manfred J. Lexer, Paul Leadley, Irma T. M. Jorritsma, Jorg Schaber, Donald F. Clark, Petra Lasch, Steve McNulty, Jianguo Wu, and Benjamin Smith. Regeneration in gap models: priority issues for studying forest responses to climate change. *Climate Change*, 51:474–508, 2001.

Xiao Qin, John N. Ivan, and Nalini Ravishanker. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis and Prevention*, 938:1–9, 2003.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

Eric Ribbens, John A. Silander Jr., and Stephen W. Pacala. Seedling recruitment in forests: calibrating models to predict patterns of tree seedling dispersion. *Ecology*, 75(6):1794–1806, 1994.

Andrew P. Robinson and Alan R. Ek. The consequences of hierarchy for modeling in forest ecosystems. *Canadian Journal of Forest Research*, 30:1837–1846, 2000.

R. Rogers and P. S. Johnson. Describing population dynamics of northern red oak. In *Symposium on Environmental Constraints and Oaks: Ecological and Physiological Aspects*, Nancy, France, 1993. Centre National de Formation Forestiere.

Robert Rogers and Paul S. Johnson. Approaches to modeling natural regeneration in oak-dominated forests. *Forest Ecology and Management*, 106:45–54, 1998.

Louis S. Santiago. Use of coarse woody debris by the plant community of a hawaiian montane cloud forest. *BIOTROPICA*, 32(4a):633–641, 2000.

J. Schweiger and H. Sterba. A model describing natural regeneration recruitment of norway spruce (picea abies(l.) karst.) in austria. *Forest Ecology and Management*, 97:107–118, 1997.

Oliver Shabenberger and Francis J. Pierce. *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press, 2002.

Stephen R. Shifley, Alan R. Ek, and Thomas E. Burk. A generalized methodology for estimating forest ingrowth at multiple threshold diameters. *Forest Science Monographs*, 39(4):776–798, 1993.

Albert R. Stage. Prognosis model for stand development. Research Paper INT–137, USDA Forest Service, 1973.

Göran Ståhl, Anna Ringvall, Jeffrey H. Gove, and Mark J. Ducey. Correction for slope in point and transect relascope sampling of downed coarse woody debris. *Forest Science Monographs*, 48(1):85–92, 2002.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics in S-Plus*. Springer-Verlag New York, Inc., 4 edition, 1994.

A.H. Welsh, R.B. Cunninham, C.F. Donnelly, and D.B. Lindenmayer. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88:297–308, 1996.

Kelvin K. W. Yau and Andy H. Lee. Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention program. *Statistics in Medicine*, 20:2907–2920, 2001.