HIERARCHICAL SEGMENTED REGRESSION MODELS

WITH APPLICATION TO A WOOD DENSITY STUDY

by

Li Xing

MSc Math, University of British Columbia, 2005

BA Applied Math, Hebei University of Technology, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the Department
of
Statistical and Actuarial Science

SIMON FRASER UNIVERSITY

Fall 2006

APPROVAL

Name: Li Xing

Degree: Master of Science

Title of thesis: Hierarchical Segmented Regression Models with Application to a Wood Density Study

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Charmaine Dean, Senior Supervisor

Dr. Carl Schwarz

Dr. Leilei Zeng
External Examiner

Date Approved: _____

Abstract

Wood density is an important characteristic of wood, and plays a major role in determining the strength of wood products. One quantity useful in calculating wood density is called area-increment, which is the estimate of the cross-sectional area of the last ring at a horizontal cut of a tree. The focus of this study is modeling of area increment as a function of a scaled measurement of the tree-height at which area increment was determined from samples taken at various heights of 60 lodgepole pine trees in British Columbia, Canada. Lodgepole pine is an important commercial species that is highly responsive to intensive management practices and is grown for a wide variety of wood products.

The relationship between area increment and scaled tree height is approximated by a hierarchical segmented regression model. Slopes of the segments vary over trees in this mixed-effect modeling framework; it is also of interest to determine whether any covariates are explanatory for variation observed. Maximum likelihood estimation is performed for inference concerning the model parameters and the model is assessed using a variety of techniques.

Acknowledgments

First I would like to express my thanks to my supervisor, Professor Charmaine Dean, who spent considerable time giving me detailed advice, going through the analysis with me and connecting me with the forest scientists.

I would also like to thank the professors in our statistics department who taught me valuable statistical knowledge and trained me in the art of consulting with subject-area researchers.

Thanks to Drs. Goudie and Parish, of the Ministry of Forests of British Columbia, for providing the data and for their valuable suggestions.

Special thanks to Darby Thompson for his help in programming at the start of this study and thanks also to all the graduate students in the department for sharing their expertise with me.

Contents

Chapter 1

Introduction

Wood density is an important attribute of wood influencing the quality of products generated from raw wood material. In particular, it has considerable influence on the strength of solid wood products and affects both the yield and fibre properties of pulp produced (for example, Zobel and van Buijtenen, 1989). There are several indices used to describe wood density, including a quantity termed area increment, which forms the focus of this study.

Area Increment, AI, is the cross-sectional area of the last ring at a horizontal cut of a tree (Goudie et al. 2004). It is defined as $AI = \pi(R_2^2 - R_1^2)$ where R_2 (R_1) is the radial measure from the pith to the outside (inside) of the last ring on bole disks removed from a tree. Note here it is assumed that rings are circular and the bias of the irregular rings is ignored due to the prohibitive cost of accounting for such. In addition, for lodgepole pine, which can have very tight rings, the bias is small. AI varies with the height at which the bole disk is taken. Since radial measures of the last ring are difficult to obtain, and to compensate for annual climate differences, AI is obtained by prediction using values from the previous 10 years using a simple quadratic regression model. The radial measures are taken from X-ray scan data performed by Forintek Canada Corporation.

Goudie et al. (2004) found that trees growing under very different conditions,

open crown to suppressed, for example, exhibit comparable patterns in AI measured at various tree heights if the bole location was scaled relative to the height of the crown centroid, which is the vertical center of foliar biomass, with half above and half below the centroid. Let x indicate the Relative Height to Crown Centroid and it is defined by

$$x = \begin{cases} \dfrac{C-h}{G-C} & \text{if } h \geq C \\ \dfrac{C-h}{C} & \text{otherwise} \end{cases} \tag{1.1}$$

where C is the crown centroid and G is the total height of the tree (See Figure 1.1). Hence x is a scaled measurement of the height of the cut with 0 at the crown centroid, 1 at the the base of the tree and -1 at the top of the tree. The relationship between AI and x for any tree is expected to follow a segmented model. The focus of this project is to describe this segmented relationship across trees using mixed effects models and to identify whether there are covariates which may explain heterogeneity in slopes of segments over trees.

1.1 Wood Density Data Structure

Data on AI and x from 60 lodgepole pine trees were provided by the British Columbia Ministry of Forests. There were a total of 729 bole disk measurements of these quantities. The data are unbalanced over trees and Table 1.1 provides the frequency of the number of measurements per tree which tells the data are not overly sparse.

Summary information on AI and x are presented in Table 1.2 and Figure 1.2. There are 29 large values of AI corresponding to 13 trees and 18 (62%) of these arise from 4 trees labeled C41, C44, C40 and J31. More than 75% of the bole disks are taken below the crown centroid.

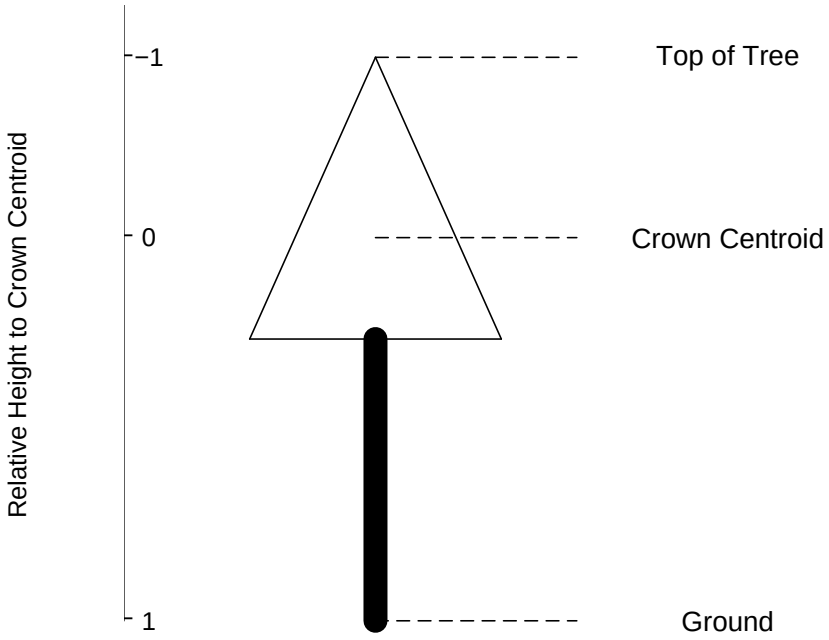

Figure 1.1: Tree Schematic

Table 1.1: Number of Observations on Each Tree

No. of Observations	8	9	10	11	12	13	14	15	18
No. of Trees (Percentage)	3 (5%)	3 (5%)	5 (8%)	2 (3%)	25 (42%)	7 (12%)	13 (22%)	1 (2%)	1 (2%)

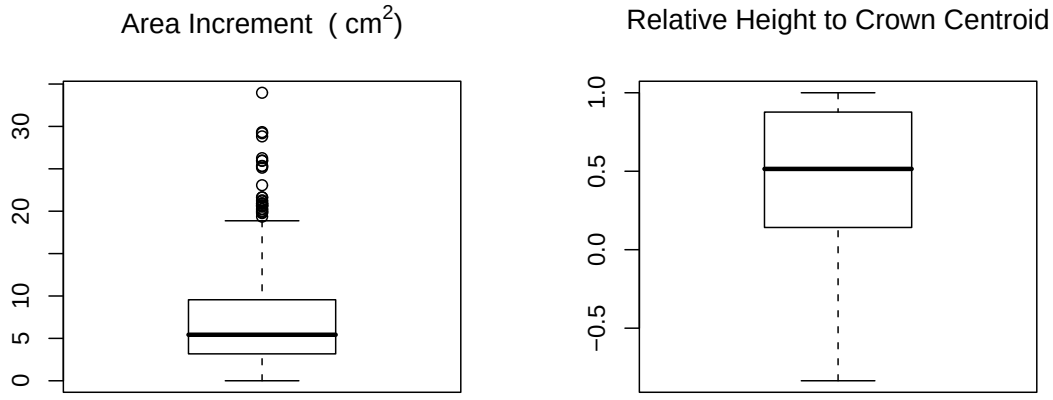

Figure 1.2: Boxplots of AI and x

Table 1.2: Summary Statistics on Area Increment (AI) and Relative Height to Crown Centroid (x)

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Area Increment (cm^2)	0.00	3.18	5.43	6.94	9.56	33.97
Relative Height to Crown Centroid	-0.84	0.14	0.52	0.44	0.88	1.00

Figure 1.3 plots AI versus x for four trees chosen because they portray the typical pattern. These plots illustrate the proposed segmented relationship between AI and x. From a scientific point of view, a mature tree has three different parts: the base, the main trunk and the crown, each of which has different growth pattern. Figure 1.3 displays fits of three-segmented regression models to the data from these four representative trees, C34 , C44, J32 and J33; estimates are obtained by maximum likelihood separately for each tree. Note the similar pattern among the trees for the relationship between AI and x for the three sections of a tree with a positive slope in the crown, AI about constant over the main trunk and then sharply increasing to the base.

Plots of the data for all 60 trees are provided in Appendix A. In the plots, trees are listed by crown type. There are quite a few trees with lack of growth in AI at the base, for example, trees labeled B22 (Figure A.1), C33 (Figure A.2), C41 (Figure A.2), J72 (Figure A.3), J93 (Figure A.3), C43 (Figure A.4) and T95 (Figure A.4). Older trees tend to exhibit this swelling at the base, so a lack of such may be due to immaturity. There is also substantial variability in the third segment which ranges from sharply increasing to sharply decreasing. We focus here on analysis of data in two parts of the tree, the main trunk and the crown, with $x < 0.95$; fits of three-segmented models to the data yield a second change point at about 0.95 and this value is also suggested as preliminary estimate of the general position of the base of the tree based on scatter

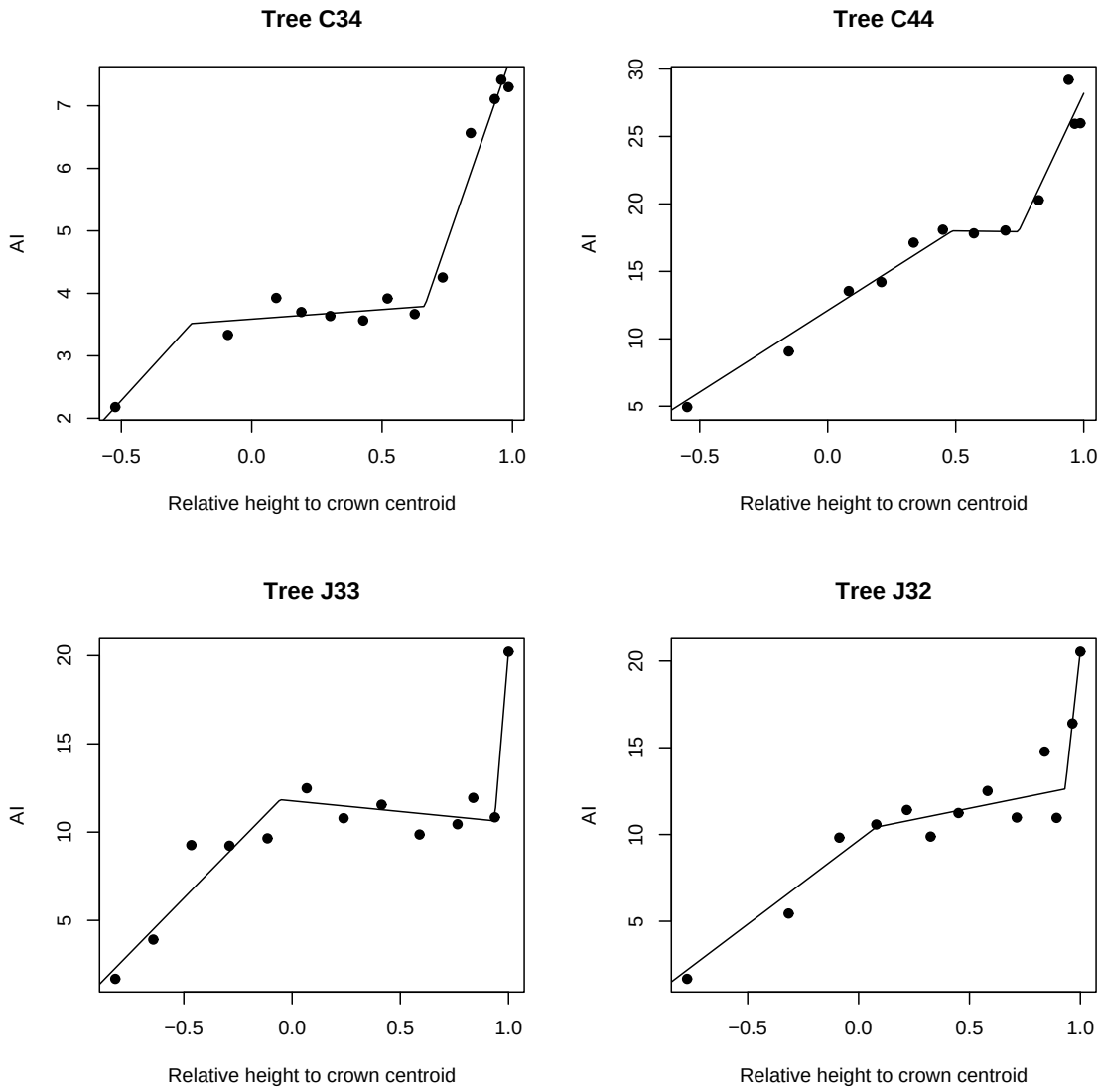

Figure 1.3: Scatter plots of AI versus Relative Height to Crown Centroid for four trees overlaid with fitted three-segmented models

plots of the data. Consultation with the scientists suggested that a model constrained to pass through (-1, 0) also seemed appropriate, i.e. AI=0 at the top of the tree. We also briefly consider analysis of the full data with x from -1 to 1 and comment on the difficulties of this analysis as well as mechanisms for further exploration of this data.

Other information on the trees, which may be useful in the analysis, are tabulated in the following three tables. Table 1.3 includes information on the characteristics of the crown; Table 1.4 includes information on the characteristics of the stem and finally Table 1.5 includes information on the properties of the tree. Of these variables, crclass, total_crown_foliage_biomass and bh-age were identified by researchers at the Ministry of Forests as being of prime interest.

Table 1.3: Characteristics of the crown

Variable	Unit	Description
foliar_volume	m^3	foliar volume (weighted by foliar age class)
crclass		Crown class (Dominant, Co-dominant, Intermediate, Suppressed, Open grown)
crown_area	m^2	projected crown area
Total_crown_foliage_biomass	kg	sum of foliar biomas (sumwt) across all internodes
Total_ht	m	total tree height (m)

Table 1.4: Characteristics of the stem

Variable	Unit	Description
bh_age	years	age at breast height at sampling
bole_volume	m^3	whole-tree bole volume
dbhob	cm	diameter outside bark @ 1.3m
ht2cent	m	vertical distance from ground to vertical center of foliar biomass
ht_2_crown_base	m	height from germination point to crown base
total_age	years	age at stump height at sampling

Table 1.5: Properties of the tree and the stand

Variable	Unit	Description
avg_diam	cm	arithmetic average stand diameter
avg_ht	m	arithmetic average stand height
basal_area	m^2/ha	stand basal area
mean_sample_bhage	years	mean breast height age of sample trees taken from stand
mean_sample_stumpage	years	mean stump age of 3-5 sample trees taken from stand
merch_vol	m^3/ha	merchantable volume/ha
stems_per_ha	No./ha	number of live stems per ha
topht_pl	m	top height of the lodgepole pine trees
total_volume	m^3/ha	stand total volume/ha

1.2 Plan of the Project

The description of hierarchical segmented models and inference for such models are presented in Chapter 2. In Chapter 3, we use a two-segmented model to analyze the wood density data of the main trunk and crown and also provide a discussion of three-segmented modeling for the full data. The project ends with a discussion of ideas for future pursuit.

Chapter 2

Hierarchical Segmented Regression Model

2.1 Introduction

Segmented regression models have been used in many other biological settings (e.g. Leites et al. 2004). Here we consider hierarchical segmented models with change points estimated as fixed effects but random slopes over individuals. We wish to investigate whether such segmented models offer a good approximation to the trend in AI growth.

The mathematical description of the hierarchical K-segmented model is provided in Section 2.2. Maximum likelihood estimation is discussed in Section 2.3.

2.2 The Model Description

Let $\boldsymbol{Y}_j = (y_{1j}, y_{2j}, \cdots, y_{n_j j})^T$ denote the response vector, representing AI measured on the jth tree and $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \cdots, x_{n_j j})^T$ be the corresponding scaled height vector, Relative Height to Crown Centroid, at which AI was measured. Here n_j is the number of the observations on the jth tree.

The model can be described in the two stages below (Laird et al. 1982 and Davidian et al. 1995).

At the first stage, for tree-specific responses, the model can be described as follows:

$$\boldsymbol{Y}_j = f(\boldsymbol{U}_j, \boldsymbol{x}_j) + \boldsymbol{\varepsilon}_j, \qquad j = 1, 2, ..., J, \tag{2.1}$$

where f is the regression function depending on the explanatory variable $\boldsymbol{x}_j$ and its tree-specific coefficients $\boldsymbol{U}_j$, $\boldsymbol{\varepsilon}_j \sim \mathbf{MVN}(\mathbf{0}, \boldsymbol{R}_{\boldsymbol{n_j}\times\boldsymbol{n_j}})$ independent of $\boldsymbol{U}_j$ with $\boldsymbol{R}$ being the variance-covariance matrix, and observations are independent over trees with J being the total number of trees.

Here we model f as a K-segmented linear model with change points at $a_1, a_2, \cdots, a_{K-1}$, $a_1 < a_2 < \cdots < a_{K-1}$:

$$f(\boldsymbol{U}_j, x) = \alpha_j + \beta_{0j}x + \beta_{1j}(x - a_1)_+ + \beta_{2j}(x - a_2)_+ + \cdots + \beta_{(K-1)j}(x - a_{K-1})_+, \tag{2.2}$$

where

$$(x - a_k)_+ = \begin{cases} 0 & \texttt{if } x \le a_k \\ x - a_k & \texttt{if } x > a_k \end{cases} \qquad (k = 1, 2, \cdots, K-1),$$

β_{0j} is the parameter representing the slope of the first segment for the jth tree, and $\sum_{i=0}^{m} \beta_{ij}$ represents the slope of the $(m+1)$th segment for the jth tree with $m = 0, 1, \cdots, K-1$ and $j = 1, 2, \cdots, J$.

At the second stage, the distributions of the random tree-specific slopes are modeled as:

$$\boldsymbol{U}_j = \begin{pmatrix} \alpha_j \\ \beta_{0j} \\ \beta_{1j} \\ \cdots \\ \beta_{(K-1)j} \end{pmatrix} \sim MVN(\boldsymbol{\gamma}, \boldsymbol{D}_{(K+1)\times(K+1)})$$

where $\boldsymbol{\gamma}$ is the mean vector with components $(\alpha, \beta_0, \beta_1, \cdots, \beta_{(K-1)})^T$ and $\boldsymbol{D}$ is the corresponding $(K+1) \times (K+1)$ variance-covariance matrix.

The mean and variance of $\boldsymbol{Y}_j$ are listed below:

$$\begin{aligned} E(\boldsymbol{Y}_j|\boldsymbol{x}_j) &= E(f(\boldsymbol{U}_j, \boldsymbol{x}_j)) = \boldsymbol{Z}_j\boldsymbol{\gamma} \\ Var(\boldsymbol{Y}_j) &= Var(f(\boldsymbol{U}_j, \boldsymbol{x}_j) + \boldsymbol{\varepsilon}_j) = \boldsymbol{Z}_j\boldsymbol{D}\boldsymbol{Z}_j^T + \boldsymbol{R}_{n_j \times n_j} \end{aligned}$$

where

$$\boldsymbol{Z}_j = \begin{pmatrix} 1 & x_{1j} & (x_{1j} - a_1)_+ & \cdots & (x_{1j} - a_{K-1})_+ \\ 1 & x_{2j} & (x_{2j} - a_1)_+ & \cdots & (x_{2j} - a_{K-1})_+ \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n_jj} & (x_{n_jj} - a_1)_+ & \cdots & (x_{n_jj} - a_{K-1})_+ \end{pmatrix}.$$

2.3 Likelihood Function

Maximum likelihood estimation of this model is based on the marginal density of the response, $\boldsymbol{y}$

$$p(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{a}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{a}) q(\boldsymbol{U}) d\boldsymbol{U} \tag{2.3}$$

where $\boldsymbol{\theta}$ is the vector of variance-covariance components of $\boldsymbol{R}$ and $\boldsymbol{D}$.

The likelihood function of this hierarchical model is

$$\begin{aligned} & L(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{a}|\boldsymbol{y}, \boldsymbol{x}) \\ = & \prod_{j=1}^{J} L_j(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{a}|\boldsymbol{y}_j, \boldsymbol{x}_j) \\ = & \prod_{j=1}^{J} \int p(\boldsymbol{y}_j|\boldsymbol{U}_j, \boldsymbol{x}_j, \boldsymbol{\theta}, \boldsymbol{a}) q(\boldsymbol{U}_j) d\boldsymbol{U}_j \\ = & \prod_{j=1}^{J} \int \frac{1}{(2\pi)^{\frac{n_j+K+1}{2}} |\boldsymbol{R}|^{\frac{1}{2}} |\boldsymbol{D}|^{\frac{1}{2}}} \exp\big(-\frac{1}{2} (\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{U}_j))^T \boldsymbol{R}^{-1} (\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{U}_j))\big) \\ & \exp(-\frac{1}{2} \boldsymbol{U}_j^T \boldsymbol{D}^{-1} \boldsymbol{U}_j) d\boldsymbol{U}_j \end{aligned}$$

Computation of the likelihood function requires numerical methods. Gaussian Quadrature (Pinheiro et al. 1995) is a popular method for obtaining integral estimation. Let $\phi(x)$ be the probability density function of a $N(0,1)$ variable. Gaussian Quadrature uses

$$\int_{-\infty}^{+\infty} g(u)\phi(u)du \approx \sum_{i=1}^{n} g(u_i)w_i$$

where u_is are the quadrature points, w_is are the weights at these points and n is the number of the quadrature points used in the approximation (Abramowitz et al. 1964). When there are several random effects to be integrated, successive applications of one-dimensional Gaussian Quadrature rules may be used to yield:

$$\begin{aligned}
& \int p(\boldsymbol{y}_j|\boldsymbol{U_j}, \boldsymbol{x_j}, \boldsymbol{\theta}, \boldsymbol{a})q(\boldsymbol{U_j})d\boldsymbol{U_j} \\
= & \int \frac{1}{(2\pi)^{\frac{n_j+K+1}{2}}|\boldsymbol{R}|^{\frac{1}{2}}|\boldsymbol{D}|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{U}_j))^T\boldsymbol{R}^{-1}(\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{U}_j))\big) \\
& \exp(-\frac{1}{2}\boldsymbol{U}_j^T\boldsymbol{D}^{-1}\boldsymbol{U}_j)d\boldsymbol{U}_j \\
= & \int \frac{1}{(2\pi)^{\frac{n_j+K+1}{2}}|\boldsymbol{R}|^{\frac{1}{2}}|\boldsymbol{D}|^{\frac{1}{2}}} \exp\big(-\frac{1}{2}(\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{U}_j))^T\boldsymbol{R}^{-1}(\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{U}_j))\big) \\
& \exp(-\frac{1}{2}\boldsymbol{z}_j^{*T}\boldsymbol{z}_j^*)d\boldsymbol{z}_j^* \\
= & \sum_{j_1=1}^{N_1}\cdots\sum_{j_h=1}^{N_h} \frac{1}{(2\pi)^{\frac{n_j+K+1}{2}}|\boldsymbol{R}|^{\frac{1}{2}}|\boldsymbol{D}|^{\frac{1}{2}}} \\
& \exp\big(-\frac{1}{2}(\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{z}^*_{j_1,j_2,\ldots,j_h}))^T\boldsymbol{R}^{-1}(\boldsymbol{y}_j - f(\boldsymbol{x}_j, \boldsymbol{a}|\boldsymbol{z}^*_{j_1,j_2,\ldots,j_h}))\big)\prod_{k=1}^{h} w_{ji}
\end{aligned}$$

where $\boldsymbol{z}^*_{j_1,j_2,\ldots,j_h} = (z^*_{j_1},...,z^*_{j_h})^T$, $h = dim(\boldsymbol{U})$ and N_i is the number the quadrature points for the ith integral on the i random effect with $i = 1, 2, \cdots, h$.

Another common approach uses so-called Restricted Maximum Likelihood (REML) estimation as discussed in Harville (1977). The REML estimator of the mean parameter is a weighted least squares estimator which makes computation easy. Because of this we initially investigated the use of REML for this hierarchical segmented analysis. In a small simulation study, based on a two-segmented model, we evaluated REML

by comparing it with Gaussian Quadrature. The REML routine was programmed in R while SAS was used to perform Gaussian Quadrature (SAS Institute Inc. 1999). Briefly, REML was found to perform reasonably well but less reliably for estimation of variance components. The bias of the estimated variance components is substantially larger for the REML estimator. In this report, Gaussian Quadrature was used for likelihood inference.

Chapter 3

Analysis of the Wood Density Data

The hierarchical two-segmented regression model was employed to investigate the relation between AI and x for values of $x < 0.95$. With the constraint that $f = 0$ when $x = -1$, the regression function becomes:

$$f(\boldsymbol{U}_j, \boldsymbol{x}) = \beta_{0j}(\boldsymbol{x} + 1) + \beta_{1j}(\boldsymbol{x} - a)_+, \tag{3.1}$$

with the variance-covariance matrix $\boldsymbol{D} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix}$, $\boldsymbol{R} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, and a is the single change point, $j = 1, 2, \cdots, 60$.

Initially, we omit a consideration of covariates. Table 3.1 presents parameter estimates of the fitted model; model-based as well as parametric bootstrap standard errors are provided. The parametric bootstrap (Efron et al. 1994) standard errors are based on estimates obtained from the analysis of 1000 sets of data of size 60 generated from the fitted model represented in the second column of Table 3.1. The table also shows the bootstrap bias. The distribution of the bootstrap estimates is provided in Figure 3.1 and there are no striking deviations from normality except for the distribution of the estimates of the change point, a. For the fitted model to the wood density data, the slope of the first segment is about twice that of the second segment. The estimate of the change point is close to the crown centroid. Since the transformation from height to the scaled version of height, Relative Height to

Crown Centroid, is such that the compression of height is different above and below the crown centroid, it may be preferable to fix the change point at zero and refit the model under this constraint. Table 3.2 provides estimates of the model parameters when $a = 0$. Estimates and standard errors are quite close to those obtained from the fit displayed in Table 3.1. Figure 3.2 illustrates the mean of the fitted model with fixed change point at 0 as well as 95% point-wise confidence intervals.

Table 3.1: Parameter Estimates from the Fit of a 2-Segmented Hierarchical Model to the Wood Density Data

Parameter	Estimate	Model-Based Std. Error	Absolute Bootstrap Bias	Bootstrap Std. Error
a	-0.09	0.05	0.00	0.03
β_0	5.93	0.57	0.00	0.55
β_1	-3.15	0.74	-0.03	0.72
σ	1.56	0.08	0.00	0.06
σ_{11}	4.14	1.21	0.02	1.12
σ_{12}	-14.21	3.96	0.03	3.47
σ_{22}	5.09	1.86	0.02	1.60

Table 3.2: Parameter Estimates from the Fit of a 2-Segmented Hierarchical Model with Change Point at 0 to the Wood Density Data

Parameter	Estimate	Model-Based Std. Error
β_0	5.72	0.53
β_1	-3.06	0.73
σ	1.56	0.08
σ_{11}	3.97	1.08
σ_{12}	-12.78	3.53
σ_{22}	5.05	1.82

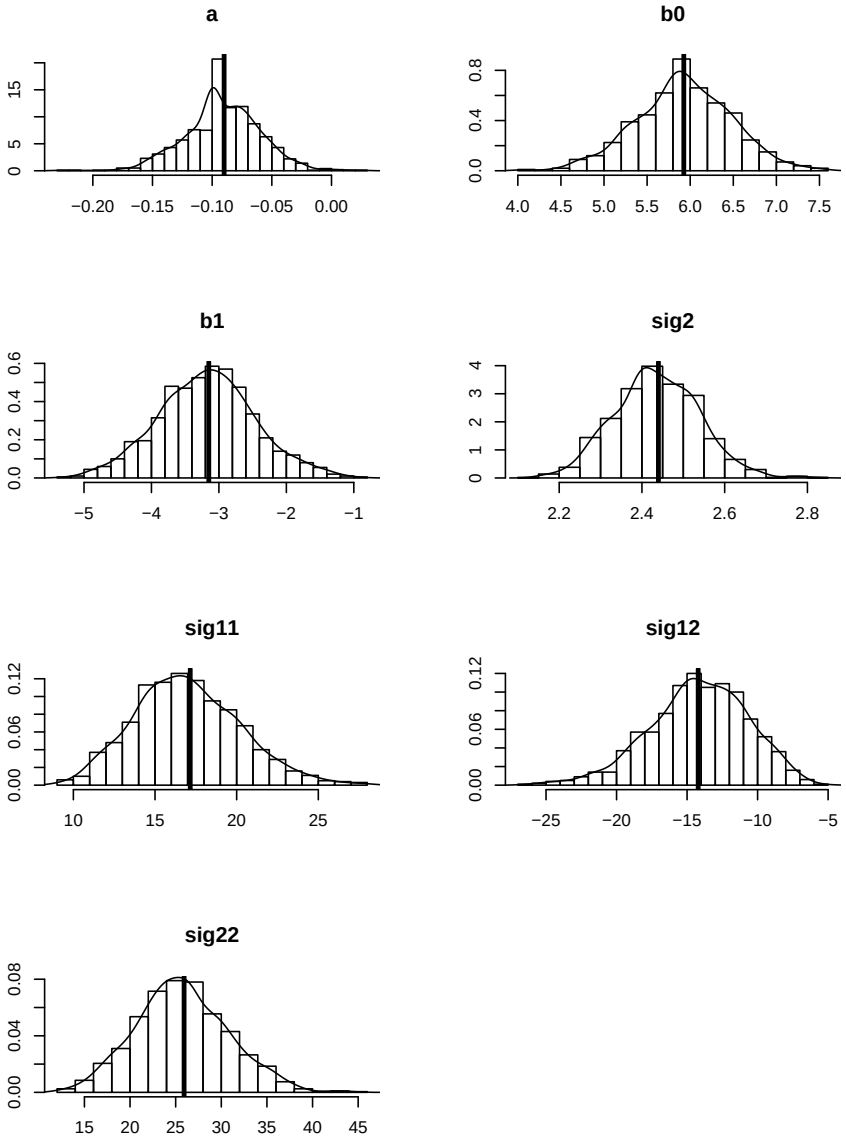

Figure 3.1: The Bootstrap Distribution of Parameter Estimates with the Vertical Line Indicating the Value from which the Data were Simulated.

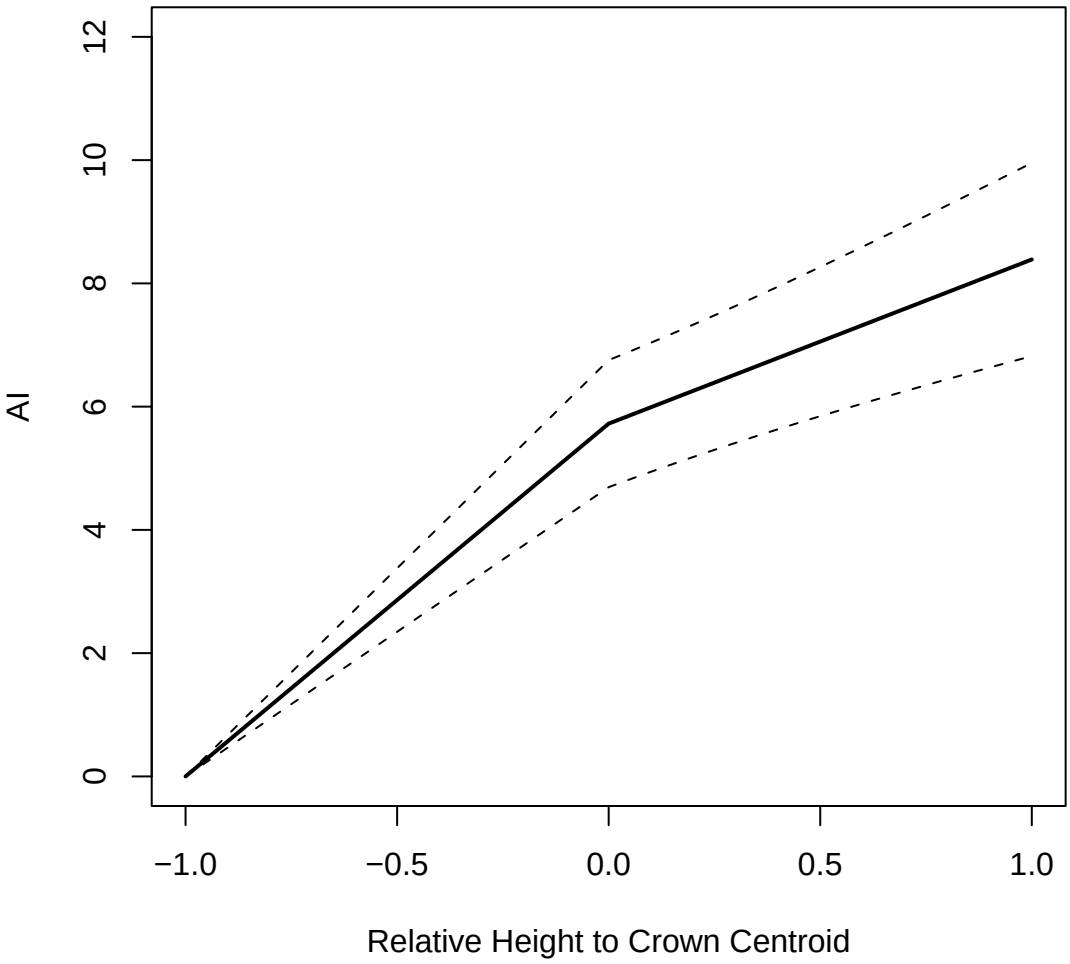

Figure 3.2: The Estimated Overall Mean Response of the Hierarchical 2-segmented Regression Model and its 95% Point-Wise Confidence Interval.

3.1 Model Assessment

In this section, re-sampling and residual analyses are conducted for the purpose of model assessment.

3.1.1 Jackknife Bootstrap

As a re-sampling method, the Jackknife Bootstrap has a variety of uses. Therneau et al. (2001) used it to identify outliers in survival analysis. Here we use this idea to look for potential "outlying trees" which may have considerable influence on the fit of the model.

The goodness-of-fit statistic of interest here is the Akaike information criterion, AIC, (Akaike, 1974) which is defined by

$$AIC = -2\log(\text{Likelihood}) + 2\dim(\Theta)$$

where Θ is the parameter space.

AIC values were calculated by refitting the model with observations from one tree deleted and repeating this for each of the 60 trees one at a time. A boxplot of these Jackknife AIC values is provided in Figure 3.3. Tree J94 is identified as a potential outlier. Other potential outliers identified are J92, J31, C41 and C40. Although J94 is considered an outlier at the tree level, note that this is particularly because of a single outlying observation instead of a general difference in trend for this tree since without this single point, the AIC value decreases from 2686 to 2557.

Figure 3.4 shows the data and the mean response of the fitted segmented model for this tree. In the next section we try to separate the effects of discordant observations versus discordant trees.

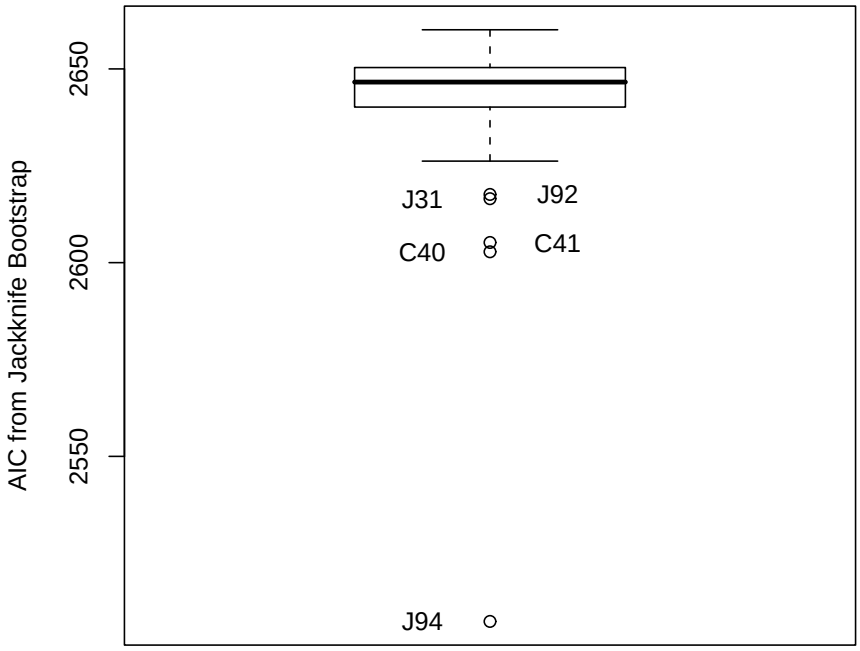

Figure 3.3: Boxplot of AIC from Jackknife Bootstrap

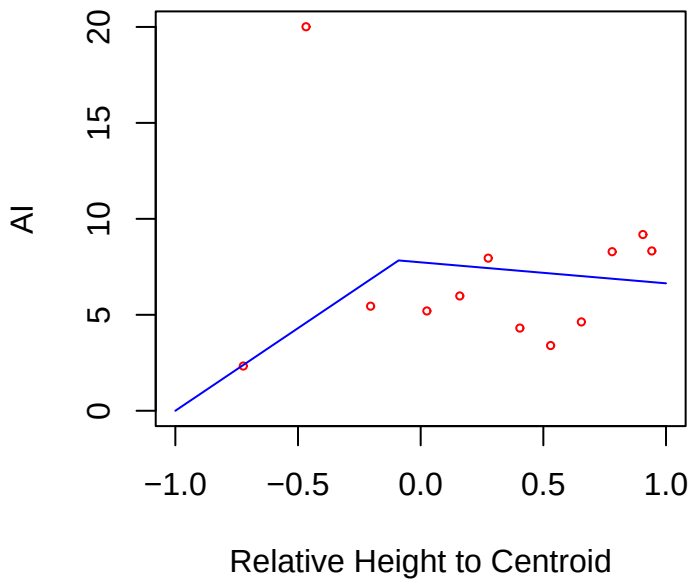

Figure 3.4: Mean Response Estimated from the Segmented Model for Tree J94

3.1.2 Residual Analysis

We consider two types of residuals in our analysis. The first is based on the difference between the response and the mean response of the fitted segmented model for each tree, $\boldsymbol{Y}_j - \boldsymbol{Z}_j\boldsymbol{\gamma}$; this describes the within-tree variation. The standardized within-tree residuals are only approximate as we standardize conditional on the fitted mean being true. The second type of residuals is based on the difference between the mean response for each tree and the overall mean, $f(\boldsymbol{U}_j, \boldsymbol{x}_j) - \boldsymbol{Z}_j\boldsymbol{\gamma}$. These between-tree residuals are based on linear combinations of the estimates of u_j (2.1) and are standardized based on the estimator of D. Figure 3.5 is a plot of the standardized within-tree residuals. Tree J94 stands out due to a large outlier and tree J92 is prominent due to its 3 outliers (2 above and 1 below the ± 2 limits). Figure 3.6 shows the standardized between-tree residuals. The four trees for whom the median values in their boxplots lie above 2 are C41, C44, C40 and J31 (shown from left to right in the plot). Those trees are identified as outliers in sense that fitted mean values for them are far from the overall mean.

Note that though the Jackknife AIC values and the residual analysis were consistent in identifying outlier trees, the residual analysis, though ad hoc, helped to isolate whether the outlier was at the tree or individual observation level.

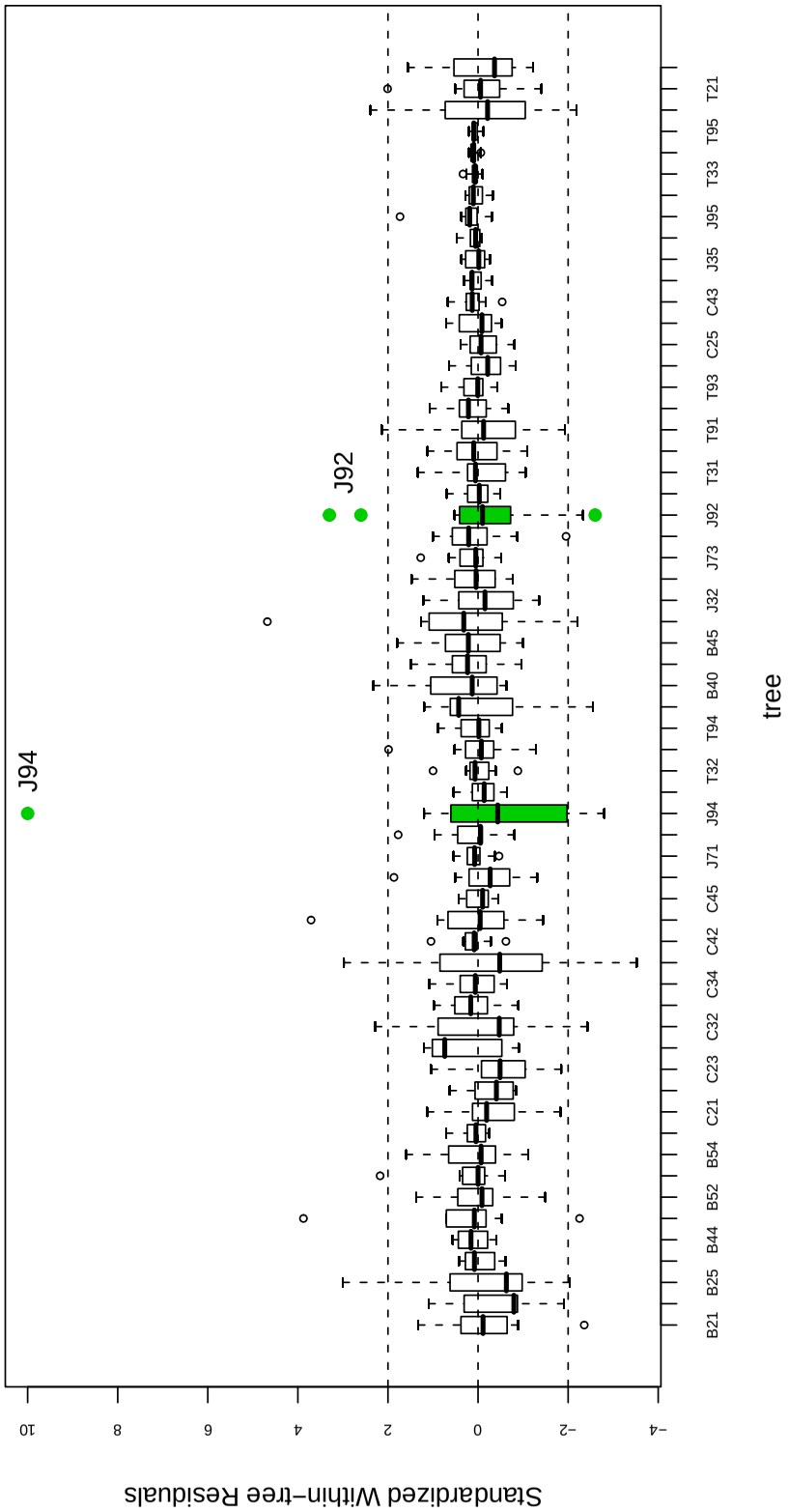

Figure 3.5: Within-tree Residuals

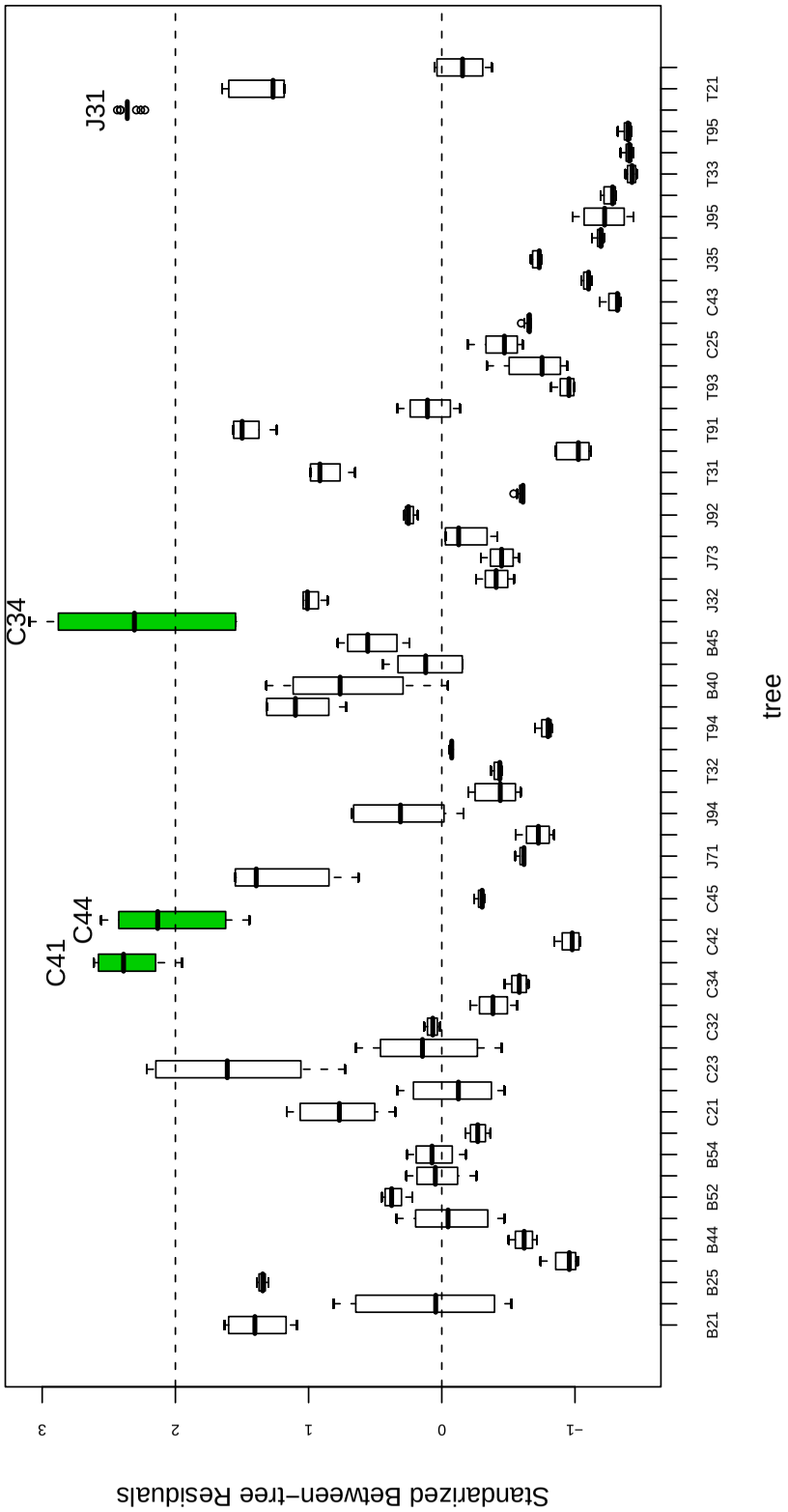

Figure 3.6: Between-tree Residuals

3.1.3 Testing the Model Intercept Constraint

AIC values comparing a model with and without the constraint that $f = 0$ when $x = -1$ were 2686 and 2740 respectively, indicating that the model with the constraint seems to fit the data better.

3.2 Investigating the Effect of Covariates

A variety of strategies was used to construct an initial small set of explanatory variables for further detailed investigation on how well they explain variability in the model. These included discussion with the scientists, backward selection and investigation of the relationship between each variable and Jackknife AIC values obtained by refitting the model with each tree omitted one at a time – Appendix B discusses why these displays may be useful. Plots of the Jackknife AIC values versus values of certain covariates from the omitted tree have been used to investigate relationships. Such relationships were observed between AIC and the seven variables plotted in Figures 3.7 and 3.8: total_ht, Crown_l_2base, crown_area, foliar_volume, dbhob, Total_crown_foliage_biomass, bole_volume. For comparison, Figure 3.9 shows corresponding plots for some of the variables for which no relationships were observed. With backward selection, total_crown_foliage_biomass and total_ht seemed important. As mentioned earlier, scientists expect total_crown_foliage_biomass to be explanatory.

Table 3.3 provides the correlation matrix for the 7 variables identified above. There are some very high correlations observed. Those variables with correlations greater than 0.70 with total_crown_foliage_biomass were not included in the regression analysis. These are crown_l_2base, crown_area, dbhob, and bole_volume. The remaining variables: total_ht, foliar_volume and total_crown_foliage_biomass have low correlations with each other and were included in the 2-segmented model to explain the variability in the slopes for the jth tree:

$\beta_{0j} = \beta^*_{0j} + c_0\text{total\_ht} + d_0\text{foliar\_volume} + e_0\text{total\_crown\_foliage\_biomass},$

$$\beta_{1j} = \beta_{1j}^* + c_1\text{total\_ht} + d_1\text{foliar\_volume} + e_1\text{total\_crown\_foliage\_biomass},$$

where $\begin{pmatrix} \beta_{0j}^* \\ \beta_{1j}^* \end{pmatrix}$ is now the subject-specific random slope vector, $j = 1, 2, \cdots, 60$, and c_is, d_is and e_is are the coefficients for the covariates of interest for $i = 1, 2$.

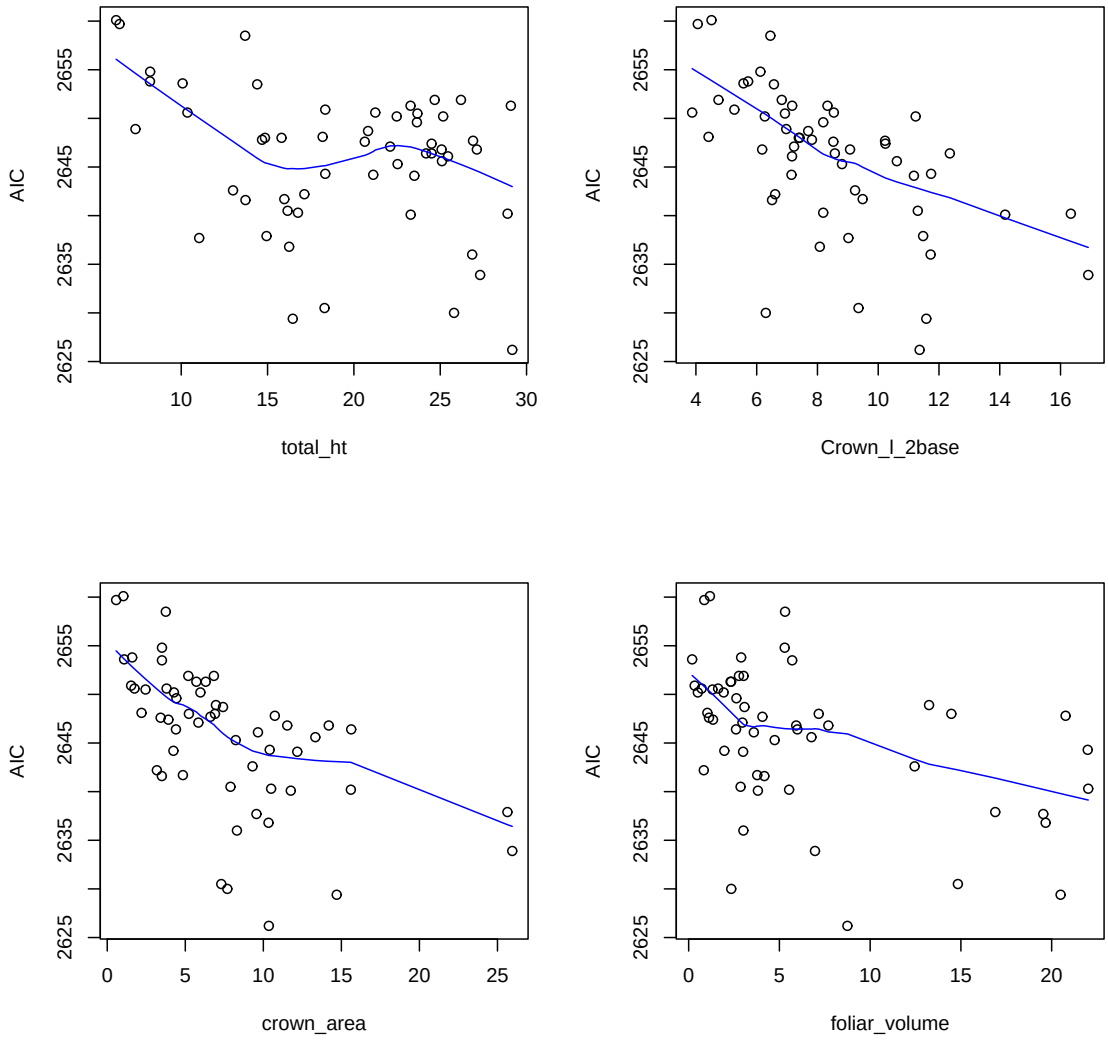

Figure 3.7: Jackknife AIC versus Selected Covariates

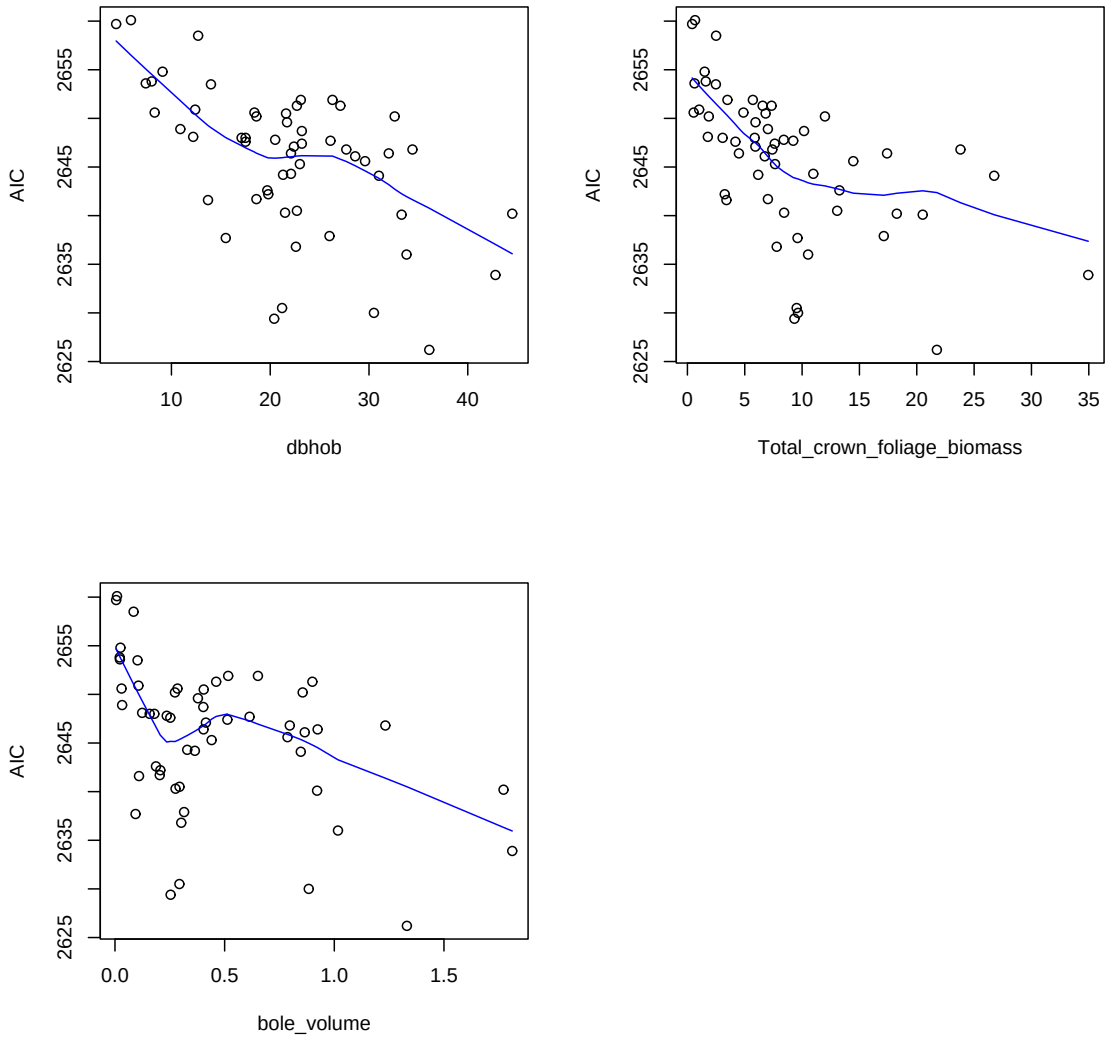

Figure 3.8: Jackknife AIC versus Selected Covariates

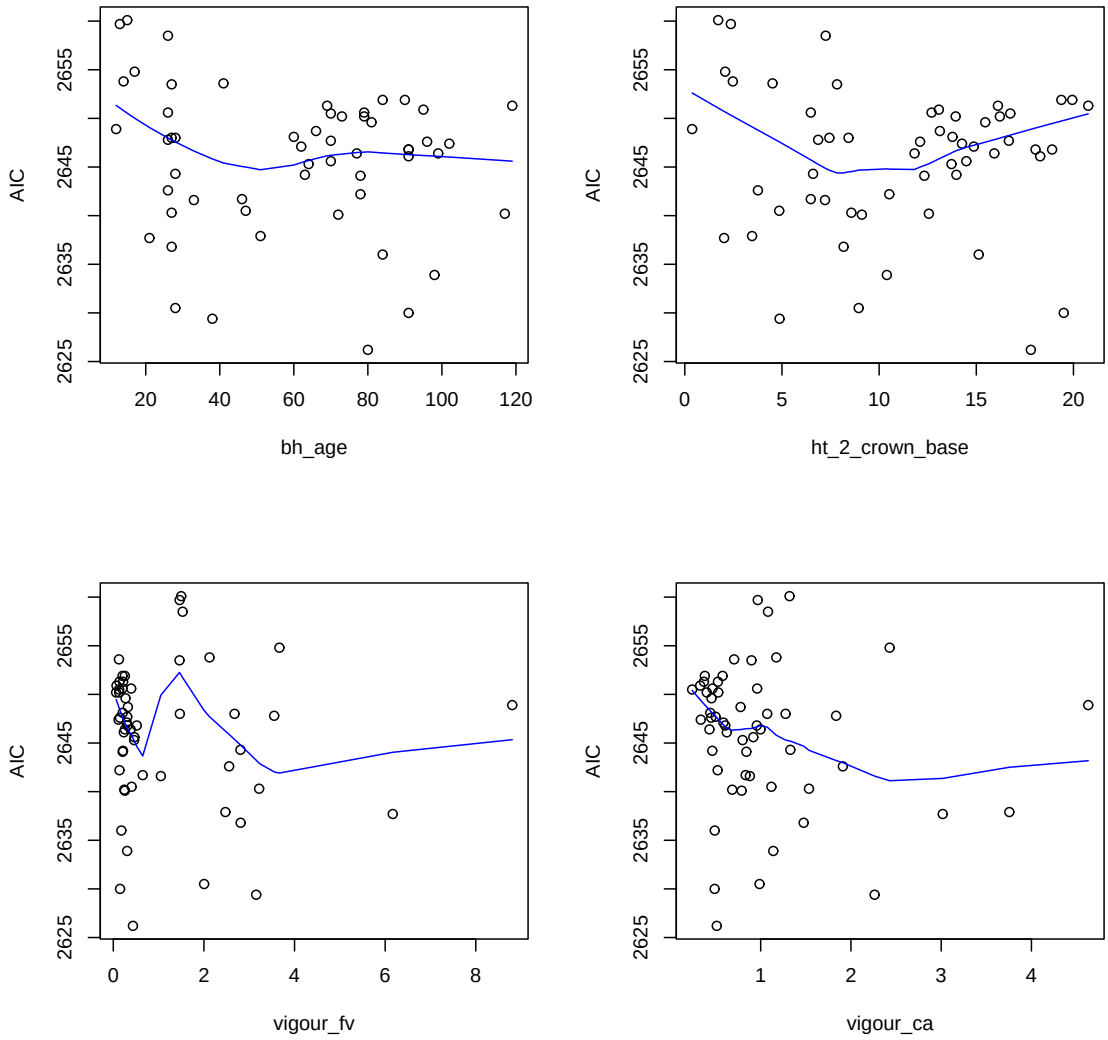

Figure 3.9: Jackknife AIC versus Selected Covariates

Table 3.3: Correlations between Covariates

	total_ht	Crown_l _2base	crown _area	foliar _volume	dbhob	Total_crown _foliage _biomass	bole _volume
total_ht	1.00	0.47	0.37	-0.15	0.83	0.47	0.80
Crown_l_2base	0.47	1.00	0.67	0.33	0.73	0.84	0.66
crown_area	0.37	0.67	1.00	0.65	0.73	0.77	0.68
foliar_volume	-0.15	0.33	0.65	1.00	0.20	0.38	0.09
dbhob	0.83	0.73	0.73	0.20	1.00	0.76	0.94
Total_crown _foliage_biomass	0.47	0.84	0.77	0.38	0.76	1.00	0.73
bole_volume	0.80	0.66	0.68	0.09	0.94	0.73	1.00

When this model was fitted, the estimates of c_0 (p value= 0.40) and c_1 (p value= 0.71) were not significant and the fitted model including only the remaining covariates effects is shown in Table 3.4. The AIC decreased from 2686.2 to 2568.3, i.e. adding these covariates in the model seems to help to explain the variation of the relationship between AI and x across trees. In addition, the estimate of σ_{11} decreases by 40% and that of σ_{22} decreases by 65%. Increasing foliar_volume increases the slope of the first segment and decreases that of the second, while increasing total_crown_foliage_biomass increases both slopes. The estimate of the change point is close to zero. The estimate of the mean slope for the first segment is positive, while that for the second segment is almost zero. Finally, note a preliminary analysis of the full data using a three-segmented model is provided in Appendix A. There is substantial variability in the third segment which can not be explained by covariates using usual regression techniques, for example, backward selection. The fitting of this model needs detailed joint investigation through modeling and discussion with the scientists. Finally we checked the constraint in the model is a reasonable assumption since without the constraint the AIC increase from 2686 to 2740.

Table 3.4: Parameter Estimates from the Fit of a 2-Segmented Hierarchical Model with Selected Covariates to the Wood Density Data

Parameter	Estimate	Model Based Std. Error
a	-0.05	0.05
β_0	1.80	0.41
β_1	-1.63	0.88
σ	1.56	0.08
σ_{11}	1.64	0.38
σ_{12}	-2.74	1.25
σ_{22}	3.33	1.21
d_0	0.41	0.04
e_0	0.11	0.03
d_1	-0.52	0.07
e_1	0.24	0.07

Chapter 4

Discussion

Exploration of an analysis of the full data with a 3-segmented model is required in followup work with the Ministry of Forests. Initially it might be helpful to start with the covariates identified as explanatory for the first two segments in the previous chapter and see which covariates might be helpful in explaining the extreme variability of the third segment. Discussion with the scientists will also help in that regard. Once a small subset of important explanatory variables is identified a full regression analysis, together with residual analyses, may provide a reasonable model. In addition, analyses using spline smoothers with covariates may provide better approximations to between- and within-tree trends, as well as the incorporation of different within- and between-tree clustering random effects. Segmented models with smooth transitions may also be considered (Toms and Lesperance, 2003). Finally the residuals identified in the analysis here need more detailed consideration. For example, the variances of error terms seem different across trees. Bayesian methods may also be considered for this analysis.

Appendix A

Three-segmented Regression Model

We also fitted a 3-segmented regression model for the Wood Density Data: with the constraint that $f = 0$ when $x = -1$, the regression function becomes:

$$f(\boldsymbol{U}_j, \boldsymbol{x}) = \beta_{0j}(\boldsymbol{x} + 1) + \beta_{1j}(\boldsymbol{x} - a_1)_+ + \beta_{2j}(\boldsymbol{x} - a_2)_+, \tag{A.1}$$

with the variance-covariance matrix $\boldsymbol{D} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22}^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33}^2 \end{pmatrix}$, $\boldsymbol{R} = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$, and $j = 1, 2, \cdots, 60$. Note there is considerable variability in the slope of the third segment and the associated standard error is quite large. Table A.1 provides estimates of the parameters in the model while Figures A.1 to A.4 provide posterior estimates of the fitted model for each tree.

Table A.1: Parameter Estimates from the Fit of a 3-Segmented Hierarchical Model to the Wood Density Data

Parameter	Estimate	Model Based Std. Error
β_0	6.20	3.83
β_1	-7.77	7.87
β_2	52.57	59.87
a_1	0.04	0.06
a_2	0.96	0.004
σ	1.60	0.09
σ_{11}	4.26	5.11
σ_{12}	-21.85	77.72
σ_{22}	10.19	29.54
σ_{13}	-2.34	98.08
σ_{23}	20.77	1295.72
σ_{33}	82.53	344.09

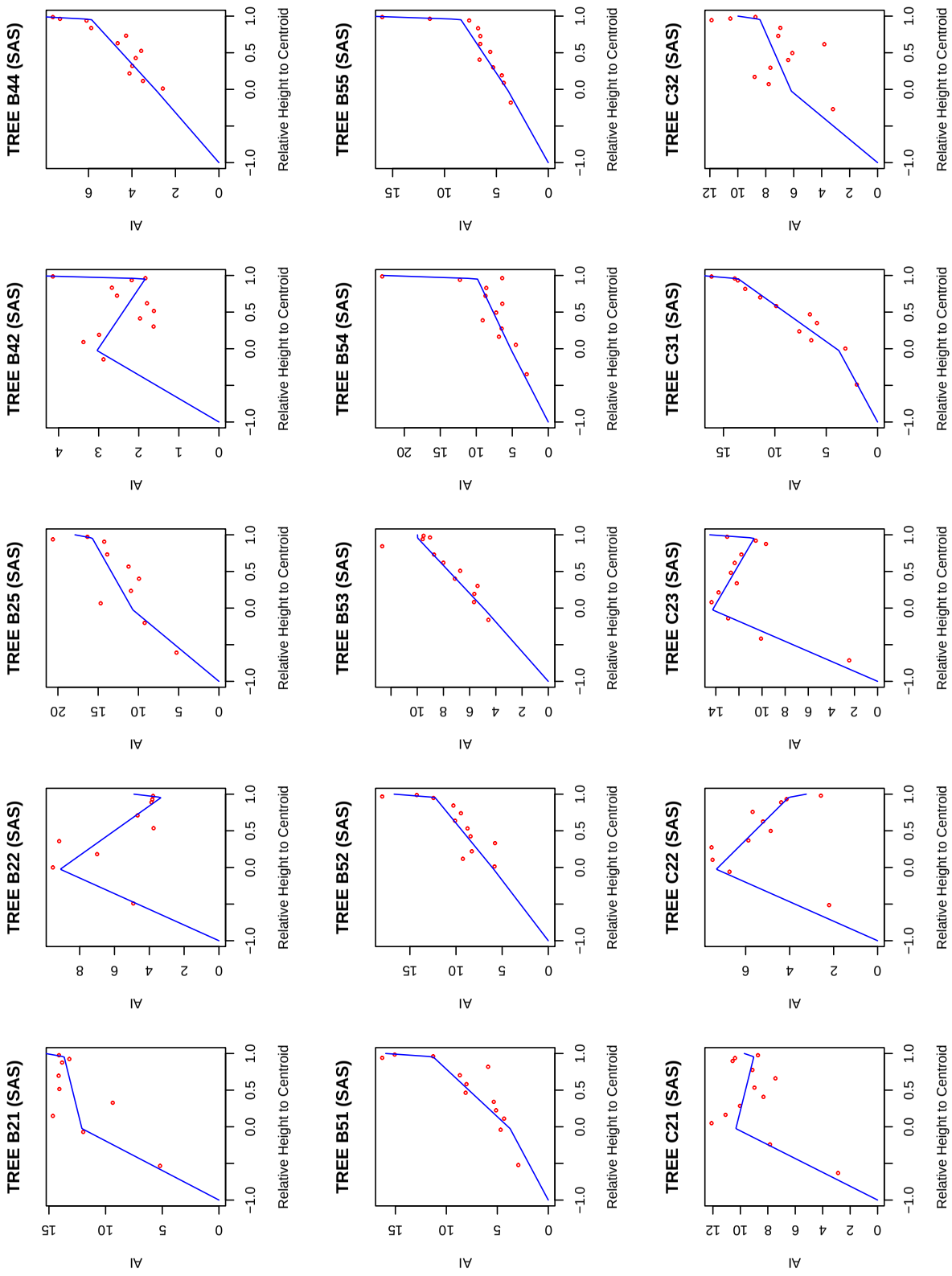

Figure A.1: Plots of the Estimated Mean Response of Each Tree from the 3-Segmented Regression Model

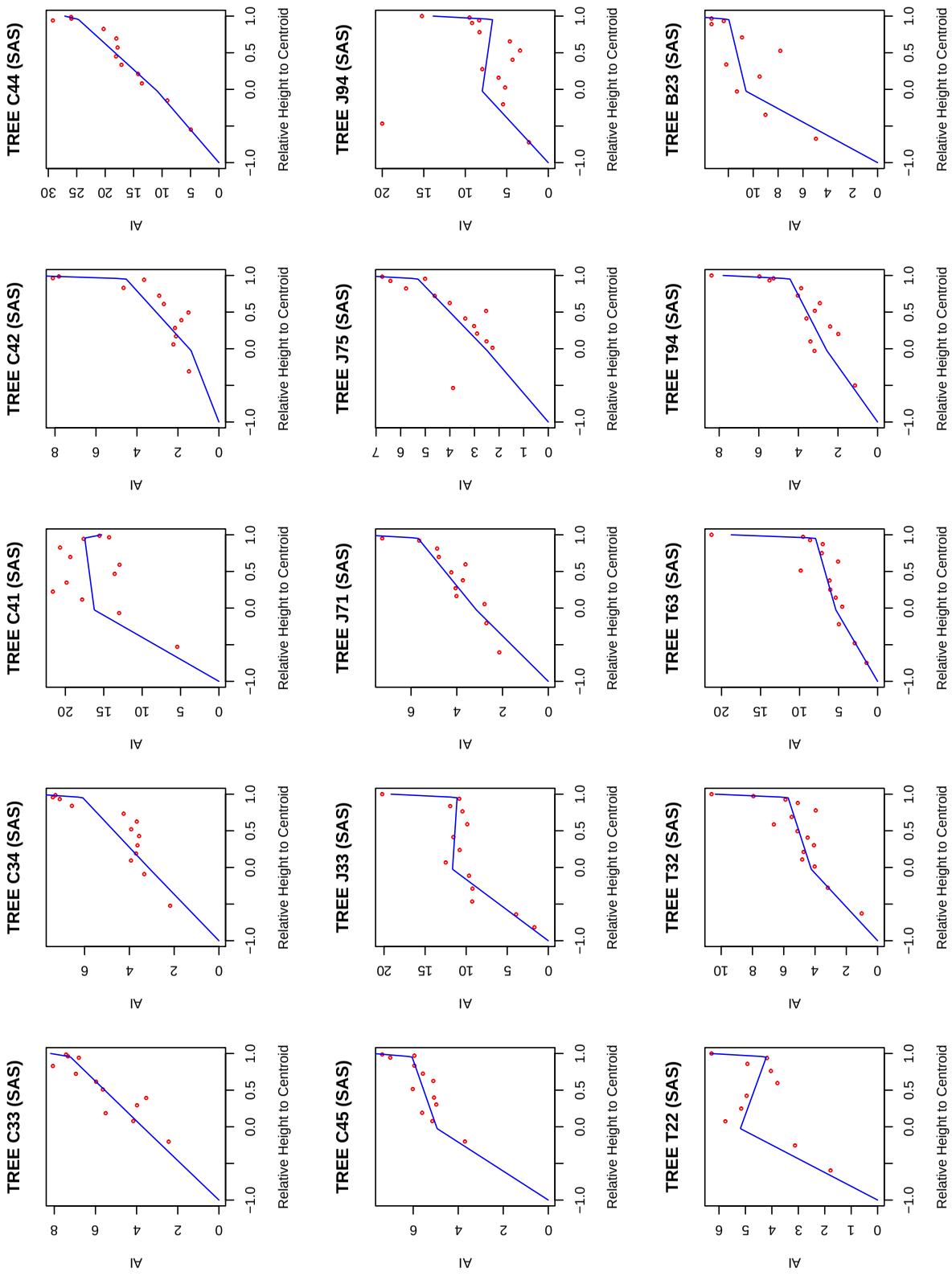

Figure A.2: Plots of the Estimated Mean Response of Each Tree from the 3-Segmented Regression Model

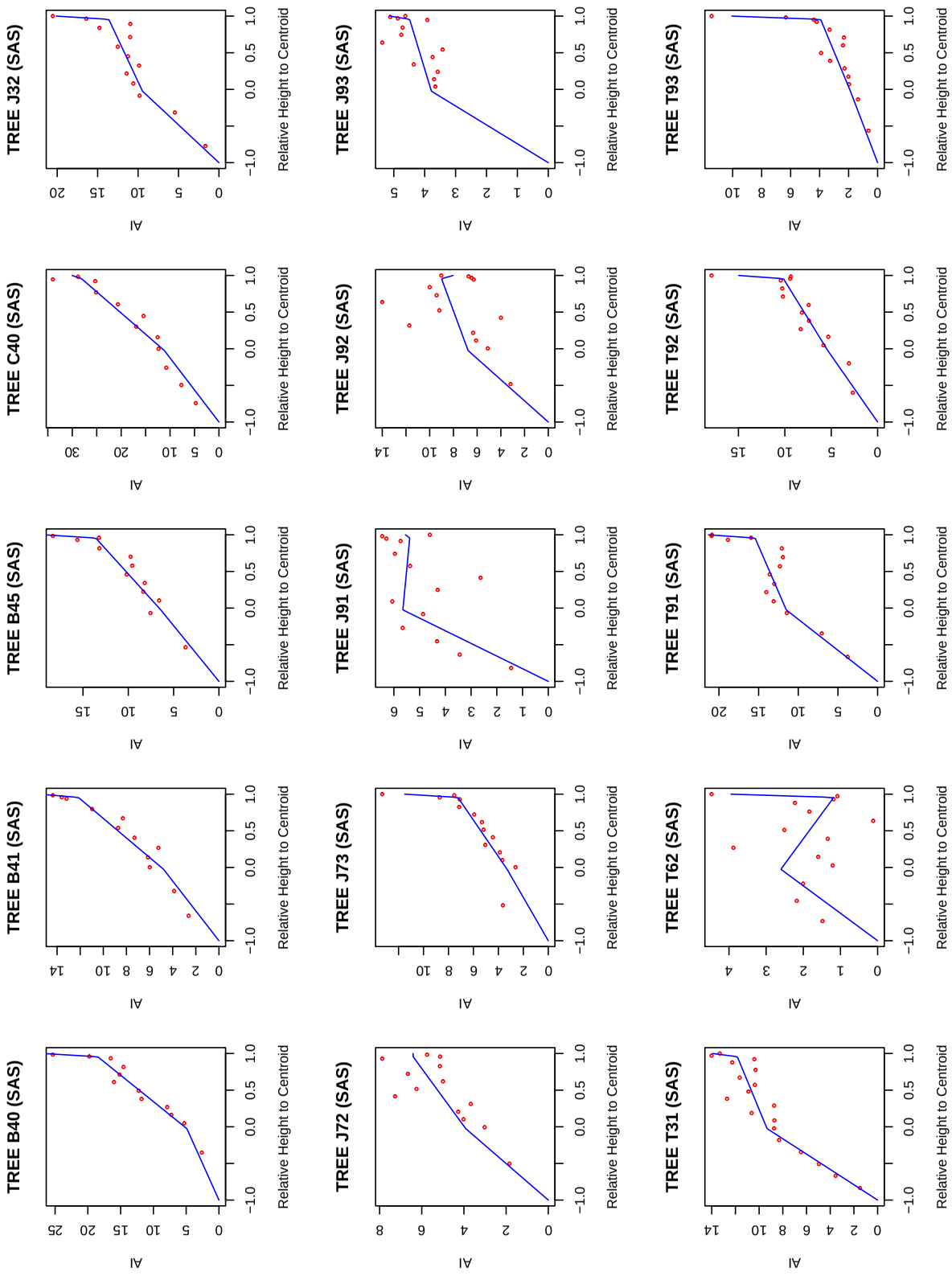

Figure A.3: Plots of the Estimated Mean Response of Each Tree from the 3-Segmented Regression Model

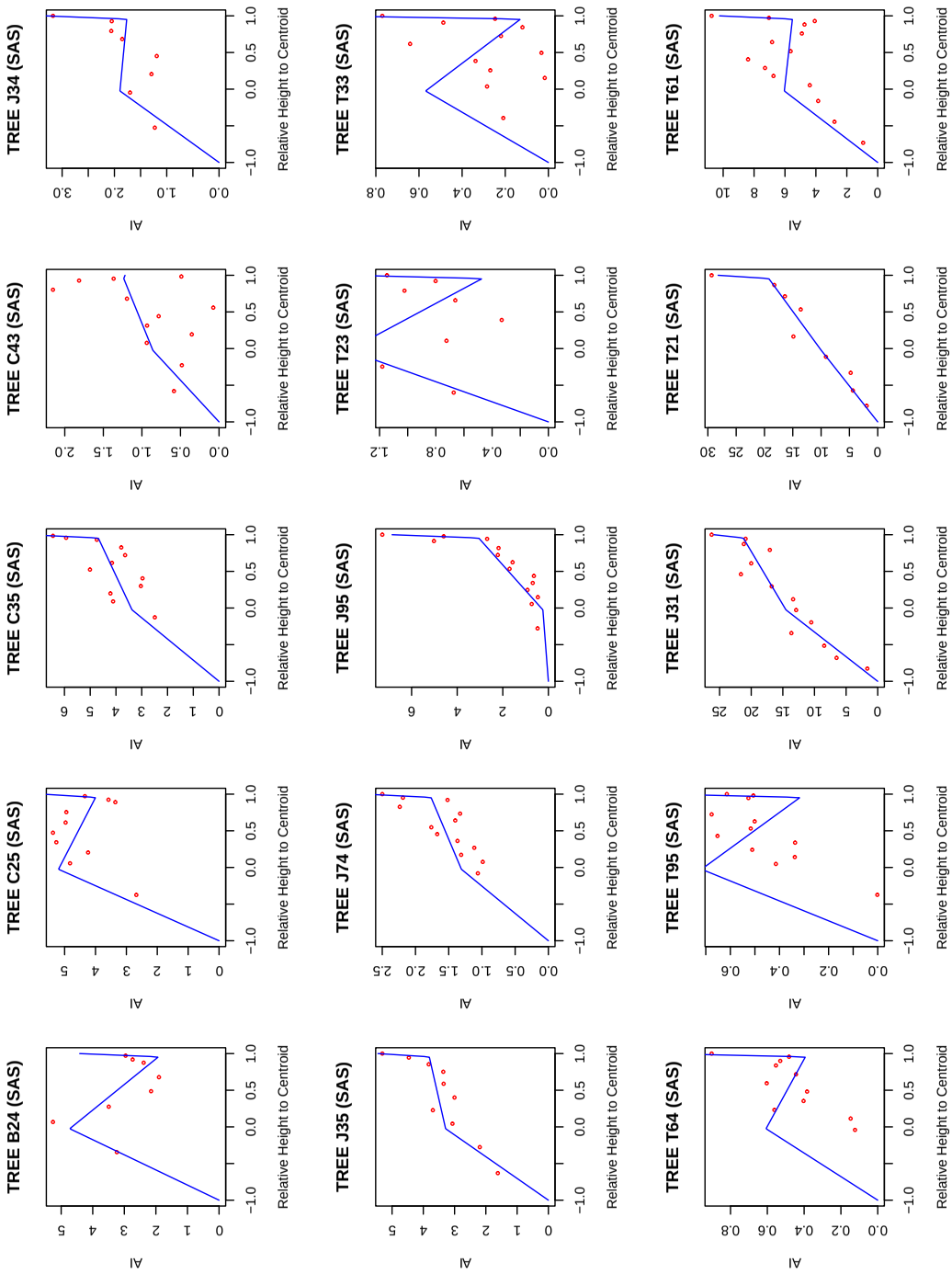

Figure A.4: Plots of the Estimated Mean Response of Each Tree from the 3-Segmented Regression Model

Appendix B

AIC and Covariates

In Section 3.2, we talked about using Jackknife AIC to look for the explanatory variables. Here we describe a simulation study using a linear regression model to show how this process of screening covariates may be useful.

Consider two predictors x and z with x randomly chosen from a uniform distribution $(-1, 1)$ and z randomly chosen from a uniform distribution $(0, m)$, where m takes values $3, 9, 15$, and 100. Our response $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{z} + \boldsymbol{\epsilon}$, where $\beta_0 = 0, \beta_1 = 1$, and $\beta_2 = 4$. The noise term, ϵ, in the linear model is distributed as a standard normal distribution. We regress y on x and evaluate whether z is a potential covariate for explaining the variation in this model. After fitting the model to the simulated data, the Jackknife bootstrap procedure, as described in Section 3.2, results in Jackknife AIC values displayed in Figures B.1 and B.2. The AIC, which is equivalent to the residual of sum squares, varies when we omit different observations. The Jackknife AIC can be shown analytically to be a quadratic function of z. Figure B.1 displays Jackknife AIC versus z for the different data sets corresponding to different m values. The larger value of m, the more contribution z makes in the linear model. Hence the clearer trends are exhibited for larger m. In comparison, Figure B.2 plots the Jackknife AIC versus x where no trends are observed.

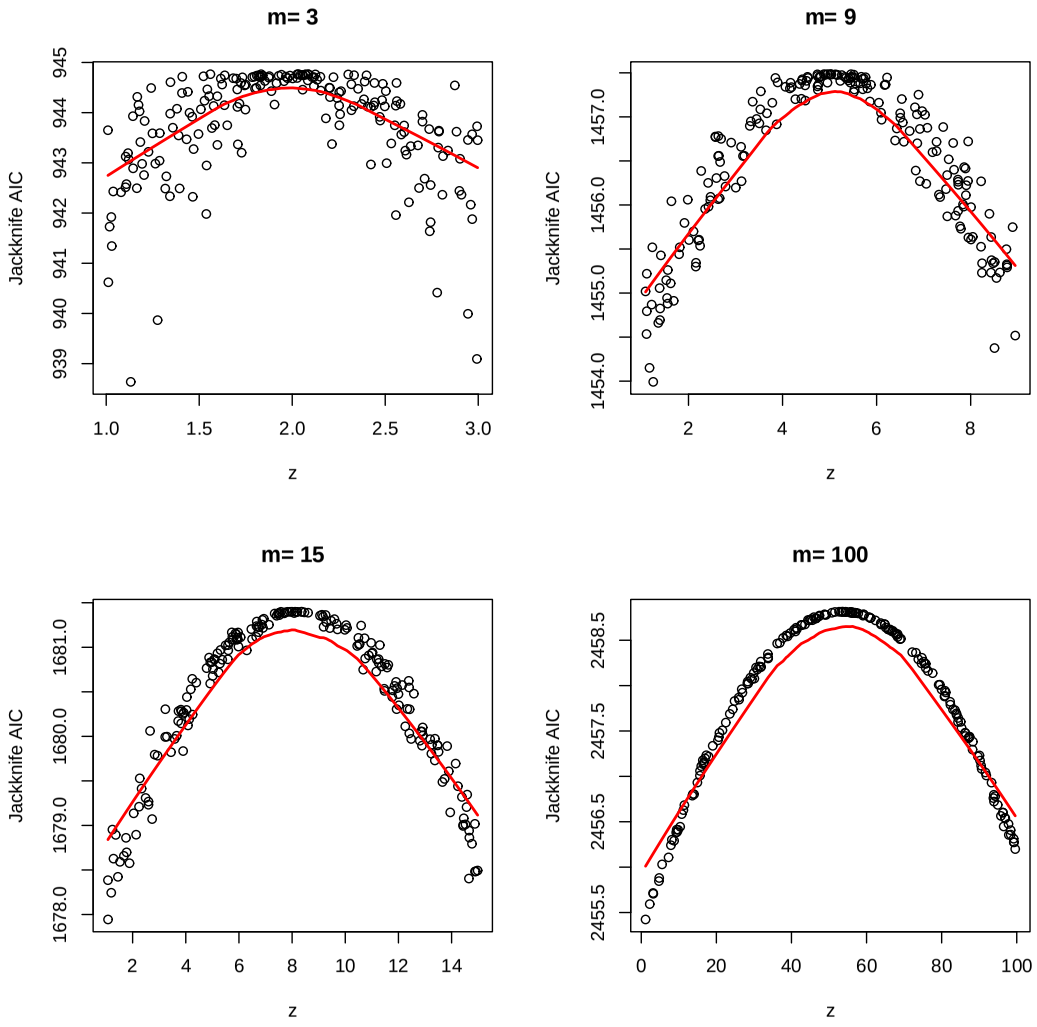

Figure B.1: Jackknife AIC versus z

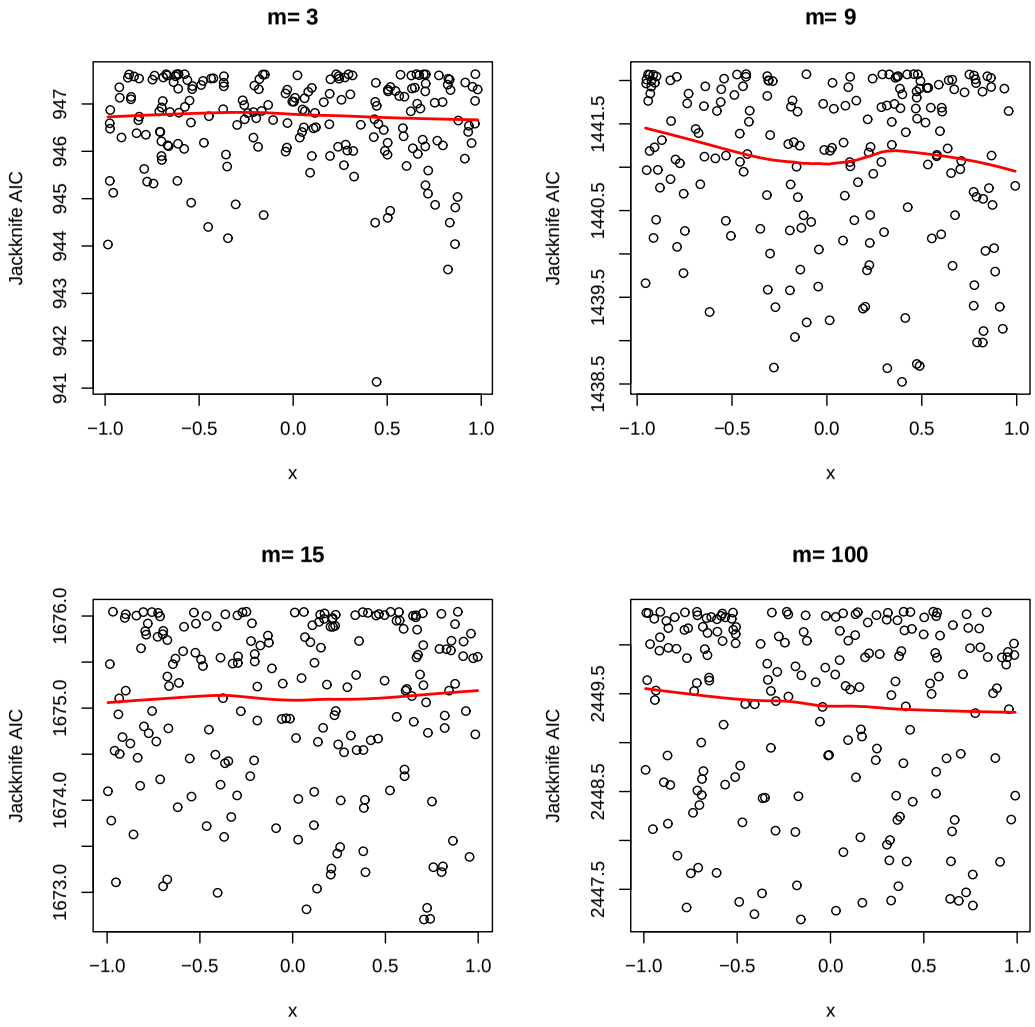

Figure B.2: Jackknife AIC versus x

Bibliography

[1] Akaike, H. (1974). A New Look at the Statistical Model Identification. IEEE Trans. Autom. Control, Volume AC-19, pp. 716-723.

[2] Abramowitz, M., Stegun, I.A. (1964). Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. New York: Dover.

[3] Casella, G., Berger, R. L. (2002). Statistical Inference. Australia ; Pacific Grove, CA.

[4] Davidian, M., Giltinan, D.M. (1995). Nonlinear Models for Repeated Measurement Data. Chapman & Hall.

[5] Efron, B., Tibshirani, R. (1994). An Introduction to the Bootstrap. Chapman & Hall /CRC.

[6] Goudie, J.W., Di Lucca, C.M. (2004). Modeling the Relationship between Crown Morphology and Wood Characteristics of Coastal Western Hemlock in British Columbia. In Fourth Workshop on the Connection between Silviculture and Wood Quality through Modeling Approaches and Simulation Software, edited by Gerard Nepveu. Equipe de Recherches sur la Qualite' des Bois, INRA, Nancy, France pp. 308-319.

[7] Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. J. Amer. Statist. Assoc., V 72, pp. 320-340.

[8] Laird, N.M., Ware, J.H. (1982). Random-Effects Models for Longitudinal Data. Biometrics 38, Dec., pp. 963-974.

[9] Leites, L.P., Robinson, A.P. (2004). Improving Taper Equations of Loblolly Pine with Crown Dimensions in a Mixed-effects Modeling Framework. Forest Science, V

50(2), April, pp. 204-212(9).

[10] Mansfield, S.D., Parish, R., Goudie, J.W., Kang, K. (In Preparation). The Effects of Crown Ratio on the Transition from Juvenile to Mature Wood Production in Lodgepole Pine in Western Canada.

[11] Pinheiro, J.C., Bates, D.M. (1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. Journal of Computation and Graphical Statistics, 4, pp. 12-35.

[12] SAS Institute Inc. (1999). http://v8doc.sas.com/sashtml/stat/chap46/index.htm

[13] Toms, J.D., Lesperance, M.L. (2003). Piecewise regression: a tool for identifying ecological thresholds. Ecology 84, pp 2034-2041.

[14] Therneau, T. M., Grambsch, P. M. (2001). Modeling Survival Data: Extending the Cox Model (Statistics for Biology and Health). Springer-Verlag, New York.

[15] Verbeke, G., Molenberghs, G. (2000). Linear mixed models for longitudinal data. Springer-Verlag, New York.

[16] Zobel, B.J., Van Buijtenen, J.P. (1989). Wood Variation: its Causes and Control. Springer-Verlag, New York.

[17] Zhang, Y., Tewarson, R.P. (1987). Least-Change Updates to Cholesky Factors Subject to the Nonlinear Quasi-Newton Condition, IMA Journal of Numerical Analysis 7, pp. 509-521.