Tree Data Analysis (ID issue fixed)

Tahmidul Islam

12/16/2020

```
library(dplyr)
library(ggplot2)
library(kableExtra)
library(tidyr)
```

Tree Data

This dataset contains features of trees (height, width, etc.) from 3 sites: Brighton, Chase Stream and Lily Bay. There are 8 species in this sample.

Table: Data overview.

```
tree <- read.csv("E:/R Project/BayesFDA/data/tree/tree.csv")
tree <- tree %>% dplyr::select(Site, ID.Code, Tree, Rep, Sp, Year, Height) %>%
  filter(!is.na(Height)) %>%
  mutate(id = paste0(Site, ID.Code, Rep, Tree), logHeight = log(Height))
head(tree) %>% kable()
```

Site	ID.Code	Tree	Rep	Sp	Year	Height	id	logHeight
Brighton	S-BS-Mlt	1	1	BS	5	4.5	BrightonS-BS-Mlt11	1.504077
Brighton	S-BS-Mlt	1	1	BS	10	10.4	BrightonS-BS-Mlt11	2.341806
Brighton	S-BS-Mlt	2	1	BS	5	6.8	BrightonS-BS-Mlt12	1.916923
Brighton	S-BS-Mlt	2	1	BS	10	13.1	BrightonS-BS-Mlt12	2.572612
Brighton	S-BS-Mlt	4	1	BS	5	4.4	BrightonS-BS-Mlt14	1.481604
Brighton	S-BS-Mlt	4	1	BS	10	10.3	BrightonS-BS-Mlt14	2.332144

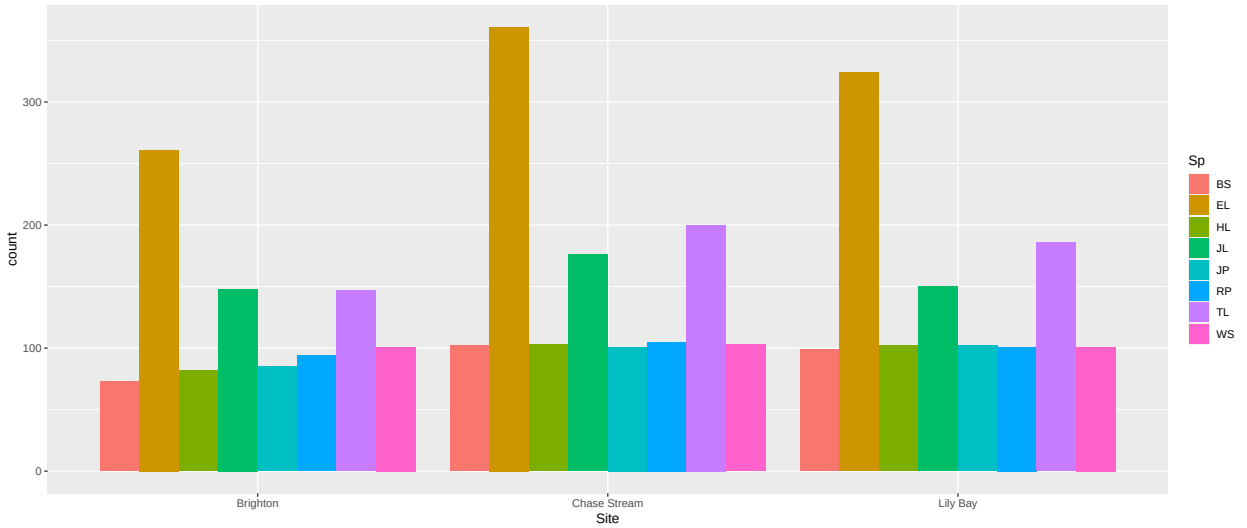
Table: Distribution of site and species.

```
sumDf <- tree %>% group_by(Site, Sp, id) %>% summarize (n = n()) %>% dplyr::select(Site, Sp)

sumDf %>% tally() %>% spread(Sp, n) %>% kable()
```

Site	BS	EL	HL	JL	JP	RP	TL	WS
Brighton	73	261	82	148	85	94	147	101
Chase Stream	102	361	103	176	101	105	200	103
Lily Bay	99	324	102	150	102	101	186	101

```
ggplot(sumDf, aes(x = Site, fill = Sp)) + geom_bar(position="dodge", stat="count")
```

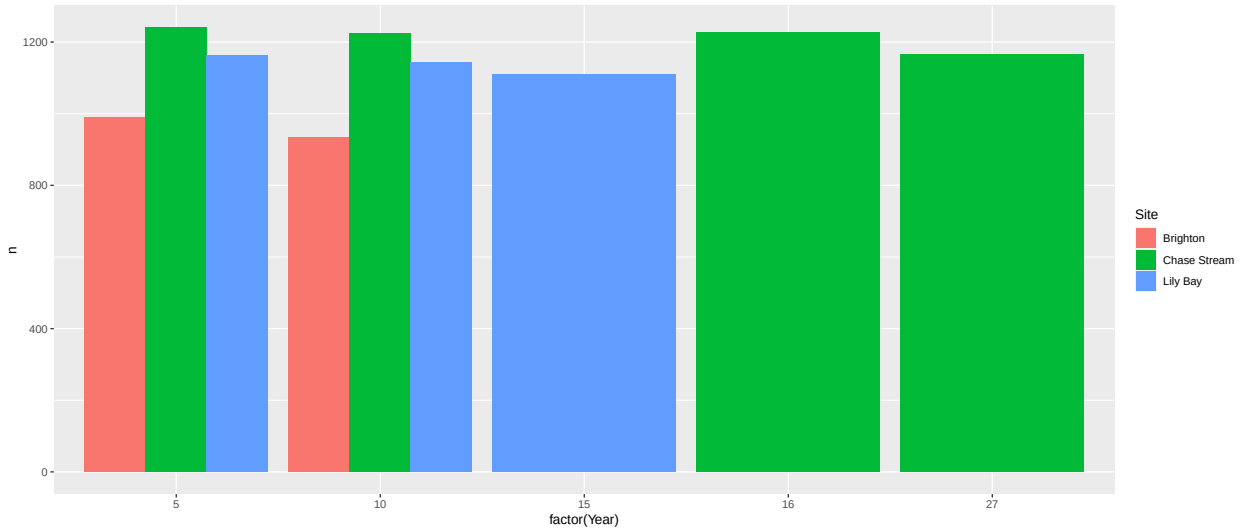


The measurements are collected few years apart: at 5, 10, 15, 16 and 27 years.

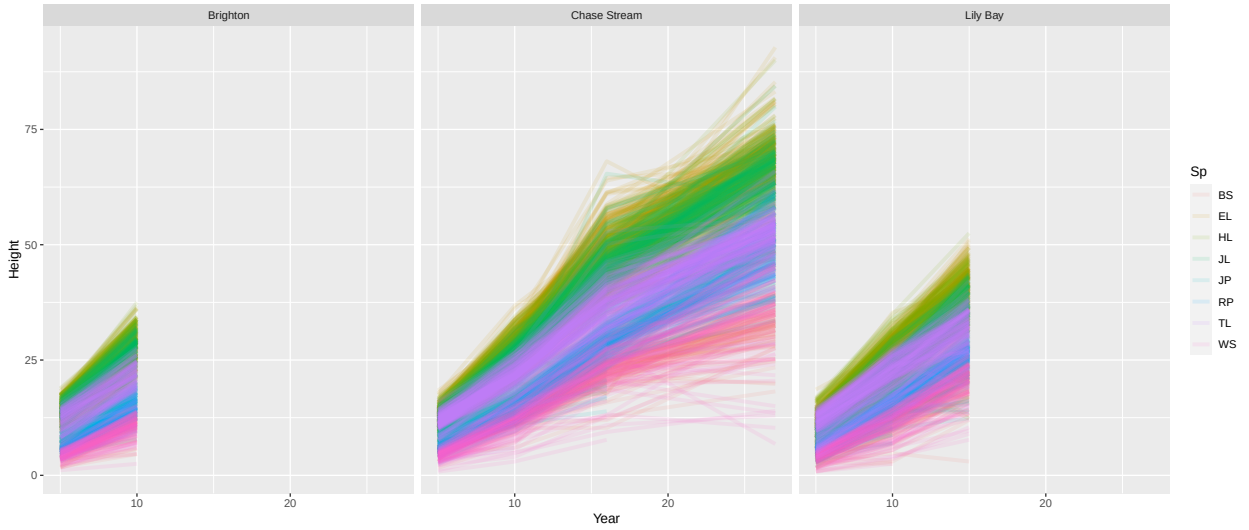
```
sumDf <- tree %>% group_by(Site, Year)  %>%  summarize(n = n())
sumDf %>% kable()
```

Site	Year	n
Brighton	5	989
Brighton	10	933
Chase Stream	5	1240
Chase Stream	10	1224
Chase Stream	16	1226
Chase Stream	27	1166
Lily Bay	5	1162
Lily Bay	10	1144
Lily Bay	15	1109

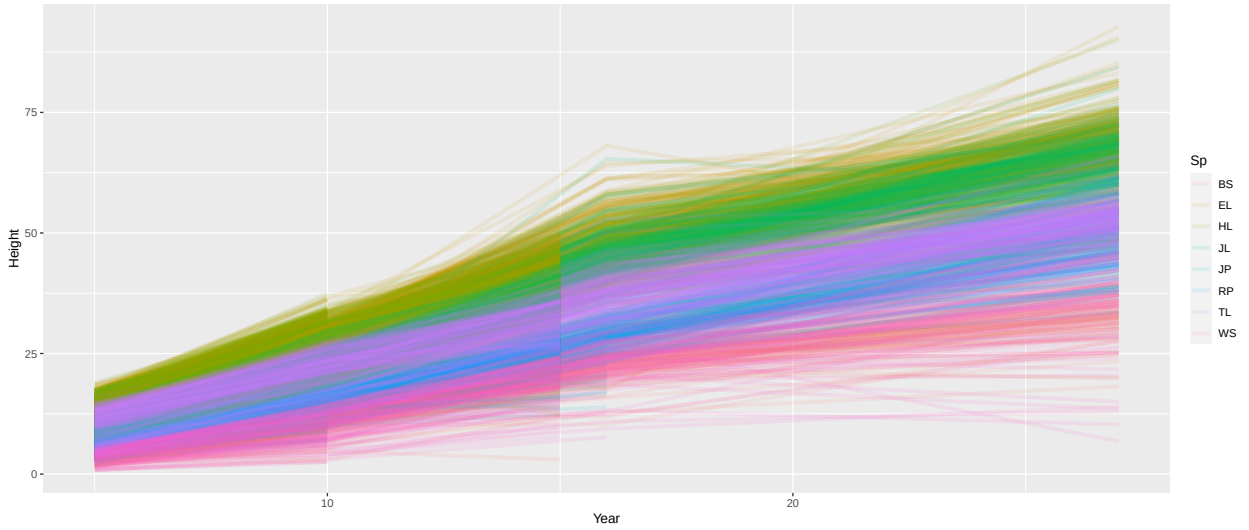
```
ggplot(sumDf, aes(x = factor(Year), y = n, fill = Site)) + geom_col(position = 'dodge', stat = 'count')
```




```
ggplot(data = tree, aes(x = Year, y = Height, col = Sp, group = id)) +
  geom_line(size = 1.5, alpha = .1) + facet_wrap(~Site)
```




```
ggplot(data = tree, aes(x = Year, y = Height, col = Sp, group = id)) +
  geom_line(size = 1.5, alpha = .1)
```

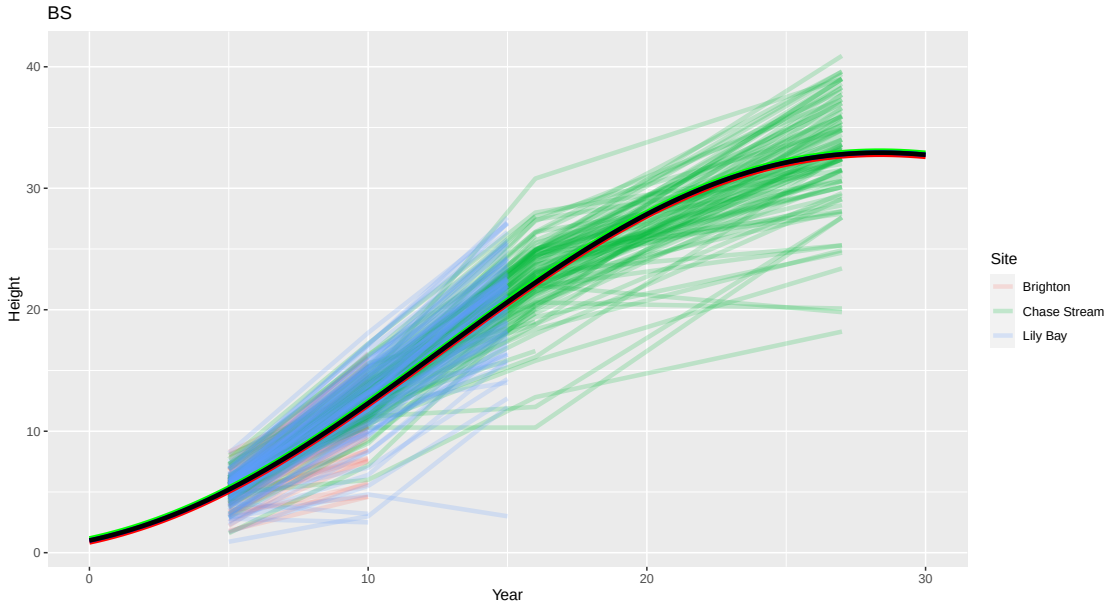


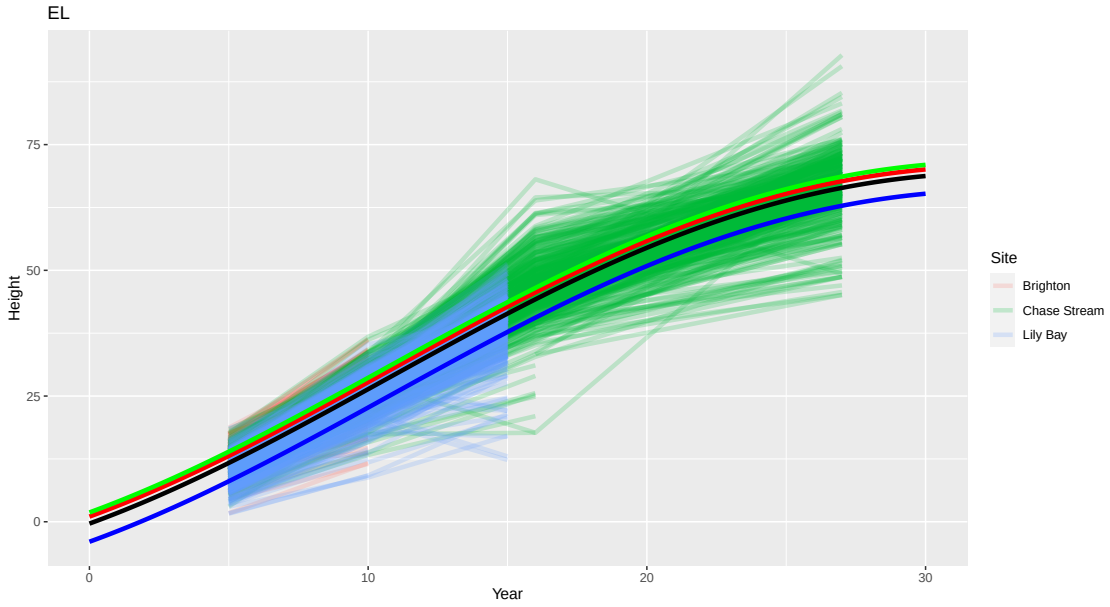
GP model: Site Specific Mean Height Estimation

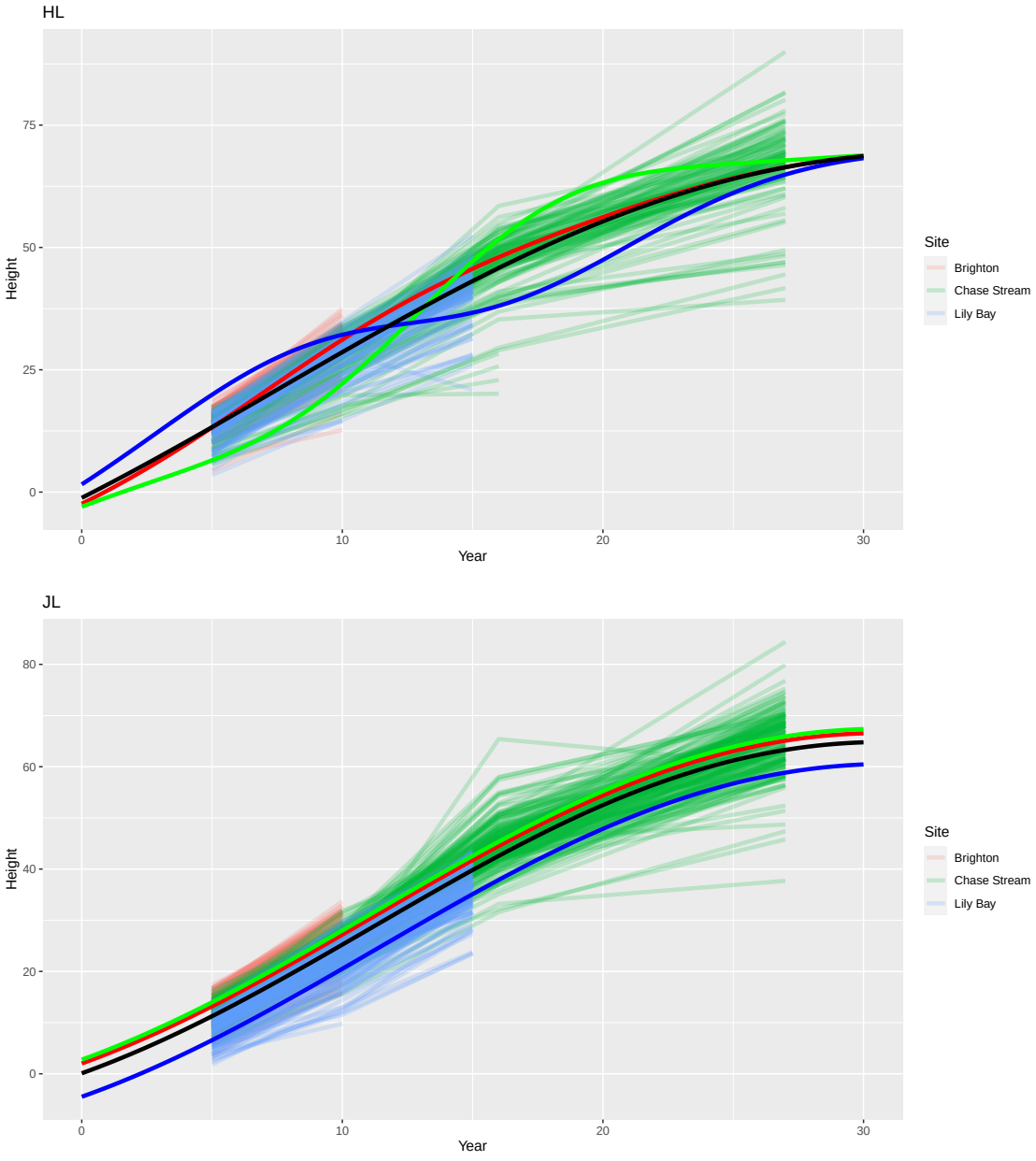
Parameter estimates

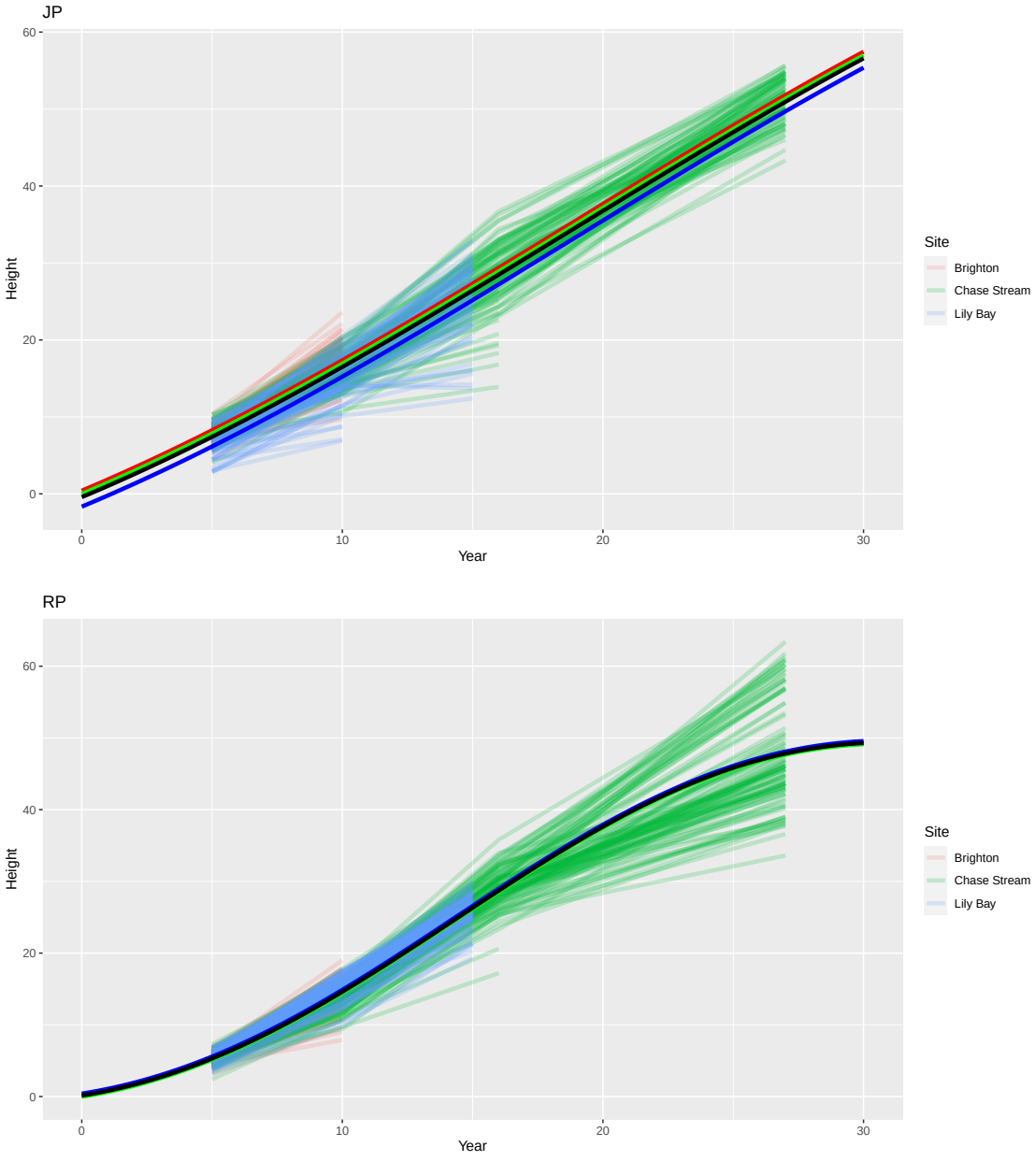
	Mean		Subject		Site		
	l^2	σ^2	l^2	σ^2	l^2	σ^2	σ^2
BS	0.69	1.79	0.17	0.06	0.14	0.00	0.03
EL	1.06	4.80	0.20	0.05	0.08	0.01	0.02
HL	2.09	5.96	0.27	0.07	11.55	0.01	0.02
JL	1.69	12.24	0.21	0.03	0.19	0.02	0.01
JP	4.36	14.42	0.18	0.02	0.10	0.01	0.02
RP	0.86	2.50	0.16	0.05	0.01	0.00	0.01
TL	0.28	1.60	0.20	0.04	0.01	0.00	0.02
WS	0.67	2.08	0.20	0.16	24.91	0.00	0.01

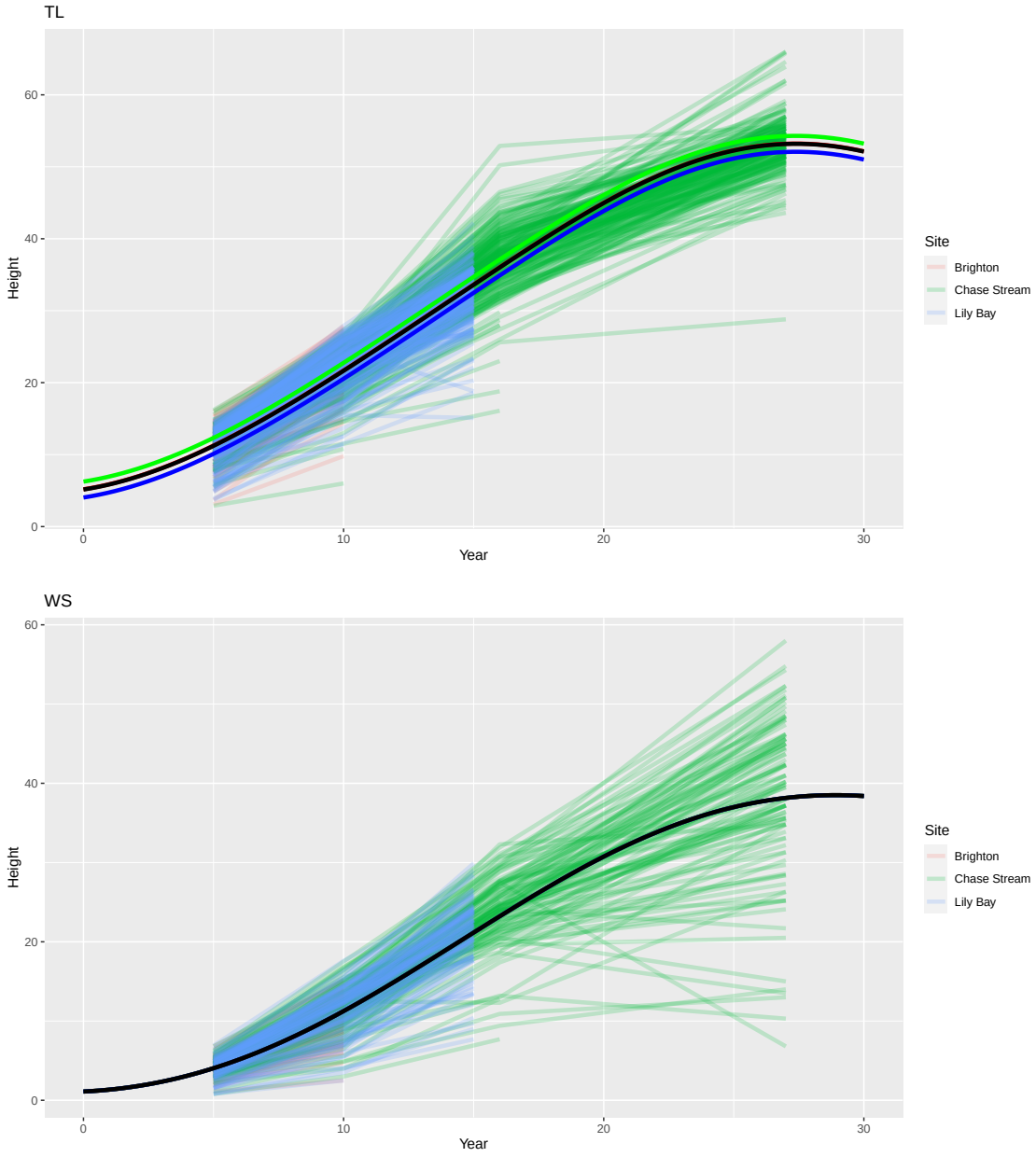
Plots of site specific estimation






HL
Height
75
50
25
0
0
10
20
30
Year
Site
Brighton
Chase Stream
Lily Bay
JL
Height
80
60
40
20
0
0
10
20
30
Year
Site
Brighton
Chase Stream
Lily Bay


JP
Height
Year
Site
Brighton
Chase Stream
Lily Bay
RP
Height
Year
Site
Brighton
Chase Stream
Lily Bay



Classification

We further predict the tree species based on their growth (height) over time. The classification result is summarized in the following table in terms of a confusion matrix. The overall accuracy for prediciting the tree species is 46.14% which is not very high. The table also shows the percentage of trees correctly classified given the species. It appears that TL trees are classified best (71% classified as TL) whereas the larches are somewhat difficult to classify (EL: 27%, JL: 24%). However, interestingly the trees are

We can aggregate the prediction results to compute the accuracy when trying to predict a tree's genus rather than the species. In this case, the overall accuracy jumps to 86.32%. Therefore, our model performs much better at predicting a tree's genus than the exact species.

Table 1: Confusion matrix for predicting tree species. The percentages in the braces give the accuracy of predicting the species given a species.

				Predicted					Total
True	BS	EL	HL	JL	JP	RP	TL	WS	
BS	149 (54.38)	0	0	0	19	11	0	95	274
EL	15	258 (27.27)	377	156	9	8	105	18	946
HL	7	37	190 (66.20)	17	6	3	24	3	287
JL	2	149	118	116 (24.47)	4	13	63	9	474
JP	43	0	0	4	160 (55.56)	25	40	16	288
RP	28	0	0	7	66	138 (46.00)	4	57	300
TL	15	23	7	40	41	12	379 (71.11)	16	533
WS	67	0	0	0	11	45	0	182 (59.67)	305

Table 2: Confusion matrix for predicting tree genus The percentages in the braces give the accuracy of predicting the genus given a species.

		Predicted		Total
True	Spruce	Larch	Pine	
BS	244 (89.05%)	0	30	274
EL	33	896 (94.71%)	17	946
HL	10	268 (93.38%)	9	287
JL	11	446 (94.09%)	17	474
JP	59	44	185 (64.24%)	288
RP	85	11	204 (68.00%)	300
TL	31	449 (84.24%)	53	533
WS	249 (81.64%)	0	56	305