# IMPUTATION BASED ON LOCAL LIKELIHOOD DENSITY ESTIMATION FOR INTERVAL CENSORED SURVIVAL DATA WITH APPLICATION TO TREE MORTALITY IN BRITISH COLUMBIA

by

Soyean Kim
Bachelor of Science, Simon Fraser University, 2006

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

SIMON FRASER UNIVERSITY
Spring 2009

# APPROVAL

**Name:** Soyean Kim

**Degree:** Master of Science

**Title of thesis:** Imputation based on local likelihood density estimation for interval censored survival data with application to tree mortality in British Columbia

**Examining Committee:** Dr. Derek Bingham
Chair

Dr. Charmaine Dean
Senior Supervisor
Simon Fraser University

Dr. Leilei Zeng
Supervisor
Simon Fraser University

Dr. Jiguo Cao
External Examiner

**Date Approved:**

# Abstract

Censored data arise in many situations including forestry and medical studies, and may take several forms. In this project, we consider imputation methods for estimating lifetimes when interval censored data are available. We investigate an imputation method based on local likelihood density estimation, where kernel smoothing is used to estimate the underlying distribution of lifetimes in order to calculate the conditional expectation of the observed lifetime. We contrast this with a simple midpoint estimator, where the imputed lifetime is the midpoint of the interval censored data. We compare the two imputation methods in the context of an analysis of tree mortality in British Columbia. The main goal of the project is to describe the relationships between tree lifetimes and important covariates such as thinning levels and species of trees while observing how the use of different imputation methods can affect the derived relationships. Additionally, we investigate the behaviour of the imputation schemes in simulation studies which vary the widths and sample size of the interval censored lifetimes.

# Acknowledgments

I would like to thank my supervisors, Dr. Charmaine Dean and Dr. Leilei Zeng for their encouragement and guidance.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Crown thinning is the selective removal of stems and branches to increase light penetration and air movement throughout the crown of a tree. It is a pruning technique primarily used on hardwood trees. The intent is to improve a tree's structure and form while making life uncomfortable for pests. The effect of crown thinning may depend on species of the tree.

In a designed experiment by the Ministry of Forests in British Columbia, three different levels of crown thinning were applied to trees in several plots of similar ages in order to assess the effect of thinning on tree mortality. The thinning treatment was applied about the time when crown thinning would begin to have an effect. At subsequent visits to the trees approximately 2-3 years apart, mortality status of the tree was recorded. There are 6 distinct sites in the experiment referred to here as *Installations* and 4 to 6 *Plots* within each site. Installations refer to the general location of the tree stand in British Columbia. The average ages at thinning are similar within the same Installation. Table A.1 in Appendix A displays the average age of trees at thinning by Installation and Plot.

The longitudinal measurements of the tree survival status are taken roughly at the same time within each Plot, but the measurement times for Plots in the same Installation are quite different. The first measurement time refers to the time at thinning. There were at most nine follow-up visits to trees; not all trees were examined at each measurement time and when trees died no subsequent visit was made to them. Missing data arise at random for live trees. Tables A.2 and A.3 in Appendix A summarize the total counts of trees and the counts of live trees at each measurement time after thinning by Plot within Installation. Those tables suggest that the proportion of missing data is small except for the last follow-up time. Table 1.1 displays the average number of follow-up visits and length of the time interval between visits by Installation. There are typically 5 follow-up visits at each Installation with the average time between visits ranging from about 2 to 4.25 years. Installations 3 and 5 had somewhat longer average time intervals between follow-up visits. Table 1.2 provides the number of dead trees at each measurement time by Installation. Note that Installation 3 has the largest number of dead trees at the first four follow-ups in part accounting for a lower number of average follow-up times (see Table 1.1).

Table A.4 in Appendix A lists the number of trees subjected to each level of the thinning treatment. The thinning treatment is based on a modified crown thinning that was conducted in the dormant season. Control trees had no thinning performed. Low Thinning refers to a thinning rate of 20%, which indicates that 20% of basal area was cut. High Thinning refers to a thinning rate of 35% of basal area being cut. Overall, there were 2195 Control trees, 1231 trees under Low Thinning, and the remaining 1149 trees were assigned High Thinning. Within each Installation, each Plot is assigned with one type of treatment with the exception of Plot 1 in Installation 2.

Let $t_{ij}$ be the age at measurement $j$ for tree $i$ and $D_{ij}$ be an indicator variable of whether tree $i$ is dead ($D_{ij}$=1) or alive ($D_{ij}$=0) at measurement $j$.

The lifetime interval ($L_i$, $R_i$) for tree $i$ is then:

$$L_i = max\{t_{ij}|D_{ij} = 0\}$$

$$R_i = min\{t_{ij}|D_{ij} = 1\}$$

For tree $i$, if $D_{ij} = 0$ at the last follow-up time (max $t_{ij}$ for tree $i$) then the interval is right censored and we let $R_i = \infty$.

We analyze the lifetime interval ($L_i$, $R_i$) since thinning defined as:

$$L_i = max\{t_{ij}|D_{ij} = 0\} - t_{i1}$$

$$R_i = min\{t_{ij}|D_{ij} = 1\} - t_{i1}$$

The data are interval censored or right censored observations where the right censored observations correspond to the lifetimes of trees that were still alive at the time of last measurement. Interval censored observations correspond to interval lifetimes for trees which died during the observation period. Table 1.3 displays the counts of the dead and live trees by Species, Installation and Treatment Levels. Douglas Fir and Western Cedar have a higher portion of right censored observations. By Treatment Level, both the Low and High Thinning groups have a larger number of right censored observations than the Control group; hence thinning reduces mortality.

Two types of imputation techniques were used to estimate the observed lifetimes for interval censored lifetimes: the interval midpoint and an imputation method based on local likelihood density estimation (Braun et al. 2005). The imputation method based on local likelihood density is useful for a one sample situation. Here, we have to account for covariate effects so we apply the method separately to each strata where a stratum is determined by each combination of tree species and treatment level. We also contrast this with a naive approach which applies the imputation method based on local likelihood density estimation without such stratification. Table 1.4 displays count of trees in each stratum. A large number of trees are Western Hemlock. We confine our analysis in this project to the main three species: Douglas Fir, Western Hemlock and Western Cedar, omitting data corresponding to trees labeled as ‘Other Species’.

In addition, we explored various scenarios for interval censored data in a simulation study where the aforementioned imputation techniques were applied. The performances of these imputation techniques were assessed under different scenarios with a varying sample size, interval width, and covariate effects.

Table 1.1: Average Number of Visits and Length of Interval between Visits by Installation

| Installation | Average Number<br>of Visits | Average Length of Interval<br>between Visits in Years |
|---:|:---|---:|
| 1 | 6.47 | 2.01 |
| 2 | 5.67 | 3.21 |
| 3 | 4.97 | 4.05 |
| 4 | 6.73 | 1.72 |
| 5 | 5.72 | 4.24 |
| 6 | 5.83 | 3.17 |

Table 1.2: Number of Dead Trees by Installation and Measurement Time (M.i refers to the (i)th follow-up time after thinning, i=1,..9)

| Installation | M.1 | M.2 | M.3 | M.4 | M.5 | M.6 | M.7 | M.8 | M.9 | Total in Sample |
|---:|:---|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 4 | 18 | 13 | 26 | 23 | 67 | 18 | 38 | | 582 |
| 2 | 30 | 139 | 99 | 69 | 104 | 78 | 51 | 46 | | 994 |
| 3 | 60 | 221 | 137 | 99 | 68 | 138 | 78 | | | 1276 |
| 4 | 3 | 15 | 18 | 23 | 36 | 44 | 21 | 36 | | 581 |
| 5 | 17 | 13 | 32 | 13 | 12 | 30 | 22 | | | 302 |
| 6 | | 15 | 11 | 13 | 20 | 23 | 11 | 44 | 62 | 840 |
| Grand Total | 114 | 421 | 310 | 243 | 263 | 380 | 201 | 164 | 62 | 4575 |

Table 1.3: Counts of Dead and Live Trees

| Group | Group level | Alive | Dead | % Dead | Total |
|---:|:---|---:|---:|---:|---:|
| By Species | Douglas Fir | 904 | 387 | 30% | 1291 |
| | Western Hemlock | 1173 | 1470 | 56% | 2643 |
| | Western Cedar | 324 | 104 | 24% | 428 |
| | Other Species | 18 | 47 | 72% | 65 |
| By Installation | 1 | 376 | 202 | 35% | 578 |
| | 2 | 379 | 585 | 61% | 964 |
| | 3 | 475 | 741 | 61% | 1216 |
| | 4 | 385 | 193 | 33% | 578 |
| | 5 | 163 | 122 | 43% | 285 |
| | 6 | 641 | 165 | 20% | 806 |
| By Treatment | Control | 968 | 1158 | 54% | 2126 |
| | Low Thinning | 722 | 473 | 40% | 1195 |
| | High Thinning | 729 | 377 | 34% | 1106 |

Table 1.4: Counts of Dead Trees by Each Combination of Species and Treatment

| Species | Treatment | Total Count of Trees |
|---:|:---|---:|
| Douglas Fir | Control | 189 |
| Douglas Fir | Low Thinning | 121 |
| Douglas Fir | High Thinning | 77 |
| Western Hemlock | Control | 860 |
| Western Hemlock | Low Thinning | 332 |
| Western Hemlock | High Thinning | 278 |
| Western Cedar | Control | 78 |
| Western Cedar | Low Thinning | 10 |
| Western Cedar | High Thinning | 16 |
| Other Species | Control | 31 |
| Other Species | Low Thinning | 10 |
| Other Species | High Thinning | 6 |

# Chapter 2

# Imputation Methodology

## Imputation based on local likelihood density estimation

Kernel density estimation is a nonparametric method for estimating a density function. It provides a simple way to find overall structure of data sets and requires no pre-specified functional form. Kernel estimators smooth out the contribution of each observed data point over a local neighborhood of that data point by using kernel weights, which depend on the proximity of an observation to the point of estimation. Applications of kernel smoothing are discussed in Wand (2006) and Ramsay and Silverman (2005). For example, Ramsay and Silverman (2005) mentions the use of kernel estimators as basis functions for fitting data. The imputation method based on local likelihood density estimation is based on an extension of kernel smoothing where the kernel weight is determined by the conditional expectation of the kernel over that interval. Braun et al. (2005) states that the main advantages of the method lies in its interpretive appeal as a kernel density estimate and that its iterative algorithm for solution provides a generalization of the self-consistency algorithms of Efron (1967), Turnbull (1976) and Li and Yu (1997). In addition, the iterative algorithm for the conditional expectation converges quickly to a unique solution and this is not dependent on initial values.

The idea of using nonparametric likelihood estimation for interval censored data is not new. Efron (1967) proposed an algorithm to obtain Kaplan and Meier (1958)'s nonparametric maximum likelihood estimator for survival function for right censored data and for more complicated censoring mechanisms such as interval censoring. Turnbull (1976) showed that when data are interval censored, the nonparametric likelihood estimator is defined up to an equivalence class of distributions over gaps called innermost intervals. Each of the Turnbull (1976)'s innermost interval is associated with a probability mass that needs to be located either to the right-hand, left-hand or mid point of the innermost interval. The selection of location of probability masses can be arbitrary and this leads to a maximum likelihood estimator that may not be unique. On the other hand, the iterative algorithm using the local likelihood density estimation offers a unique solution that converges quickly.

When data are interval censored, our goal here is to estimate the unobserved lifetime. We let $X_i$ be a lifetime that lies in an interval $I_i$=$(L_i, R_i]$ for a subject $i$. The estimated lifetime $\hat{X}_i$ can be calculated by taking the conditional expectation of $X_i$ given that $X_i$ lies in the interval $(L_i, R_i]$:

$$\hat{X}_i = E\{X_i | X_i \in (L_i, R_i]\} = \frac{\int_{L_i}^{R_i} x f(x) dx}{\int_{L_i}^{R_i} f(x) dx} \tag{2.1}$$

In order to estimate the expected value above, the underlying density $f(x)$ of the lifetimes needs to be estimated. Braun et al. (2005) proposes the following to estimate the underlying density, $f(x)$.

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} E\{K_h(X_i - x) | I_i\} \tag{2.2}$$

Equation 2.2 is the extension of the usual kernel density estimate for the conditional expectation where a lifetime, $X_i$, is not directly observed. $K(\cdot)$ is a symmetric probability density function with the bandwidth $h$ controlling the amount of smoothing; $x$ is the location of the kernel and $I_i$ indicates the interval where the lifetime $X_i$ lies. The conditional distribution itself is unknown. Solving Equation 2.2 involves the kernel density estimate itself and this results in a fixed point equation of $\hat{f}$ as follows:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{f}}\{K_h(X_i - x)|I_i\} \tag{2.3}$$

The function $f(x)$ is discretized so that it is effectively a vector forming a grid of points for $x$ and $\hat{f}(x)$, leading to fixed point iteration. At the $jth$ step of the iteration, the kernel density estimate is updated by:

$$\hat{f}_j(x) = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{f}_{j-1}}\{K_h(X_i - x)|I_i\} \tag{2.4}$$

The conditional expectation taken with $\hat{f}_{j-1}(t)$ is of the following form:

$$E_{\hat{f}_{j-1}}\{K_h(X_i - x)|I_i\} = \int_{I_i} K_h(X_i - x)\hat{f}_{j-1;i}(t)dt$$

The default initial value of $\hat{f}_0$ follows a uniform distribution unless specified otherwise. The conditional density over the $ith$ interval at the $jth$ step is:

$$\hat{f}_{j-1;i}(t) = 1(t \in I_i)\frac{\hat{f}_{j-1}(t)}{c_{j-1;i}} \tag{2.5}$$

where the normalizing constant for the conditional density over $ith$ interval at the $jth$ step is:

$$c_{j-1;i} = \int_{I_i} \hat{f}_{j-1}(t)dt$$

Since we have a grid of points for $x$ and the corresponding set of $\hat{f}(x)$ values that are discrete, the above expectation can be approximated by Riemann sum. Similarly, the expectation is taken for all intervals and the average value of all the expectations over all intervals is the estimated density $\hat{f}_j(x)$ at the $jth$ step.

The iterative algorithm to solve Equation 2.4 is implemented in the ICE package (Braun et al. 2005) in R. The output of the ICE package consists of a grid of points for $x$ and the corresponding probabilities. In this project, a particular value of $x$ represents a point in lifetime. The probabilities are to be estimated for 200 points of $x$'s in our study. These $x$ values along with the corresponding probabilities form a density estimate for the underlying density of tree lifetimes.

Once the density is estimated, the conditional expectation of a lifetime $X_i$ is calculated by the following Riemann sum:

$$\hat{X}_i = \frac{\sum_{i=1}^{N} x_i I_{\{L_i < x_i < R_i\}} \hat{f}(x_i)}{\sum_{i=1}^{N} I_{\{L_i < x_i < R_i\}} \hat{f}(x_i)}. \tag{2.6}$$

The convergence of the iteration algorithm above can be proven via the contraction mapping theorem (Ortega 1976).

The bandwidth $h$ can be estimated by the function 'dpik()' in R. The function utilizes the direct plug-in rules described in Wand (2006). Plug-in bandwidth selection is based on "Pluging in" estimates of the unknown quantities that appear in the formulae for the asymptotically optimal bandwidth. The asymptotically optimal bandwidth is derived from minimizing the asymptotic mean integrated error (AMISE). The mean integrated error criterion globally measures the distance between the kernel estimator and $f$. The selected bandwidth is inversely proportional to a quantity which is a measure of curvature of $f$. Thus, for a density with little curvature, little smoothing will be optimal. The gaussian kernel will be used in our study.

In the following sections, we employ local likelihood density estimation to smooth interval lifetimes of dead trees in order to build a density estimator and hence the imputed lifetimes. Here this is appropriate since all the live trees are right censored at approximately the same end of follow-up time, to the right of the interval lifetimes. Since the smoothing approach relies on local smoothing, lifetimes far away to the right will not affect the imputation.

# Chapter 3

# Analysis of tree mortality data

In this section, we use imputation methods for the analysis of the interval censored tree mortality data. The imputation methods employed are: (i) midpoint (MI), (ii) local likelihood density imputation applied to the data as a whole (LDI), (iii) local likelihood density imputation within strata defined by species and thinning levels (SLDI).

Figure 3.1 displays the distributions of the imputed lifetimes. The average imputed lifetimes using the three imputation methods are : MI-10.34 years (sd 7.62), LDI- 10.20 years (sd 7.40) and SLDI - 10.16 years (sd 7.49). Though, there are some slight differences among the imputed lifetimes from the three methods, overall there is generally considerable agreement.

Such close similarity among the imputed data sets may be due to interval lengths being relatively small here; we explore this in Chapter 4. Figure 3.2 shows the distribution of the interval lengths for the 2008 dead trees. It suggests that among 2008 dead trees, 63% of the interval lifetimes have the lengths less than 3 years.

Figure 3.3 displays boxplots of imputed lifetimes by Species and Treatment. The first row of the plot includes the boxplots by species using the three methods. The second row displays the boxplots in the first row arranged by Type of Species. Row 3 shows the boxplots by Treatment and finally, the fourth row displays the boxplots organized by level

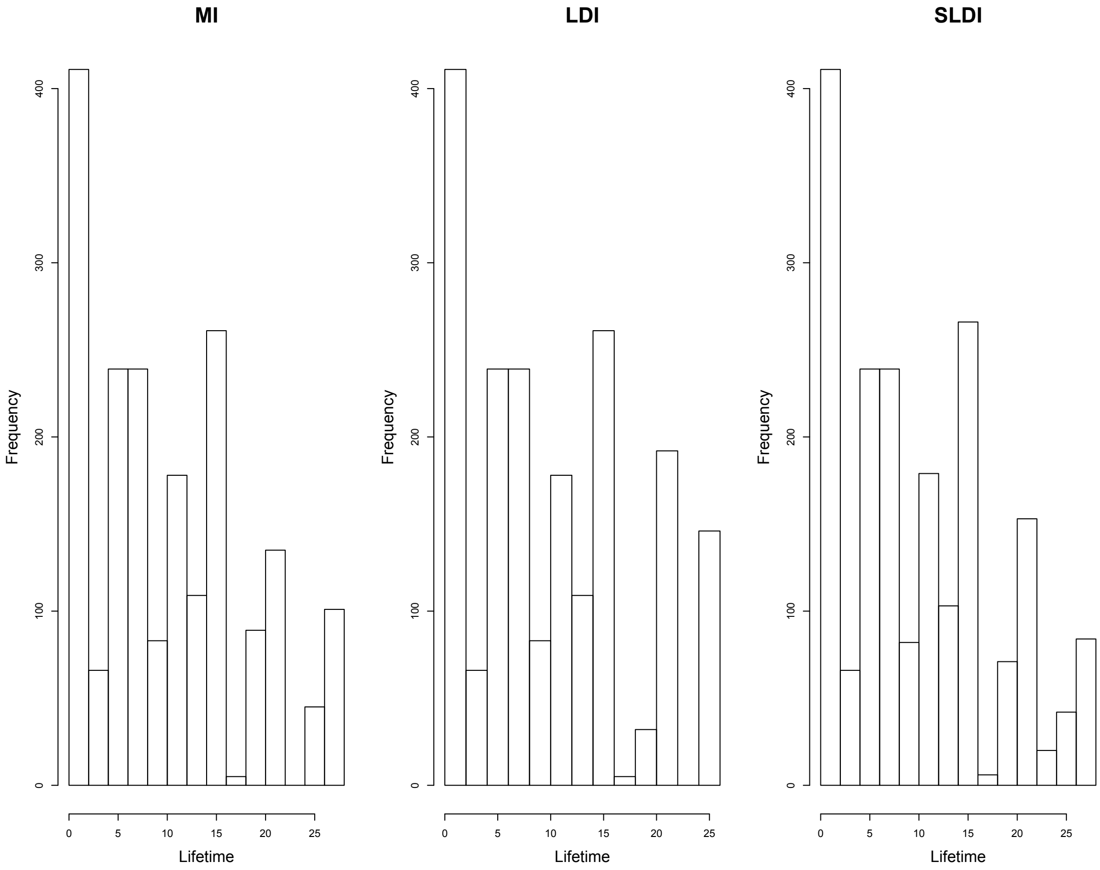


Figure 3.1: Histograms of Imputed Tree Lifetimes by the Imputation Methods

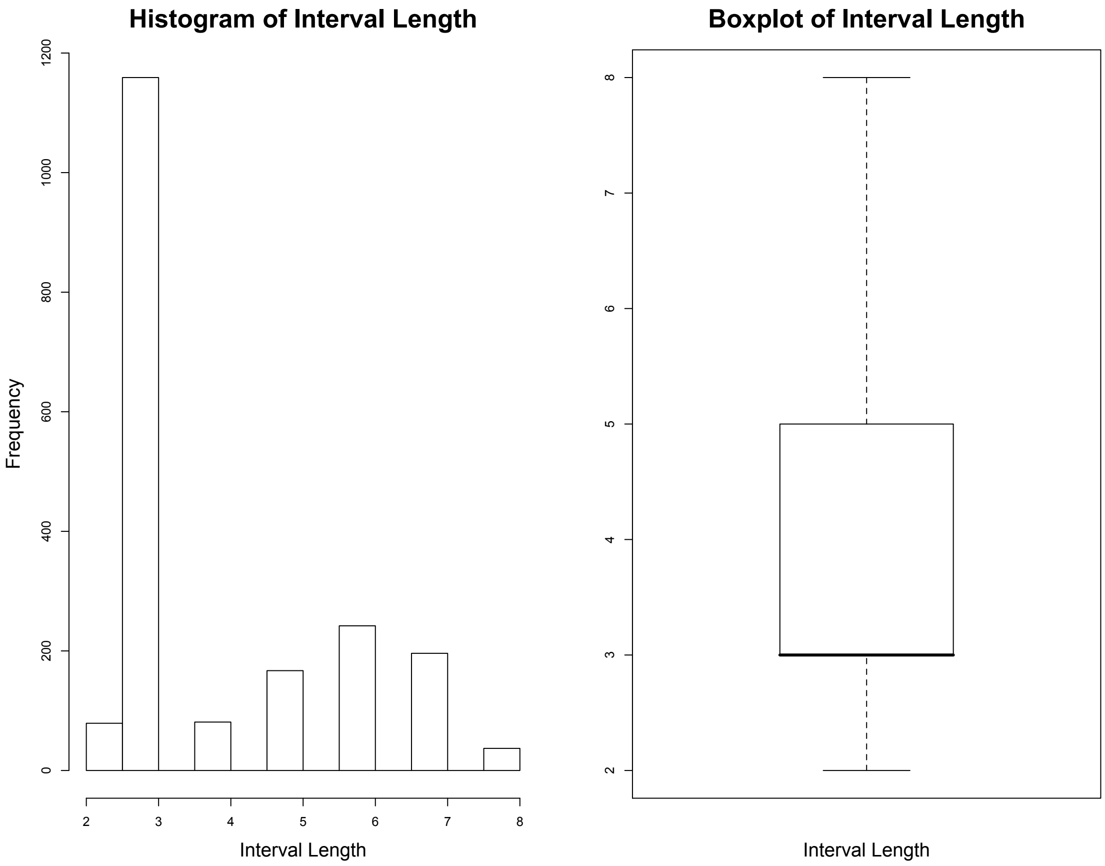


Figure 3.2: Distribution of the Interval Lengths in the Tree Mortality data

of Treatment. Figure 3.3 suggests that there seems to be effects due to both Species and Treatment. The three imputation methods produced very similar results. The imputed values using the midpoint method shows more variation than those from the other two methods.

Figure 3.4 displays Kaplan Meyer survival curves by the three imputation methods. About 40 % of trees die around 15 years after thinning . The Kaplan Meyer curves using all three imputation methods are almost identical.

Figure 3.5 displays Kaplan Meyer survival curves by Species and Treatment using the three methods. The first row displays the Kaplan Meyer curves by Species using the three methods. In the second row, the survival curves in Row 1 are organized by type of Species showing Douglas Fir, Western Hemlock and Western Cedar from left to right. Row 3 displays the Kaplan Meyer curves by Treatment using the three methods. Finally, Row 4 displays the survival curves organized by Level of Treatment showing the Control group, Low Thinning and High Thinning from left to right. Figure 3.5 suggests that survival differs amongst the three species. Western Hemlock seems to have higher mortality while Douglas Fir and Western Cedar seem to have similar survival experience. Figure 3.5 also suggests that there exist treatment effects with higher level of thinning yielding lower mortality.

Figure 3.6 displays the estimated underlying density of the tree lifetimes using LDI. The histogram on the right hand side next to the estimated density displays the distribution of the midpoints of the interval lifetimes. It shows a similar pattern to the estimated density using LDI. The estimated density is heavily skewed to the right with the highest peak occurring around 2 years. This indicates most of the trees that died before the end of follow-up have lifetimes around 2 years since the time of thinning.

Figure 3.7 displays the histograms of midpoints of the interval lifetimes by strata as determined by each combination of Species and Treatment. The underlying densities among the three species look quite different as shown in the histograms. The distributions seem somewhat different across Species and across different Levels of treatment. Figure 3.8

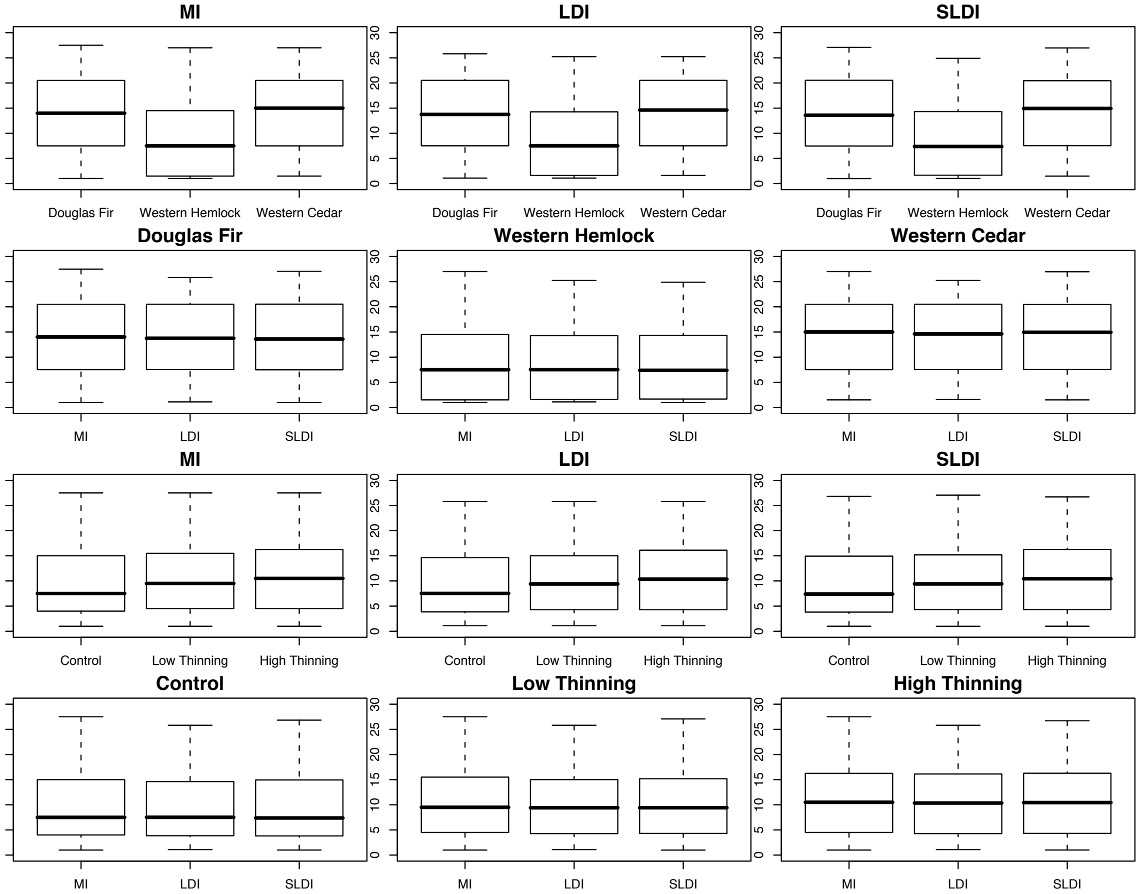


Figure 3.3: Boxplots of Imputed Tree Lifetimes by Species and Treatment

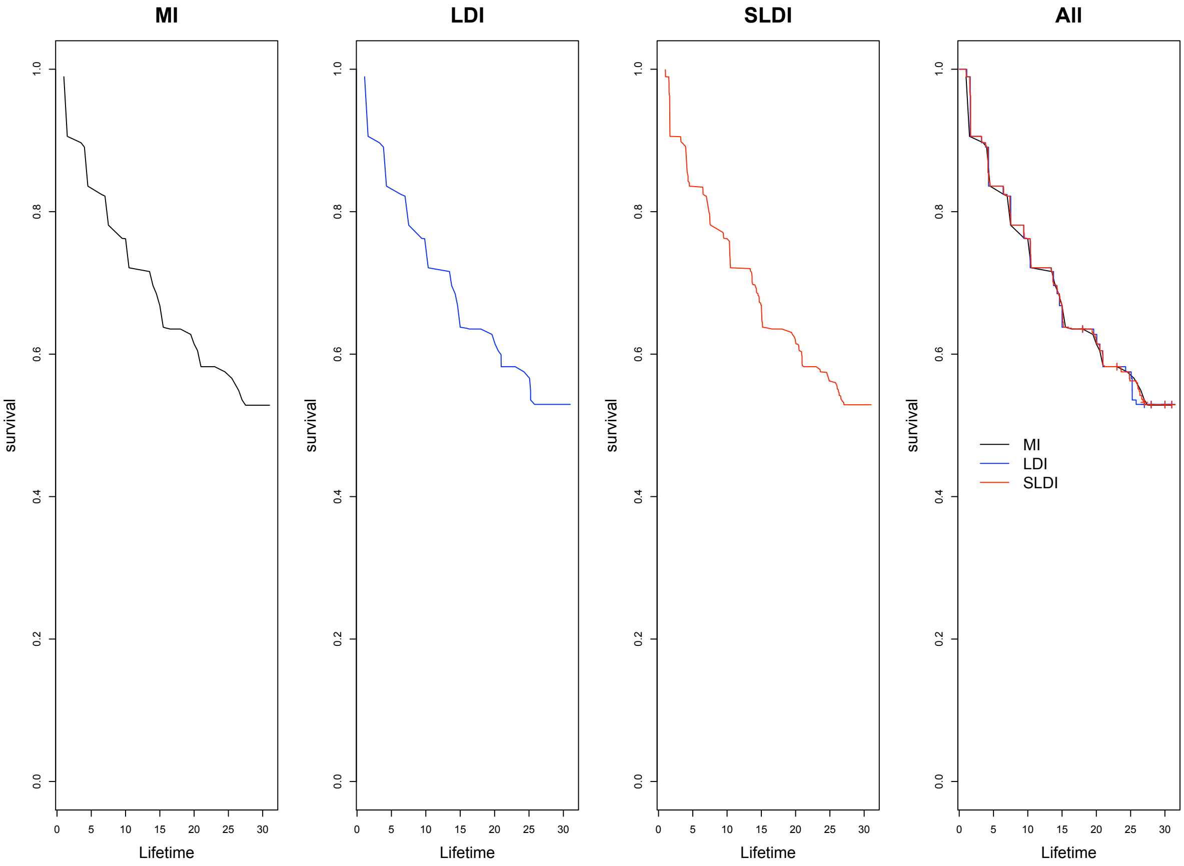


Figure 3.4: Kaplan Meyer Survival Curves by Three Imputation Methods

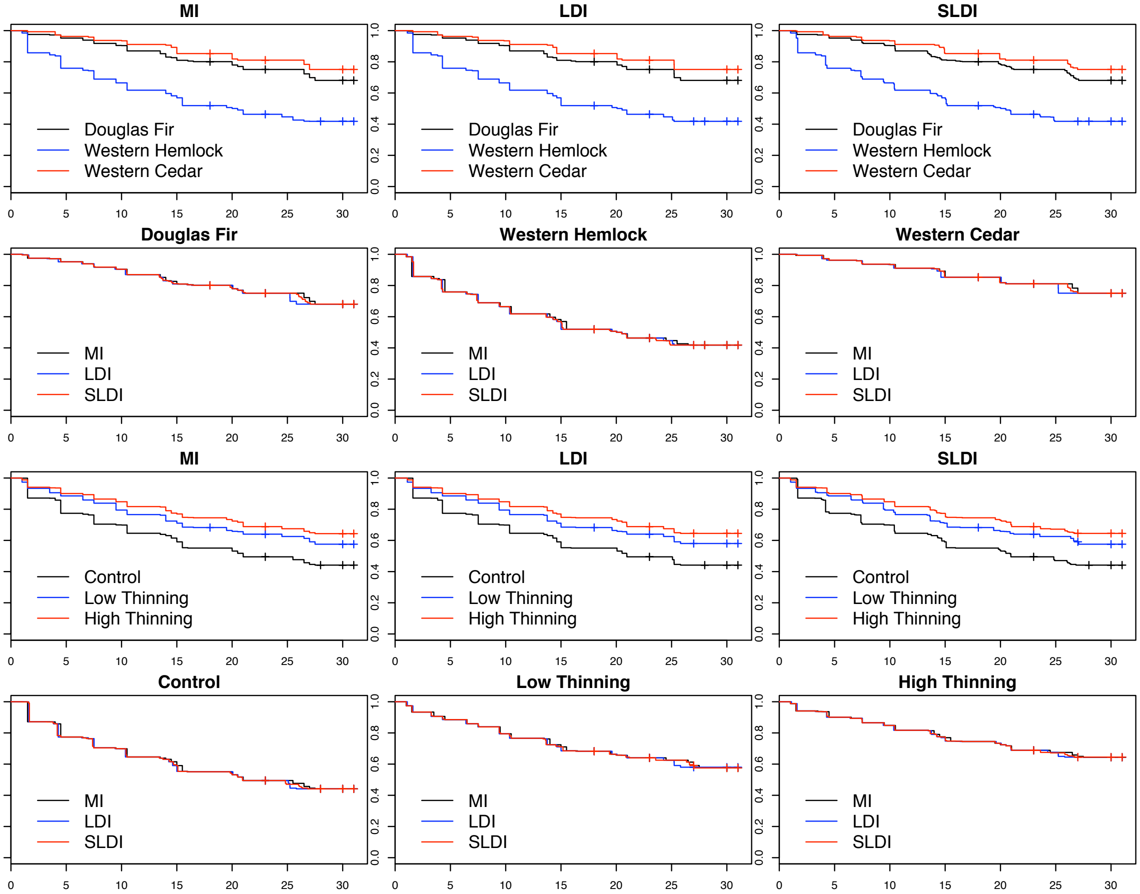


Figure 3.5: Kaplan Meyer Survival Curves by Species and Treatment

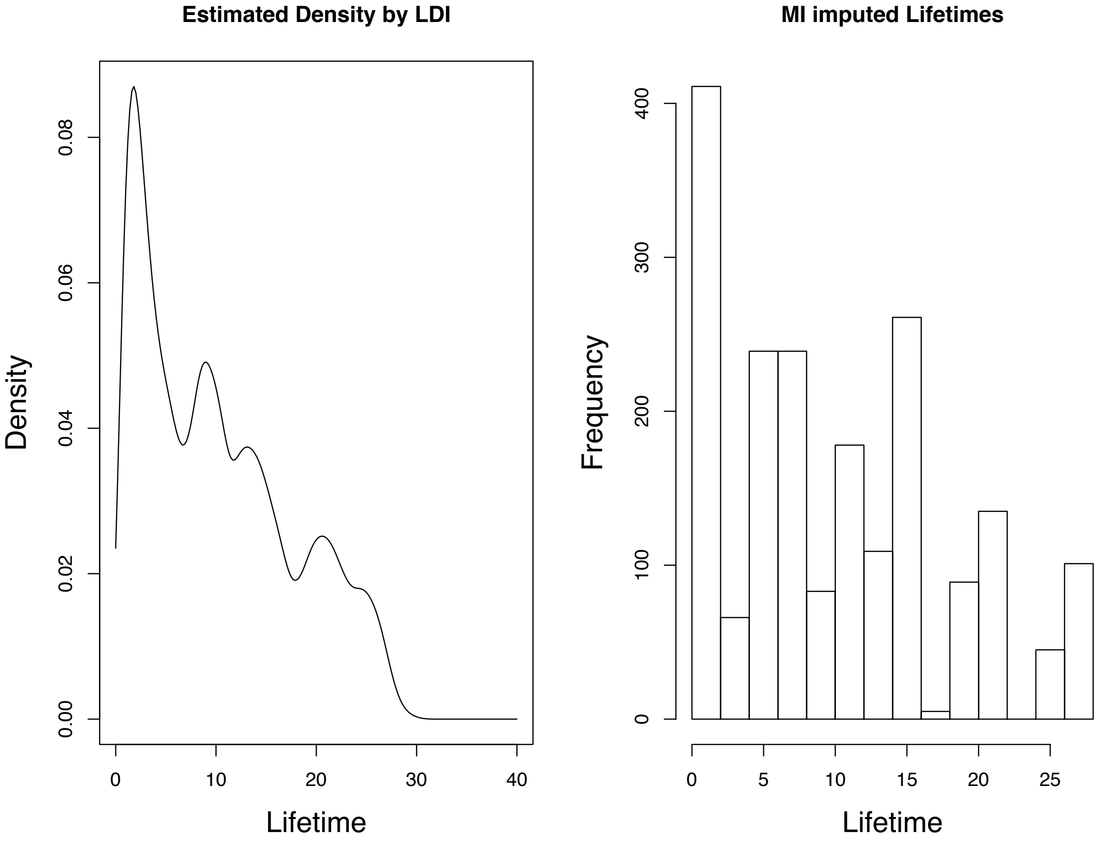


Figure 3.6: Estimated Underlying Distribution of Tree Lifetimes by LDI method and the Histogram of the Midpoints of the Interval Lifetimes

displays the estimated underlying density for each stratum using SLDI. The estimated densities vary by each combination of Species and Treatment Levels. It should be noted that sample size matters when estimating a density, and that the number of lifetime intervals used in the estimation of underlying density for each stratum were different. Strata for Western Hemlock have the largest sample sizes, especially the Western Hemlock and Control combination (860 lifetime intervals); their density estimates are shown in Row 2. Strata for Western Cedar have the smallest sample sizes and their density estimates are shown in Row 3.

Figure 3.9 shows the absolute differences in imputed lifetimes from the use of the three imputation methods. Most of the differences are quite small (less than 0.5 years). Some large differences occur between the estimates obtained from MI and LDI. The average difference between MI and LDI was 0.24 and the average difference between MI and SLDI was 0.19. Similarly, the average difference between LDI and SLDI was 0.14. This suggests that stratification when implementing LDI to incorporate the covariate effects has an impact. The larger differences among the three methods may be linked to the interval lengths in the tree mortality data. Figure 3.10 displays the absolute differences by interval length. All three plots show an increasing pattern in the differences as interval length increases.

Regression analysis was performed using the SURVREG procedure in R. The SURVREG procedure fits parametric accelerated failure time models to survival data that may be left, right, or interval censored. The parametric model is of the form

$$\log(T) = y = x'\beta + \sigma\epsilon$$

where $y$ is usually and is here the log of the failure time variable, $x$ is a vector of covariate values, $\beta$ is a vector of unknown regression parameters, $\sigma$ is an unknown scale parameter, and $\epsilon$ is an error term; $y$ can be specified as Weibull or Exponential distributions. For the Weibull model, note that the survival function is

$$S(t) = Pr(T \geqslant t) = \exp(-\exp(\frac{y - x'\beta}{\sigma}))$$

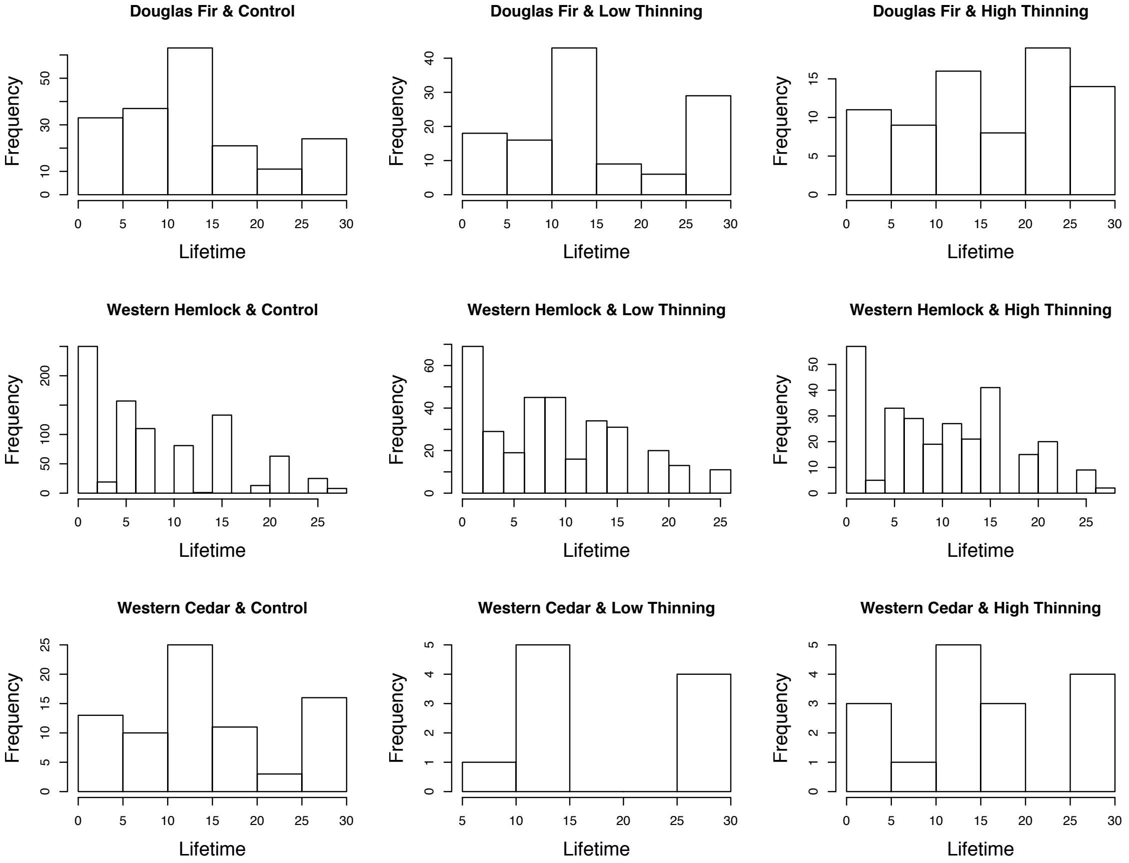


Figure 3.7: Histograms of the Midpoints of the Interval Lifetimes by Strata

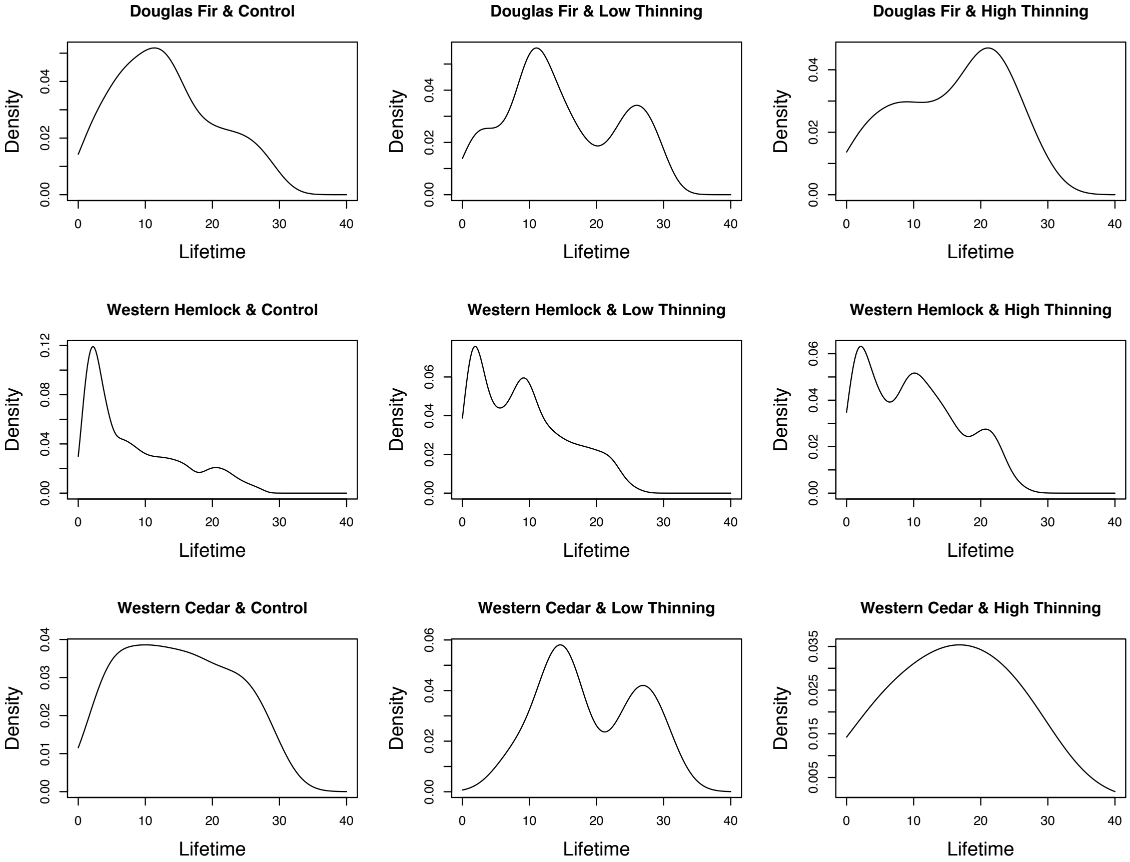
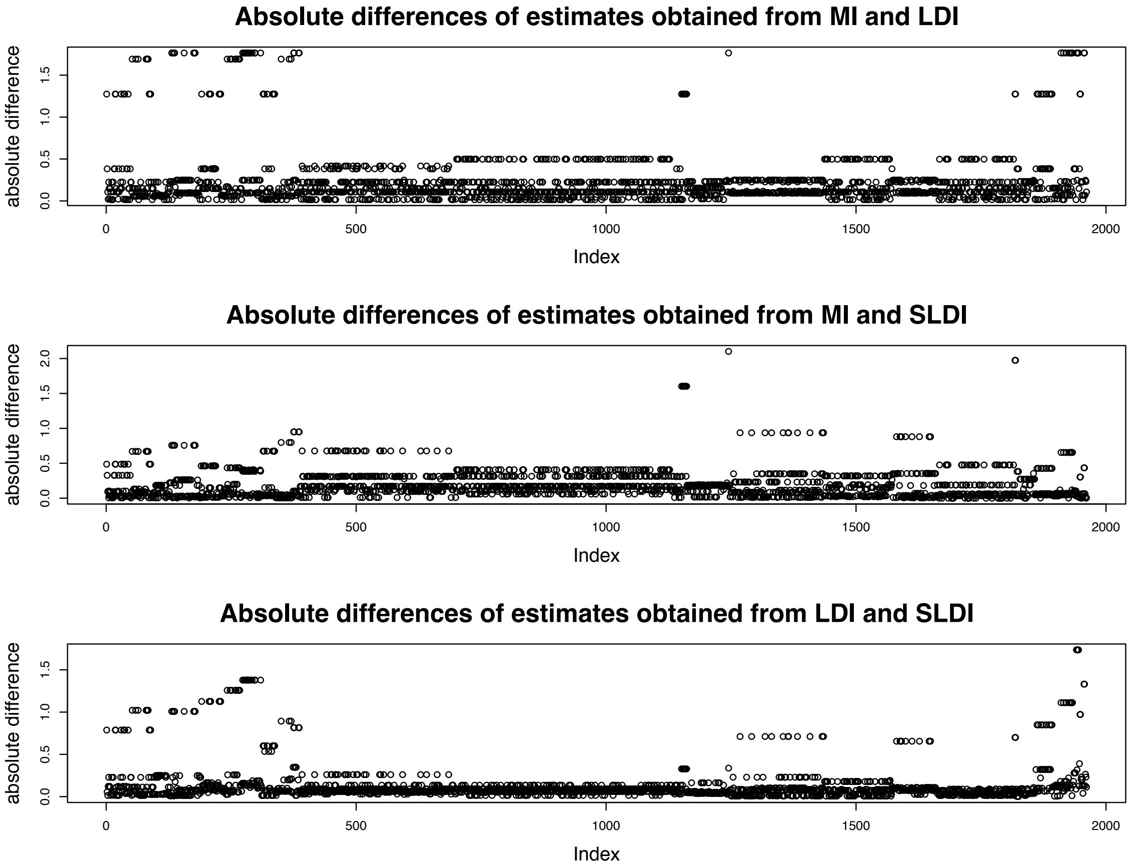


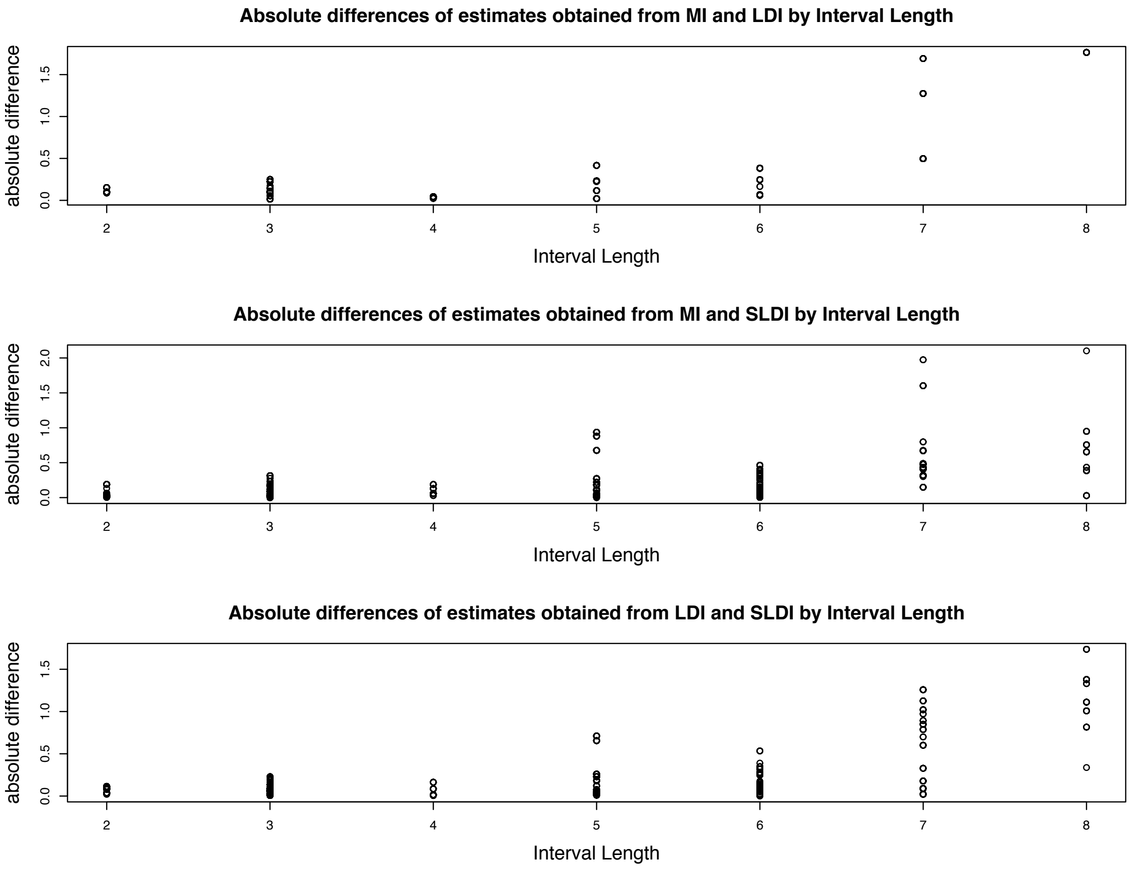
Figure 3.8: Estimated Underlying Distributions of Tree Lifetimes by Strata


Absolute differences of estimates obtained from MI and LDI

Absolute differences of estimates obtained from MI and SLDI

Absolute differences of estimates obtained from LDI and SLDI

Figure 3.9: Absolute Differences of Estimates obtained from the Three Imputation Methods



Absolute differences of estimates obtained from MI and LDI by Interval Length

Absolute differences of estimates obtained from MI and SLDI by Interval Length

Absolute differences of estimates obtained from LDI and SLDI by Interval Length

Figure 3.10: Differences of Estimates obtained from the Three Imputation Methods plotted against Interval Length

. See Kalbfleisch and Prentice (1980) and Elandt-Johnson and Johnson (1980) for more details.

The Weibull model is a proportional hazard model with the logarithm of the hazard being of the form $h(t) = h_0(t)e^{-x'\beta}$. Regression coefficient estimates from the COXPH procedure are expected to be of the same magnitude but opposite in sign to those from fitting the Weibull model using the SURVREG procedure.

The Weibull and Cox proportional hazard models were fit to the three sets of imputed values produced by the three imputation methods. Species, Treatment and their interaction were included in the models as covariates. Table 3.1 displays the estimated covariate effects and their standard errors from fitting the Weibull model to the three imputed data sets and from using a full likelihood approach based directly on the interval and right censored lifetimes. For the full likelihood approach using a Weibull model, note that the likelihood function becomes:

$$\prod_{i=1}^{n}[S(L_i) - S(R_i)]^{Z_i}[S(L_i)]^{1-Z_i}$$

where $Z_i$ is an indicator variable for an observation being interval censored. The partial likelihood function for the proportional hazards model is:

$$\prod_{i=1}^{n}\{\frac{\exp(-X_i'\beta)}{\sum_{L\in R_i}\exp(-X_i'\beta)}\}^{Z_i}$$

where $R_i$ is the risk set corresponding to the imputed lifetime $t_i$. The risk set is the set of all trees alive and uncensored at $t_i$.

The estimated effects are relative to reference categories; the reference categories for Species and Treatment are Douglas Fir and Control accordingly. Both the estimated effects and the standard errors are similar using all three imputation methods and they are quite comparable with corresponding values from a full likelihood analysis. Table 3.2 displays the estimated effects using the proportional hazard model. The estimated effects and their standard deviations show consistent results and are similar in magnitude as those from the Weibull model. In the following discussion, estimates are based on those derived from

SLDI imputation and the Cox model.

*Western Hemlock relative to Douglas Fir* the relative risk of mortality is larger by a factor of 2.66 when no thinning is applied. With Low Thinning, the relative risk of mortality is larger by a factor of 2.25. With High Thinning, the relative risk of mortality is larger by a factor of 2.59. However, neither treatment effect is significant. The relative risk of mortality of Western Hemlock relative to Douglas Fir seems to be larger regardless of the Level of Treatment and thinning treatment does not significantly affect the relative risk of mortality.

*Western Cedar relative to Douglas Fir* the relative risk of mortality is not significantly different from that of Douglas Fir when there is no thinning applied (the relative risk is close to 1). The relative risk of mortality is smaller by a factor of 0.28 when Low Thinning is applied. With High Thinning, the relative risk is smaller by 0.58. The relative risk of mortality is minimized when Low Thinning is applied.

*Low Thinning relative to the Control group* the relative risk of mortality is smaller by a factor of 0.81 when the treatment is applied to Douglas Fir. The relative risk is smaller by a factor of 0.68 when the thinning was applied to Western Hemlock. The relative risk is smaller by a factor of 0.22 with Western Cedar. The relative risk of mortality of Low Thinning group relative to Control group is minimized with Western Cedar.

*High Thinning relative to the Control group* the relative risk of mortality is smaller by a factor of 0.57 when the treatment is subjected to Douglas Fir. The relative risk is smaller by a factor of 0.55 when high thinning is applied to Douglas Fir. The relative risk is smaller by a factor of 0.32 when high thinning is applied to Western Cedar. The relative risk of mortality of High Thinning group is minimized with Western Cedar.

More thinning seems to improve the chance of survival in trees by minimizing the relative risk of mortality except for Western Cedar where the relative risk is minimized with Low Thinning. The effectiveness of thinning depends on Species as the interaction between Types of Species and the Level of Treatment affects the relative risk of mortality in trees. Interpretation is similar using other imputation methods.

Figure 3.11 displays diagnostic plots to check for the assumption of the proportional hazards model. The first column displays, by Species, $\log(-\log \tilde{S}(t))$ versus $\log(t)$ where $\tilde{S}(t)$ is the Kaplan Meyer survival function. The second column displays $\log(-\log \tilde{S}(t))$ versus $\log(t)$ by the Level of Treatment. Imputed values obtained from MI are displayed in the first row, those from LDI are displayed in the second row, and the last row displays the imputed values from SLDI. As the parallel curves would suggest that the assumption of proportional hazard is met, we conclude that there is no striking evidence of departures from the proportional hazard assumption here.

In order to assess the adequacy of the Weibull model, a residual analysis was performed. If a lifetime $T_i$ has a survival function $S(t; x_i, \beta)$, then the residual defined as $-\log(S(t; x_i, \beta))$ has a unit exponential distribution. Let $\hat{e}_i$=$-\log(S(t; x_i, \hat{\beta}))$ for lifetimes and $-\log(S(t; x_i, \hat{\beta})) + 1$ for censored times. Then, $\hat{e}_i$ estimates the residuals $e_i$ and $\hat{e}_i$ should behave approximately like a unit exponential. Thus, we plot ordered residuals $\hat{e}_i$ versus the expected exponential order statistics i.e Savage scores; the values of the expected exponential order statistics are $\sum_{r=1}^{i}(n - r + 1)^{-1}$. The plot should be roughly a straight line when our original Weibull model is adequate (Lawless 2003). In addtion, we can treat the residuals as a set of possible censored observations and derive their Kaplan Meyer estimates $S^*(t)$. A plot of $-\log(S^*(t))$ versus $\log(t)$ should be roughly a straight line when the original model is adequate.

Figure 3.12 displays such residual plots to assess the adequacy of the Weibull model. The first column plots the ordered residuals versus expected exponential order statistics using the imputed values from the three imputation methods. The second column plots the ordered residuals versus their Kaplan Meyer estimates using the imputed values from the three imputation methods. Both plots show departures from linearity in their tails. The deviation from linearity may be due to the heavy amount censoring present in the tree mortality data (over 55% censoring). With such a large amount of censoring, the usefulness

Table 3.1: Estimated Effects and SE by the Three Imputed Methods and Full Likelihood Approach using the Weibull Model.

| Imputation Method | Variable | Coefficient ($\hat{\beta}$) | SE | Relative Risk ($e^{\hat{\beta}}$) |
|---|---|---|---|---|
| MI | Western Hemlock | -1.12 | 0.09 | 0.33 |
| | Western Cedar | -0.01 | 0.15 | 0.99 |
| | Low Thinning | 0.22 | 0.13 | 1.24 |
| | High Thinning | 0.61 | 0.15 | 1.84 |
| | Western Hemlock*Low Thinning | 0.21 | 0.14 | 1.23 |
| | Western Cedar*Low Thinning | 1.41 | 0.39 | 4.10 |
| | Western Hemlock*High Thinning | 0.05 | 0.16 | 1.05 |
| | Western Cedar*High Thinning | 0.63 | 0.33 | 1.88 |
| LDI | Western Hemlock | -1.12 | 0.09 | 0.33 |
| | Western Cedar | -0.02 | 0.15 | 0.98 |
| | Low Thinning | 0.21 | 0.13 | 1.23 |
| | High Thinning | 0.61 | 0.15 | 1.84 |
| | Western Hemlock*Low Thinning | 0.21 | 0.14 | 1.23 |
| | Western Cedar*Low Thinning | 1.41 | 0.39 | 4.10 |
| | Western Hemlock*High Thinning | 0.05 | 0.16 | 1.05 |
| | Western Cedar*High Thinning | 0.63 | 0.33 | 1.88 |
| SLDI | Western Hemlock | -1.12 | 0.09 | 0.33 |
| | Western Cedar | -0.01 | 0.15 | 0.99 |
| | Low Thinning | 0.22 | 0.13 | 1.24 |
| | High Thinning | 0.61 | 0.15 | 1.84 |
| | Western Hemlock*Low Thinning | 0.21 | 0.14 | 1.23 |
| | Western Cedar*Low Thinning | 1.40 | 0.39 | 4.06 |
| | Western Hemlock*High Thinning | 0.05 | 0.16 | 1.05 |
| | Western Cedar*High Thinning | 0.63 | 0.33 | 1.88 |
| Full Likelihood | Western Hemlock | -1.17 | 0.09 | 0.31 |
| | Western Cedar | -0.02 | 0.15 | 0.98 |
| | Low Thinning | 0.23 | 0.13 | 1.26 |
| | High Thinning | 0.64 | 0.16 | 1.90 |
| | Western Hemlock*Low Thinning | 0.22 | 0.15 | 1.25 |
| | Western Cedar*Low Thinning | 1.49 | 0.41 | 4.44 |
| | Western Hemlock*High Thinning | 0.05 | 0.17 | 1.05 |
| | Western Cedar*High Thinning | 0.66 | 0.35 | 1.93 |

Table 3.2: Estimated Effects and SE by the Three Imputed Methods using the Proportional Hazard Model.

| Imputation Method | Variable | Coefficient ($\hat{\beta}$) | SE | Relative Risk ($e^{\hat{\beta}}$) |
|---|---|---|---|---|
| MI | Western Hemlock | 0.98 | 0.08 | 2.66 |
| | Western Cedar | 0.02 | 0.13 | 1.02 |
| | Low Thinning | -0.20 | 0.12 | 0.82 |
| | High Thinning | -0.56 | 0.14 | 0.57 |
| | Western Hemlock*Low Thinning | -0.18 | 0.13 | 0.84 |
| | Western Cedar*Low Thinning | -1.29 | 0.36 | 0.28 |
| | Western Hemlock*High Thinning | -0.03 | 0.15 | 0.97 |
| | Western Cedar*High Thinning | -0.57 | 0.31 | 0.57 |
| LDI | Western Hemlock | 0.97 | 0.08 | 2.64 |
| | Western Cedar | 0.02 | 0.13 | 1.02 |
| | Low Thinning | -0.20 | 0.12 | 0.82 |
| | High Thinning | -0.56 | 0.14 | 0.57 |
| | Western Hemlock*Low Thinning | -0.18 | 0.13 | 0.84 |
| | Western Cedar*Low Thinning | -1.29 | 0.36 | 0.28 |
| | Western Hemlock*High Thinning | -0.04 | 0.15 | 0.96 |
| | Western Cedar*High Thinning | -0.57 | 0.31 | 0.57 |
| SLDI | Western Hemlock | 0.98 | 0.08 | 2.66 |
| | Western Cedar | 0.02 | 0.13 | 1.02 |
| | Low Thinning | -0.21 | 0.12 | 0.81 |
| | High Thinning | -0.56 | 0.14 | 0.57 |
| | Western Hemlock*Low Thinning | -0.17 | 0.13 | 0.84 |
| | Western Cedar*Low Thinning | -1.28 | 0.36 | 0.28 |
| | Western Hemlock*High Thinning | -0.03 | 0.15 | 0.97 |
| | Western Cedar*High Thinning | -0.56 | 0.31 | 0.57 |

of such residual analysis is limited. Figure A.1 in Appendix A displays examples of residual analysis based on simulated data sets with various amount of censoring. As the amount of censoring increases, deviation from linearity becomes more obvious. We acknowledge that the residual plots suggest some concern for the lack of fit of the Weibull model. Nevertheless, the imputation methods produced very similar results under all analysis.

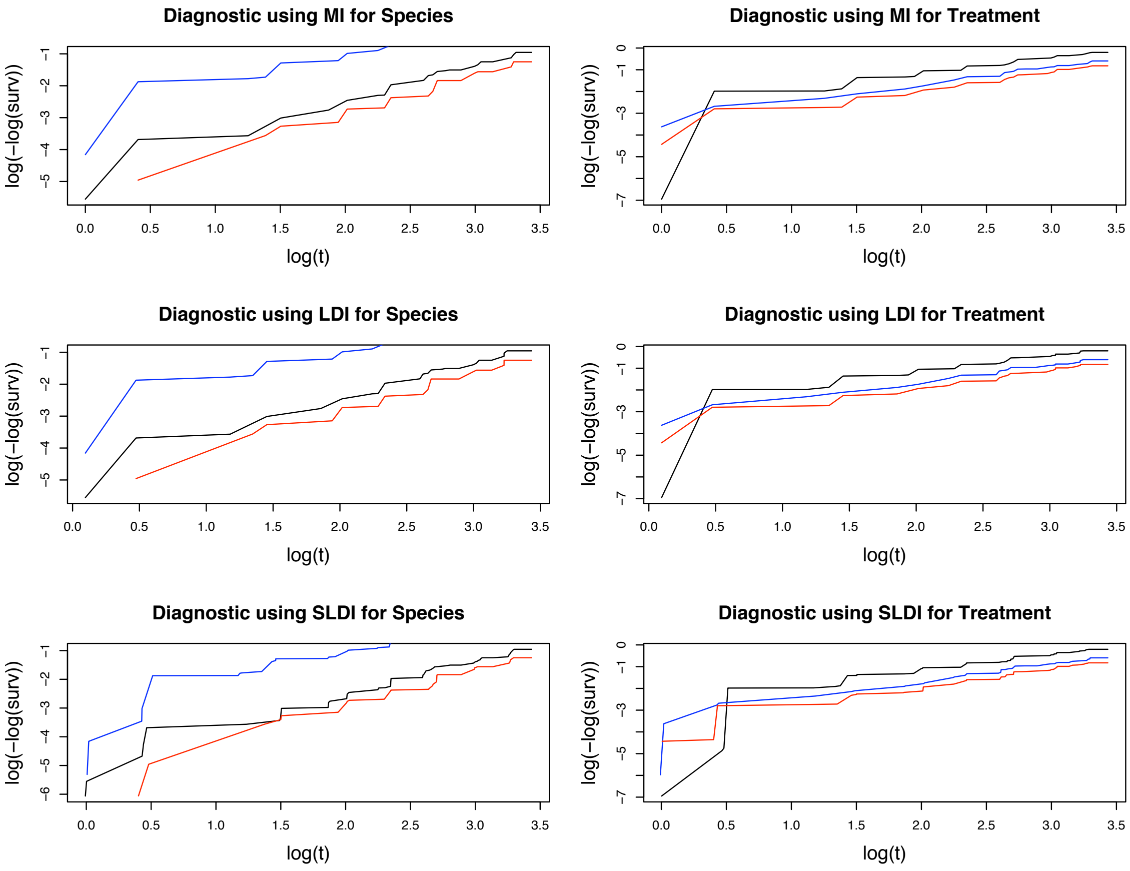


Figure 3.11: Diagnostic Plots to check for Proportional Hazard Assumption using the Three Imputed values

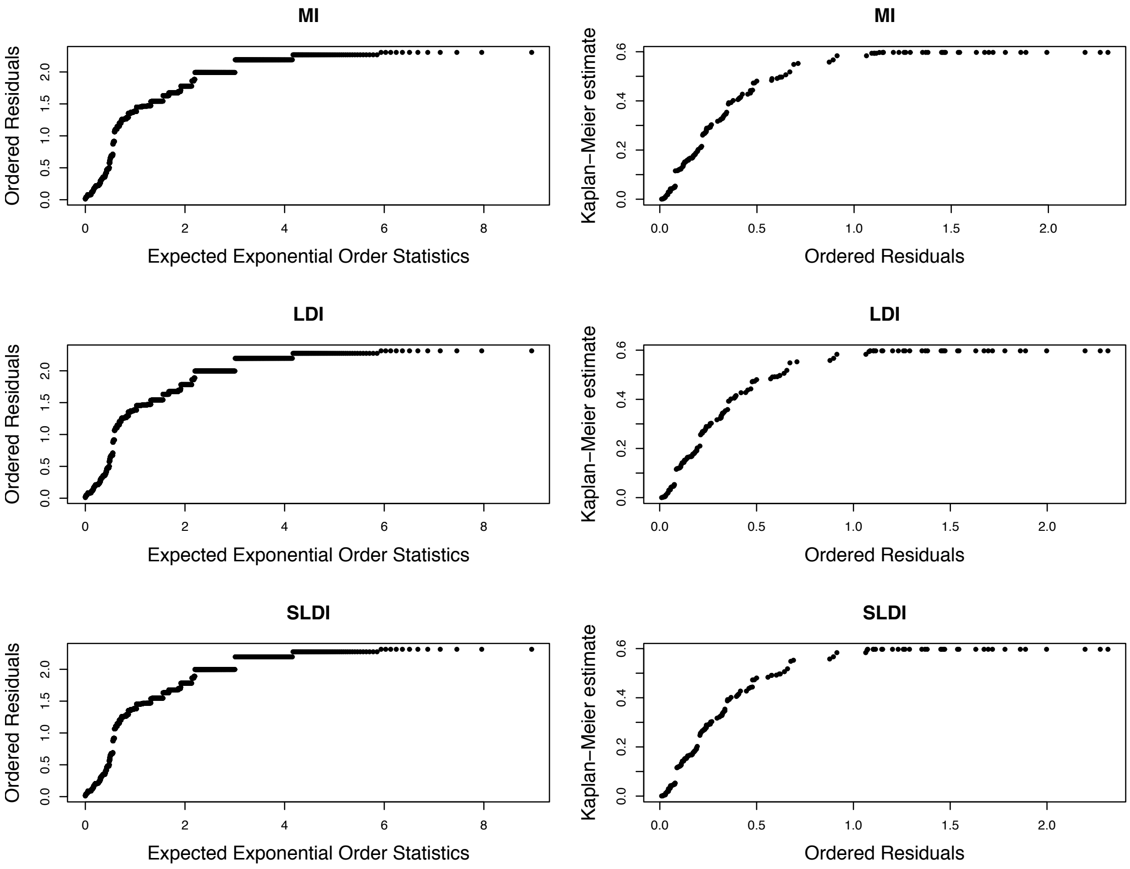


Figure 3.12: Residual Diagnostic Plots to check for Weibull Regression Model using the Three Imputed values

# Chapter 4

# Simulation Study

Simulation studies were executed in order to compare the three imputation methods. The three methods compared are: (i) midpoint (MI), (ii) local likelihood density imputation applied to the data as a whole (LDI), (iii) local likelihood density imputation within strata defined by species and thinning levels (SLDI). The performance of the three methods was compared, in terms of their ability to produce precise and accurate estimates of covariate effects in a Weibull survival analysis.

In a complete factorial design, datasets containing the true lifetime since thinning were generated from a Weibull distribution with scale parameter $\theta$ and the shape parameter equal to 5 where the scale parameter is defined as:

$$\theta = e^{-X'\beta}$$

$\beta$ is a vector containing the true parameter estimates and the Weibull distribution takes the following form:

$$S(t) = \theta e^{-5\theta t}$$

In the datasets, three types of species were present as in the analysis of the tree mortality data: Douglas Fir, Western Hemlock and Western Cedar, each representing 30, 60 and 10

% of the trees in the datasets. Within each type of species, three levels of treatment were randomly assigned: 50 % applied to Control group, and 25 % each for Low Thinning and High Thinning treatments.

Intervals of tree lifetimes were simulated based on the measurement times. The measurement time $V_j$ for the $j$ th visit varies among the trees. The time intervals between the visit times are modeled as exponential; here, the time between subsequent visits $V_j - V_{j-1}$ follows the distribution $exp(\lambda)$.

Under each scenario, the simulated lifetime intervals for tree $i$ ($L_i$, $R_i$) are:

if $T_i < max\{V_j\}$

$$L_i = max\{V_j | V_j \leqslant T_i | \delta_j = 1\}$$

$$R_i = min\{V_j | V_j > T_i | \delta_j = 1\}$$

if $T_i > max\{V_j\}$

$$L_i = max\{V_j | \delta_j = 1\}$$

$$R_i = \infty$$

where $\delta_j = 1$ when the tree is visited at the $j$th visit $V_j$, and otherwise is 0.

Five factors were initially identified potentially important in their effect on the estimation of the lifetimes using the local likelihood density imputation (LDI): the sample size (n), the covariate effect ($\beta$), the probability of a tree being missed at the visit (p), the interval width ($\lambda$), and bandwidth of the kernel (h). We set p as 10% so as to have the simulated datasets comparable with the tree mortality data. In the tree mortality data, not all the trees are visited for each measurement. This suggests in our simulation scenario that measurements of trees are omitted for such visits; we assume such missingness happens at random. To mimic this, the probability of missing visits for each tree will be modeled through a Bernoulli random variable $\delta_j$ where the probability of missing a visit for a tree was p=10%, approximated from the tree mortality data. We fix the bandwidth as 0.4 so that the amount of smoothing is controlled in the estimation of the underlying distribution regardless of the sample size per stratum.

The full factorial design was performed with the following three factors:

- Sample size (n)

  Sample size has two levels : n=100 and n=1000, each representing small and large samples.

- Interval width ($\lambda$)

  Interval width has two levels: $\lambda = 2$ and $\lambda = 5$, each representing small and large intervals. This reflects the time between subsequent visits.

- Covariate effect ($\beta$)

  The covariate effect has three levels, each representing no effect ($\beta$=0,0,0,0), moderate effect ($\beta$=-0.1,0.1,0.1,0.1), and large covariate effect ($\beta$=-1,0.5,0.5,0.5). The parameter $\beta_1$ represents the effect of Western Hemlock relative to Douglas Fir, $\beta_2$ for the effect of Western Cedar relative to Douglas Fir, $\beta_3$ for the effect of Low Thinning relative to Control, and $\beta_4$ represents the effect of High Thinning relative to Control.

The three aforementioned imputation methods were applied to the simulated data sets to compare their performance. The Weibull accelerated failure time regression model was fit to the three sets of imputed data under each scenario in order to produce the parameter estimates for our covariates. Table 4.1 displays the twelve scenarios that were examined based on 1000 simulated data sets.

Figure 4.1 displays distributions of the estimates of $\beta$ obtained from using the imputation methods and scenarios in the case of $\beta = (0, 0, 0, 0)$; Table 4.2 provides summary statistics for these distributions. All three imputation methods performed well in terms of producing estimates that are close to the true values. The estimates obtained from LDI has the smallest standard deviation. As sample size increases, the overall precision improves for all three methods. As the interval width increases, however, MI produces estimates that are considerably more variable. MI is more sensitive to interval width than LDI or SLDI estimates.

Table 4.1: Summary of Simulation Design by Scenario

| Scenario | Sample size | Interval width | Covariate efffect |
|---|---|---|---|
| 1 | n=100 | $\lambda$=2 | $\beta = (0, 0, 0, 0)$ |
| 2 | n=100 | $\lambda$=2 | $\beta = (-0.1, 0.1, 0.1, 0.1)$ |
| 3 | n=100 | $\lambda$=2 | $\beta = (-1, 0.5, 0.5, 0.5)$ |
| 4 | n=100 | $\lambda$=5 | $\beta = (0, 0, 0, 0)$ |
| 5 | n=100 | $\lambda$=5 | $\beta = (-0.1, 0.1, 0.1, 0.1)$ |
| 6 | n=100 | $\lambda$=5 | $\beta = (-1, 0.5, 0.5, 0.5)$ |
| 7 | n=1000 | $\lambda$=2 | $\beta = (0, 0, 0, 0)$ |
| 8 | n=1000 | $\lambda$=2 | $\beta = (-0.1, 0.1, 0.1, 0.1)$ |
| 9 | n=1000 | $\lambda$=2 | $\beta = (-1, 0.5, 0.5, 0.5)$ |
| 10 | n=1000 | $\lambda$=5 | $\beta = (0, 0, 0, 0)$ |
| 11 | n=1000 | $\lambda$=5 | $\beta = (-0.1, 0.1, 0.1, 0.1)$ |
| 12 | n=1000 | $\lambda$=5 | $\beta = (-1, 0.5, 0.5, 0.5)$ |

Figure 4.2 displays distributions of estimates for the scenarios when moderate covariate effects are present; Table 4.2 provides a summary of these estimates. The use of SLDI slightly outperforms the other two methods as the estimates obtained from SLDI are closer to the true values of $\beta$ and have smaller variances compared to the other two methods. An increase in sample size contributes to improved precision but it does not seem to affect the overall accuracy. As the interval width increases, the overall variation in estimates increases for all three methods.

Figure 4.3 displays estimates obtained for those scenarios when large covariate effects are present while Table 4.2 summarizes the estimates. Except SLDI, all other methods failed to produce estimates that are close to the true values of $\beta$ and biases are quite large here. An increase in sample size does not reduce bias but reduces variability. An increase in interval width accompanied with large covariate effects leads to poor performance by MI and LDI. SLDI seems less affected by the increase in interval width as it still produced unbiased estimates with relatively small variances.

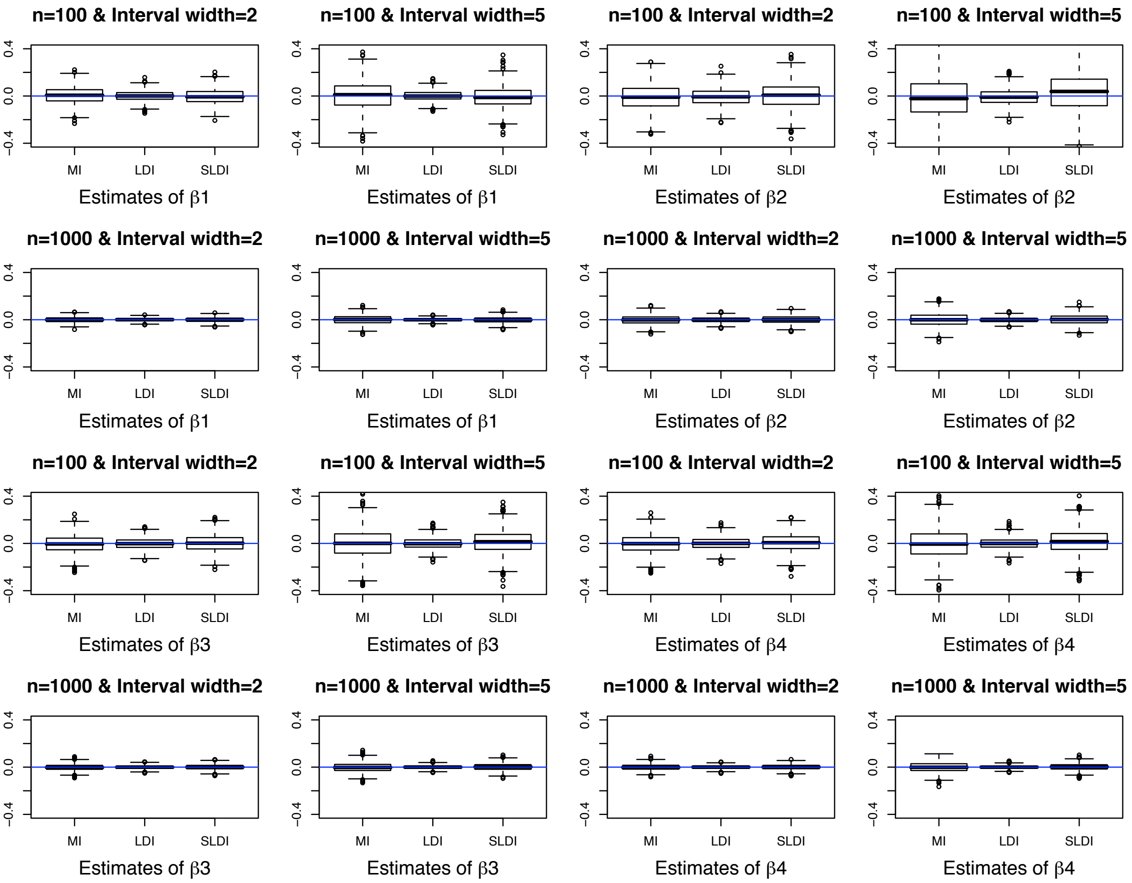


Figure 4.1: Boxplots of Estimates of $\beta$ by Imputation Methods and by Scenario for No Covariate Effects. The blue horizontal line indicates the true value of $\beta$

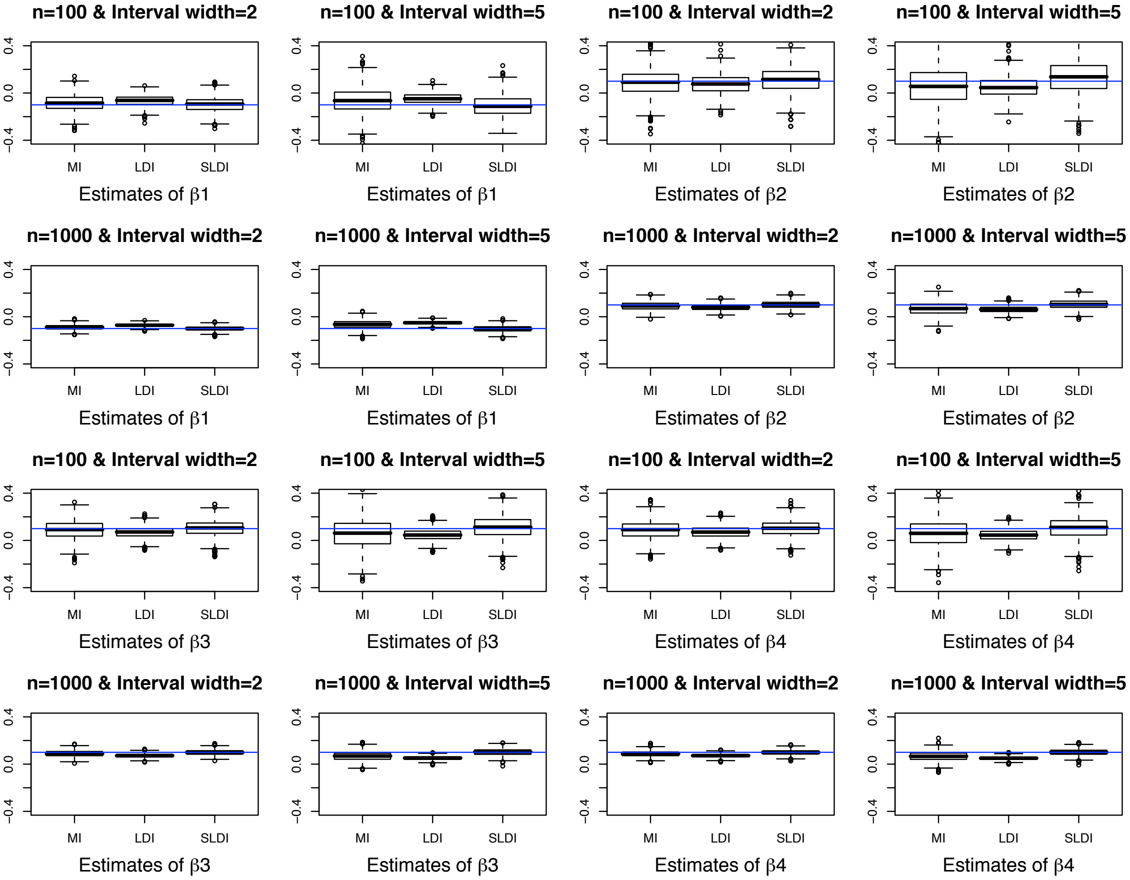


Figure 4.2: Boxplots of Estimates of $\beta$ by Imputation Methods and by Scenario for Moderate Covariate Effects. The blue horizontal line indicates the true value of $\beta$

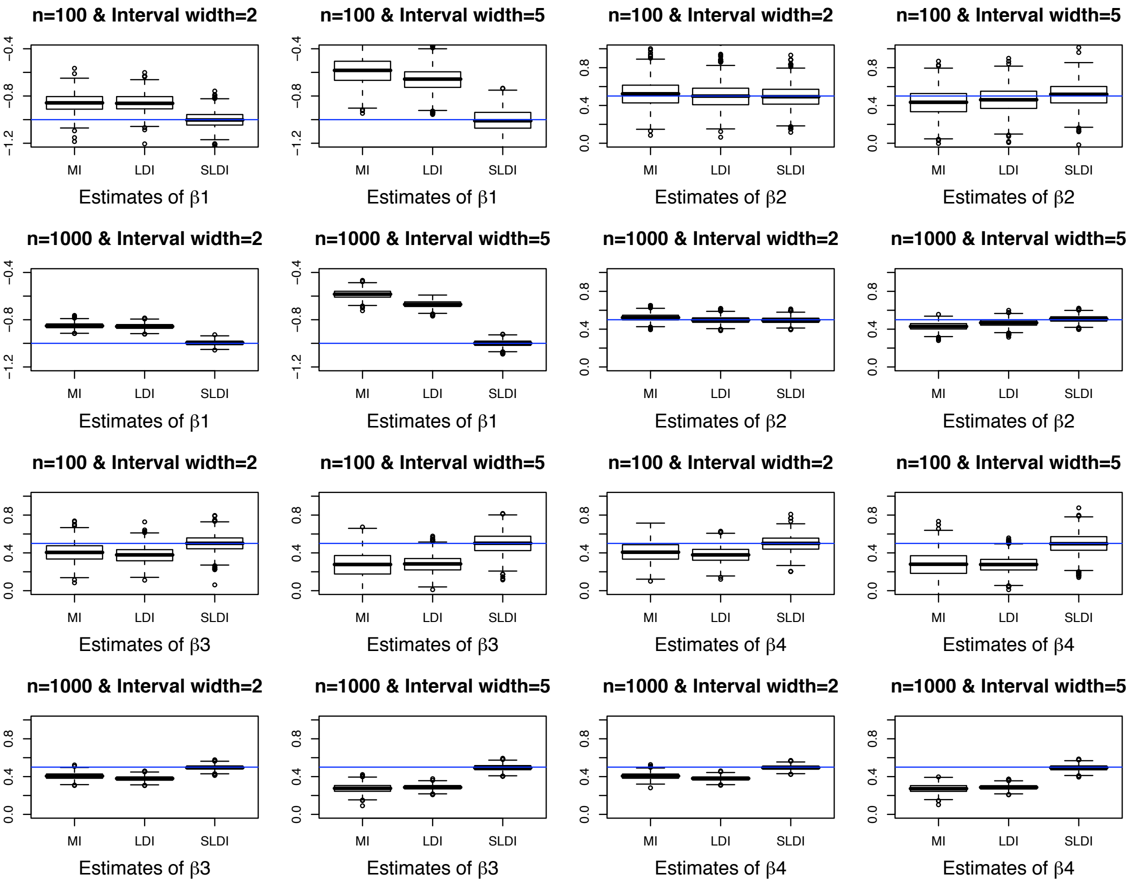


Figure 4.3: Boxplots of Estimates of $\beta$ by Imputation Methods and by Scenario for Large Covariate Effects. The blue horizontal line indicates the true value of $\beta$

Table 4.2: Mean and Standard Deviation (SD) of $\hat{\beta}$ by Imputation Methods and Scenario

| | LDI | | SLDI | | MI | | LDI | | SLDI | | MI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| | $(\beta = (\mathbf{0, 0, 0, 0}), \mathbf{n} = \mathbf{100}, \lambda = \mathbf{2})$ | | | | | | $(\beta = (\mathbf{0, 0, 0, 0}), \mathbf{n} = \mathbf{100}, \lambda = \mathbf{5})$ | | | | | |
| $\beta_1$ | 0.001 | 0.043 | -0.005 | 0.062 | 0.004 | 0.070 | 0.001 | 0.042 | -0.012 | 0.090 | 0.004 | 0.116 |
| $\beta_2$ | -0.008 | 0.070 | 0.002 | 0.107 | -0.011 | 0.108 | -0.009 | 0.065 | 0.031 | 0.176 | -0.019 | 0.180 |
| $\beta_3$ | -0.002 | 0.048 | 0.001 | 0.071 | -0.005 | 0.077 | -0.001 | 0.047 | 0.013 | 0.096 | -0.001 | 0.125 |
| $\beta_4$ | 0.000 | 0.049 | 0.005 | 0.072 | -0.002 | 0.077 | 0.000 | 0.049 | 0.016 | 0.103 | -0.002 | 0.124 |
| | $(\beta = (\mathbf{0, 0, 0, 0}), \mathbf{n} = \mathbf{1000}, \lambda = \mathbf{2})$ | | | | | | $(\beta = (\mathbf{0, 0, 0, 0}), \mathbf{n} = \mathbf{1000}, \lambda = \mathbf{5})$ | | | | | |
| $\beta_1$ | 0.000 | 0.014 | -0.000 | 0.020 | 0.000 | 0.023 | 0.000 | 0.013 | -0.001 | 0.025 | -0.000 | 0.037 |
| $\beta_2$ | 0.001 | 0.022 | 0.001 | 0.032 | -0.001 | 0.037 | -0.001 | 0.021 | 0.003 | 0.042 | -0.001 | 0.058 |
| $\beta_3$ | 0.000 | 0.015 | 0.000 | 0.021 | -0.002 | 0.026 | -0.000 | 0.015 | 0.002 | 0.028 | -0.001 | 0.039 |
| $\beta_4$ | -0.000 | 0.015 | 0.001 | 0.021 | -0.001 | 0.025 | 0.000 | 0.015 | 0.002 | 0.028 | -0.001 | 0.041 |
| | $(\beta = (\mathbf{-0.1, 0.1, 0.1, 0.1}), \mathbf{n} = \mathbf{100}, \lambda = \mathbf{2})$ | | | | | | $(\beta = (\mathbf{-0.1, 0.1, 0.1, 0.1}), \mathbf{n} = \mathbf{100}, \lambda = \mathbf{5})$ | | | | | |
| $\beta_1$ | -0.065 | 0.047 | -0.096 | 0.064 | -0.084 | 0.068 | -0.049 | 0.046 | -0.110 | 0.088 | -0.062 | 0.110 |
| $\beta_2$ | 0.076 | 0.084 | 0.109 | 0.107 | 0.089 | 0.111 | 0.049 | 0.089 | 0.133 | 0.152 | 0.062 | 0.170 |
| $\beta_3$ | 0.070 | 0.050 | 0.102 | 0.070 | 0.088 | 0.079 | 0.048 | 0.050 | 0.112 | 0.097 | 0.059 | 0.127 |
| $\beta_4$ | 0.070 | 0.050 | 0.103 | 0.070 | 0.090 | 0.078 | 0.046 | 0.047 | 0.108 | 0.094 | 0.061 | 0.121 |
| | $(\beta = (\mathbf{-0.1, 0.1, 0.1, 0.1}), \mathbf{n} = \mathbf{1000}, \lambda = \mathbf{2})$ | | | | | | $(\beta = (\mathbf{-0.1, 0.1, 0.1, 0.1}), \mathbf{n} = \mathbf{1000}, \lambda = \mathbf{5})$ | | | | | |
| $\beta_1$ | -0.072 | 0.014 | -0.100 | 0.019 | -0.089 | 0.022 | -0.051 | 0.014 | -0.102 | 0.025 | -0.066 | 0.036 |
| $\beta_2$ | 0.080 | 0.026 | 0.101 | 0.031 | 0.091 | 0.034 | 0.063 | 0.028 | 0.105 | 0.040 | 0.069 | 0.055 |
| $\beta_3$ | 0.072 | 0.016 | 0.099 | 0.021 | 0.089 | 0.025 | 0.051 | 0.016 | 0.102 | 0.028 | 0.068 | 0.040 |
| $\beta_4$ | 0.071 | 0.016 | 0.099 | 0.021 | 0.087 | 0.024 | 0.051 | 0.015 | 0.101 | 0.027 | 0.066 | 0.041 |
| | $(\beta = (\mathbf{-1, 0.5, 0.5, 0.5}), \mathbf{n} = \mathbf{100}, \lambda = \mathbf{2})$ | | | | | | $(\beta = (\mathbf{-1, 0.5, 0.5, 0.5}), \mathbf{n} = \mathbf{100}, \lambda = \mathbf{5})$ | | | | | |
| $\beta_1$ | -0.860 | 0.077 | -1.003 | 0.069 | -0.859 | 0.080 | -0.660 | 0.098 | -1.007 | 0.097 | -0.588 | 0.119 |
| $\beta_2$ | 0.497 | 0.135 | 0.491 | 0.120 | 0.524 | 0.148 | 0.460 | 0.138 | 0.516 | 0.125 | 0.430 | 0.138 |
| $\beta_3$ | 0.378 | 0.087 | 0.501 | 0.085 | 0.406 | 0.106 | 0.281 | 0.091 | 0.501 | 0.115 | 0.275 | 0.142 |
| $\beta_4$ | 0.380 | 0.085 | 0.499 | 0.085 | 0.410 | 0.105 | 0.278 | 0.087 | 0.496 | 0.110 | 0.277 | 0.139 |
| | $(\beta = (\mathbf{-1, 0.5, 0.5, 0.5}), \mathbf{n} = \mathbf{1000}, \lambda = \mathbf{2})$ | | | | | | $(\beta = (\mathbf{-1, 0.5, 0.5, 0.5}), \mathbf{n} = \mathbf{1000}, \lambda = \mathbf{5})$ | | | | | |
| $\beta_1$ | -0.857 | 0.023 | -0.996 | 0.021 | -0.851 | 0.025 | -0.669 | 0.028 | -0.999 | 0.027 | -0.584 | 0.036 |
| $\beta_2$ | 0.497 | 0.036 | 0.495 | 0.034 | 0.523 | 0.040 | 0.468 | 0.040 | 0.508 | 0.034 | 0.429 | 0.041 |
| $\beta_3$ | 0.380 | 0.025 | 0.496 | 0.025 | 0.406 | 0.033 | 0.287 | 0.026 | 0.495 | 0.032 | 0.275 | 0.045 |
| $\beta_4$ | 0.380 | 0.024 | 0.495 | 0.024 | 0.405 | 0.032 | 0.286 | 0.026 | 0.493 | 0.031 | 0.274 | 0.045 |

# Chapter 5

# Summary

The interval midpoint (MI) and an imputation method based on local likelihood density estimation were employed to analyze interval censored tree mortality data. A focus of the analysis was to assess the effect of species and thinning levels on lifetimes of trees. Since the imputation method based on local likelihood density estimation (LDI) does not handle potential covariate effects, we implemented it by each strata based on each combination of Species and Treatment (SLDI). In terms of estimated effects of Species and Treatment on tree lifetimes, all three methods produced similar results. The imputed data sets were very close to each other as were the resulting covariate estimates. As we suspected that this was due to small interval widths or relatively small covariate effects that are present in the tree mortality data, we also conducted a simulation study to investigate further the factors that affect the performance of these three imputation methods. The factors of particular interests were sample size, interval width and the size of covariate effects. The simulation studies showed that where there are no covariate effects, all three methods performed well in producing unbiased estimates with similar standard errors. The estimates based on the use of SLDI produced larger standard errors, as expected, since the smoothing is performed on stratified samples and therefore based on a smaller number of observations. Further, MI was most sensitive to interval width as interval width increased, there were significant increases in variation of the estimates. With increased covariate effects, SLDI outperforms the other two methods by producing estimates for $\beta$ which are far less biased. The simulation studies suggested that both MI and LDI failed to produce unbiased estimates when covariate effects are large. Increased sample size had an impact on precision of estimates in general. The dominant factor to determine the overall performance of the imputation methods was the size of the covariate effects (although both interval width and sample size had some impact on precision and accuracy of estimates). SLDI performed well regardless of the size of covariate effects, interval size and sample size. This suggests that potential covariate effects cannot be ignored in implementing the local likelihood imputation method, and that when the covariate effects are handled properly, the local likelihood imputation method is more robust and reliable than MI in imputing true lifetimes.

# Chapter 6

# Discussion

Through the simulation studies, the imputation method based on local likelihood density estimation was proven to perform well when covariate effects were handled with stratification in its implementation while the approach failed when such covariate effects were ignored. In regression settings in survival analysis where the effects of covariates are investigated, the imputation method based on local likelihood density estimation requires a means to incorporate covariate effects in its implementation. A more efficient approach than stratification may de developed to incorporate such effects. This is particularly true in the presence of continuous covariates. Another point that requires attention concerns variation of interval widths within a data set. In our simulation studies, the widths of intervals were controlled to be about the same within the data set. The density estimate acquired by the imputation method based on local likelihood density estimation needs further investigation to assess whether it can describe the true underlying density properly in situations where the interval widths vary significantly within a data set.

# Appendix A

# Tables and Figures

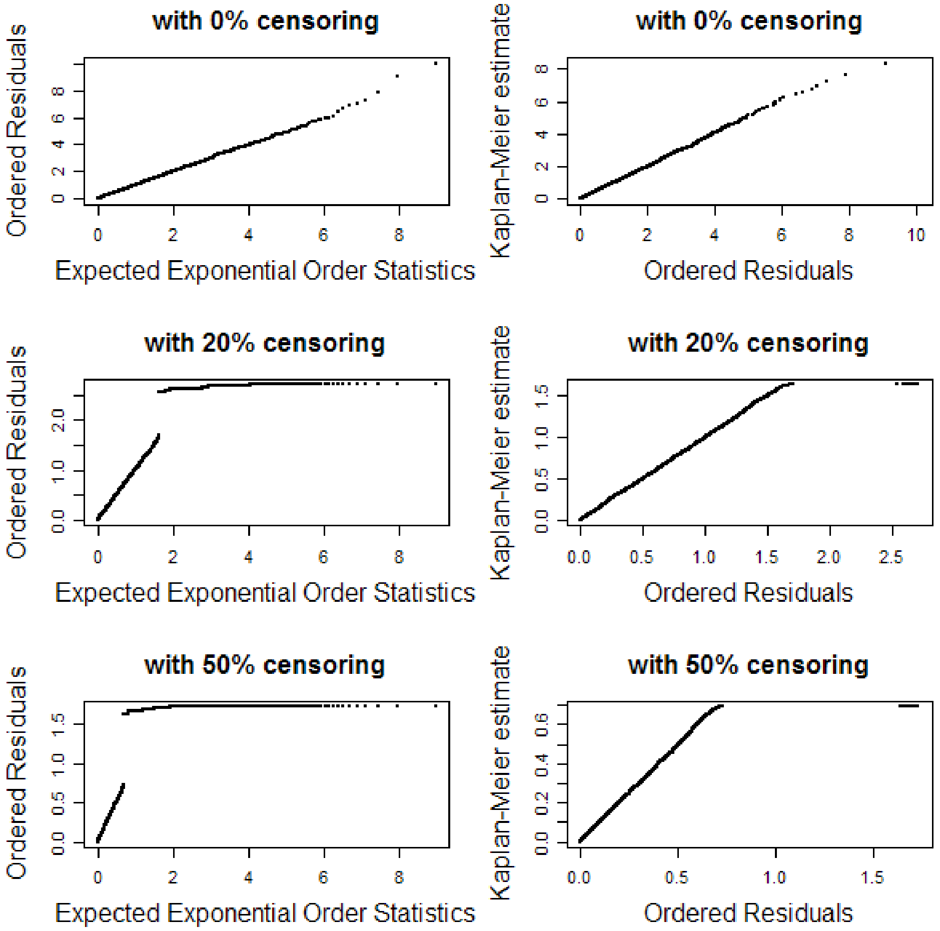


Figure A.1: Diagnostic Plots to check for Weibull Assumption using Simulated Data with Various amount of Censoring

Table A.1: Average Age of Trees at Thinning by Installation and Plot

| Installation | Plot | Age at thinning |
|---|---|---|
| 1 | 1 | 40 |
| | 2 | 40 |
| | 3 | 42 |
| | 4 | 38 |
| | 5 | 36 |
| | 6 | 39 |
| 2 | 1 | 33 |
| | 2 | 34 |
| | 3 | 35 |
| | 4 | 35 |
| | 5 | 35 |
| | 6 | 35 |
| 3 | 1 | 37 |
| | 2 | 35 |
| | 3 | 36 |
| | 4 | 40 |
| | 5 | 40 |
| | 6 | 39 |
| 4 | 1 | 35 |
| | 2 | 36 |
| | 3 | 33 |
| | 4 | 36 |
| | 5 | 35 |
| | 6 | 36 |
| 5 | 1 | 46 |
| | 2 | 50 |
| | 3 | 40 |
| | 4 | 41 |
| 6 | 1 | 28 |
| | 2 | 28 |
| | 3 | 27 |
| | 4 | 27 |
| | 5 | 26 |
| | 6 | 27 |

Table A.2: Counts of Trees by Installation, Plot and Measurement Time (M.i refers to the (i)th follow-up time after thinning, i=1,...,9)

| Installation | Plot | M.1 | M.2 | M.3 | M.4 | M.5 | M.6 | M.7 | M.8 | M.9 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 149 | 147 | 143 | 140 | 135 | 129 | 115 | 105 |  | 159 |
|  | 2 | 50 | 49 | 49 | 50 | 52 | 53 | 57 | 57 |  | 65 |
|  | 3 | 91 | 91 | 92 | 88 | 87 | 84 | 87 | 87 |  | 114 |
|  | 4 | 51 | 51 | 53 | 53 | 51 | 52 | 51 | 52 |  | 62 |
|  | 5 | 61 | 61 | 62 | 64 | 62 | 62 | 59 | 57 |  | 70 |
|  | 6 | 102 | 104 | 98 | 98 | 84 | 74 | 60 | 59 |  | 112 |
| Subtotal |  | 504 | 503 | 497 | 493 | 471 | 454 | 429 | 417 |  | 582 |
| 2 | 1 | 101 | 101 | 95 | 93 | 92 | 77 | 69 | 61 |  | 101 |
|  | 2 | 148 | 148 | 140 | 132 | 118 | 100 | 81 | 66 |  | 148 |
|  | 3 | 177 | 169 | 149 | 128 | 115 | 88 | 79 | 74 |  | 177 |
|  | 4 | 221 | 211 | 177 | 146 | 131 | 118 | 103 | 94 |  | 221 |
|  | 5 | 91 | 87 | 80 | 78 | 73 | 69 | 56 | 52 |  | 91 |
|  | 6 | 258 | 250 | 186 | 151 | 130 | 103 | 89 | 79 |  | 256 |
| Subtotal |  | 996 | 966 | 827 | 728 | 659 | 555 | 477 | 426 |  | 994 |
| 3 | 1 | 162 | 155 | 133 | 116 | 104 | 95 | 81 |  |  | 162 |
|  | 2 | 144 | 134 | 114 | 105 | 98 | 91 | 75 |  |  | 144 |
|  | 3 | 153 | 149 | 128 | 120 | 109 | 100 | 84 |  |  | 153 |
|  | 4 | 347 | 331 | 257 | 219 | 185 | 165 | 125 |  |  | 347 |
|  | 5 | 321 | 301 | 235 | 183 | 157 | 147 | 107 |  |  | 321 |
|  | 6 | 149 | 146 | 128 | 115 | 106 | 93 | 81 |  |  | 149 |
| Subtotal |  | 1276 | 1216 | 995 | 858 | 759 | 691 | 553 |  |  | 1276 |
| 4 | 1 | 84 | 84 | 84 | 80 | 79 | 73 | 67 | 65 |  | 84 |
|  | 2 | 131 | 130 | 122 | 116 | 103 | 91 | 77 | 69 |  | 132 |
|  | 3 | 77 | 76 | 75 | 76 | 74 | 72 | 69 | 64 |  | 80 |
|  | 4 | 55 | 55 | 52 | 52 | 51 | 49 | 46 | 43 |  | 55 |
|  | 5 | 100 | 99 | 99 | 96 | 94 | 90 |  |  |  | 102 |
|  | 6 | 116 | 116 | 118 | 114 | 111 | 111 | 103 | 100 |  | 128 |
| Subtotal |  | 563 | 560 | 550 | 534 | 512 | 486 | 362 | 341 |  | 581 |
| 5 | 1 | 55 | 51 | 50 | 49 | 49 | 47 | 43 |  |  | 55 |
|  | 2 | 74 | 64 | 63 | 58 | 55 | 53 | 43 |  |  | 74 |
|  | 3 | 81 | 79 | 72 | 61 | 58 | 54 | 48 |  |  | 81 |
|  | 4 | 92 | 91 | 87 | 72 | 65 | 61 | 51 |  |  | 92 |
| Subtotal |  | 302 | 285 | 272 | 240 | 227 | 215 | 185 |  |  | 302 |
| 6 | 1 | 73 | 73 | 73 | 70 | 70 | 68 | 67 | 65 | 52 | 76 |
|  | 2 | 79 | 80 | 75 | 76 | 73 | 67 | 66 | 64 | 59 | 83 |
|  | 3 | 116 | 116 | 110 | 108 | 107 | 100 | 93 | 94 | 88 | 118 |
|  | 4 | 138 | 138 | 137 | 132 | 128 | 125 | 117 | 112 | 102 | 226 |
|  | 5 | 89 | 91 | 93 | 93 | 91 | 91 | 90 | 91 | 88 | 173 |
|  | 6 | 86 | 86 | 86 | 87 | 88 | 87 | 84 | 85 | 84 | 164 |
| Subtotal |  | 581 | 584 | 574 | 566 | 557 | 538 | 517 | 511 | 473 | 840 |
| Grand Total |  | 4222 | 4114 | 3715 | 3419 | 3185 | 2939 | 2523 | 1695 | 473 | 4575 |

Table A.3: Counts of Live Trees at Measurement Times (M.i refers to the (i)th follow-up time after thinning, i=1,...,9) by Installation and Plot

| Installation | Plot | M.1 | M.2 | M.3 | M.4 | M.5 | M.6 | M.7 | M.8 | M.9 |
|---:|---|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 1 | 147 | 143 | 138 | 134 | 129 | 108 | 105 | 94 | |
| | 2 | 49 | 46 | 48 | 50 | 52 | 50 | 57 | 55 | |
| | 3 | 91 | 88 | 88 | 86 | 82 | 72 | 86 | 78 | |
| | 4 | 51 | 50 | 53 | 51 | 49 | 47 | 51 | 45 | |
| | 5 | 60 | 61 | 60 | 62 | 62 | 53 | 56 | 53 | |
| | 6 | 102 | 97 | 97 | 84 | 74 | 57 | 56 | 54 | |
| Subtotal | | 500 | 485 | 484 | 467 | 448 | 387 | 411 | 379 | |
| 2 | 1 | 101 | 95 | 93 | 92 | 77 | 69 | 61 | 55 | |
| | 2 | 148 | 140 | 132 | 118 | 100 | 81 | 66 | 59 | |
| | 3 | 169 | 149 | 128 | 115 | 88 | 79 | 74 | 70 | |
| | 4 | 211 | 177 | 146 | 131 | 118 | 103 | 94 | 79 | |
| | 5 | 87 | 80 | 78 | 73 | 69 | 56 | 52 | 48 | |
| | 6 | 250 | 186 | 151 | 130 | 103 | 89 | 79 | 69 | |
| Subtotal | | 966 | 827 | 728 | 659 | 555 | 477 | 426 | 380 | |
| 3 | 1 | 155 | 133 | 116 | 104 | 95 | 81 | 72 | | |
| | 2 | 134 | 114 | 105 | 98 | 91 | 75 | 69 | | |
| | 3 | 149 | 128 | 120 | 109 | 100 | 84 | 74 | | |
| | 4 | 331 | 257 | 219 | 185 | 165 | 125 | 99 | | |
| | 5 | 301 | 235 | 183 | 157 | 147 | 107 | 89 | | |
| | 6 | 146 | 128 | 115 | 106 | 93 | 81 | 72 | | |
| Subtotal | | 1216 | 995 | 858 | 759 | 691 | 553 | 475 | | |
| 4 | 1 | 84 | 84 | 80 | 79 | 73 | 67 | 65 | 60 | |
| | 2 | 130 | 122 | 115 | 103 | 91 | 77 | 69 | 59 | |
| | 3 | 76 | 74 | 75 | 74 | 71 | 69 | 64 | 58 | |
| | 4 | 55 | 52 | 52 | 51 | 49 | 46 | 43 | 42 | |
| | 5 | 99 | 99 | 96 | 94 | 88 | 80 | | | |
| | 6 | 116 | 114 | 114 | 110 | 104 | 103 | 100 | 86 | |
| Subtotal | | 560 | 545 | 532 | 511 | 476 | 442 | 341 | 305 | |
| 5 | 1 | 51 | 50 | 49 | 49 | 47 | 43 | 41 | | |
| | 2 | 64 | 63 | 58 | 55 | 53 | 43 | 40 | | |
| | 3 | 79 | 72 | 61 | 58 | 54 | 48 | 42 | | |
| | 4 | 91 | 87 | 72 | 65 | 61 | 51 | 40 | | |
| Subtotal | | 285 | 272 | 240 | 227 | 215 | 185 | 162 | | |
| 6 | 1 | 73 | 72 | 70 | 68 | 68 | 67 | 65 | 52 | 49 |
| | 2 | 79 | 75 | 75 | 73 | 67 | 65 | 63 | 59 | 51 |
| | 3 | 116 | 110 | 107 | 107 | 100 | 93 | 93 | 88 | 82 |
| | 4 | 138 | 137 | 132 | 128 | 125 | 117 | 112 | 100 | 90 |
| | 5 | 89 | 90 | 93 | 91 | 90 | 89 | 89 | 88 | 72 |
| | 6 | 86 | 85 | 86 | 86 | 87 | 84 | 84 | 80 | 67 |
| Subtotal | | 581 | 569 | 563 | 553 | 537 | 515 | 506 | 467 | 411 |
| Grand Total | | 4108 | 3693 | 3405 | 3176 | 2922 | 2559 | 2322 | 1531 | 411 |
| Total Visited (from A.2) | | 4222 | 4114 | 3715 | 3419 | 3185 | 2939 | 2523 | 1695 | 473 |
| Proportion of Missing Trees | | 7.7% | 7.8% | 8.0% | 8.3% | 8.7% | 8.8% | 11.3% | 35.9% | 80.9% |

Table A.4: Counts of Trees by Installation, Plot and Thinning Level

| Installation | Plot | Control | Low Thinning | High Thinning | Total |
|---|---|---|---|---|---|
| 1 | 1 | 159 | | | 159 |
| | 2 | | | 65 | 65 |
| | 3 | | 114 | | 114 |
| | 4 | | 62 | | 62 |
| | 5 | | | 70 | 70 |
| | 6 | 112 | | | 112 |
| 2 | 1 | 2 | | 99 | 101 |
| | 2 | | 148 | | 148 |
| | 3 | | 177 | | 177 |
| | 4 | 221 | | | 221 |
| | 5 | | | 91 | 91 |
| | 6 | 256 | | | 256 |
| 3 | 1 | | | 162 | 162 |
| | 2 | | 144 | | 144 |
| | 3 | | 153 | | 153 |
| | 4 | 347 | | | 347 |
| | 5 | 321 | | | 321 |
| | 6 | | | 149 | 149 |
| 4 | 1 | | 84 | | 84 |
| | 2 | 132 | | | 132 |
| | 3 | | | 80 | 80 |
| | 4 | | | 55 | 55 |
| | 5 | | 102 | | 102 |
| | 6 | 128 | | | 128 |
| 5 | 1 | | | 55 | 55 |
| | 2 | | | 74 | 74 |
| | 3 | 81 | | | 81 |
| | 4 | 92 | | | 92 |
| 6 | 1 | | | 76 | 76 |
| | 2 | | 83 | | 83 |
| | 3 | 118 | | | 118 |
| | 4 | 226 | | | 226 |
| | 5 | | | 173 | 173 |
| | 6 | | 164 | | 164 |
| Grand Total | | 2195 | 1231 | 1149 | 4575 |

Table A.5: Correspondence Table for Relabeling

| Labeling by Ministry | Labeling by Ministry | Our labeling | Our labeling |
|---:|:---|---:|---:|
| Installation | Plot | Installation | Plot |
| 8 | 1 | 1 | 1 |
| | 2 | 1 | 2 |
| | 5 | 1 | 3 |
| | 10 | 1 | 4 |
| | 13 | 1 | 5 |
| | 14 | 1 | 6 |
| 28 | 2 | 2 | 1 |
| | 8 | 2 | 2 |
| | 13 | 2 | 3 |
| | 15 | 2 | 4 |
| | 25 | 2 | 5 |
| | 26 | 2 | 6 |
| 31 | 1 | 3 | 1 |
| | 2 | 3 | 2 |
| | 3 | 3 | 3 |
| | 4 | 3 | 4 |
| | 5 | 3 | 5 |
| | 6 | 3 | 6 |
| 43 | 3 | 4 | 1 |
| | 4 | 4 | 2 |
| | 5 | 4 | 3 |
| | 7 | 4 | 4 |
| | 8 | 4 | 5 |
| | 9 | 4 | 6 |
| 67 | 1 | 5 | 1 |
| | 2 | 5 | 2 |
| | 3 | 5 | 3 |
| | 4 | 5 | 4 |
| 71 | 3 | 6 | 1 |
| | 5 | 6 | 2 |
| | 11 | 6 | 3 |
| | 14 | 6 | 4 |
| | 15 | 6 | 5 |
| | 16 | 6 | 6 |

# Bibliography

Braun, J., T. Duchesne, and J. E. Stafford (2005). Local likelihood density estimation for interval censored data. *The Canadian Journal of Statistics 33*, 39–60.

Efron, B. (1967). Two sample problem with censored data. *University of California Press 4*, 831–853.

Elandt-Johnson, R. and N. Johnson (1980). *Survival Models and Data Analysis*. New York, NY: John Wiley & Sons, Inc.

Kalbfleisch, J. and R. Prentice (1980). *The Statistical Analysis of Failure Time Data*. New York, NY: John Wiley  Sons, Inc.

Kaplan, E. and P. Meier (1958). Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association 53*, 457–481.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley-Interscience.

Li, W. and Q. Yu (1997). Em algorithm for smoothing the self consistent estimator of survival functions with interval censored data. *Scandinavian Journal of Statistics 24*, 531–542.

Ortega, J. (1976). *Numerical Analysis: A Second Course*. New York, NY: Academic Press.

Ramsay, J. and B. Silverman (2005). *Functional data analysis*. New York, NY: Springer.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored, and truncated data. *Journal of the Royal Statistical Society Series B 38*, 290–295.

Wand, M. (2006). *Kernel Smoothing*. Chapman & Hall/CRC.