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

MONTANA STATE UNIVERSITY
DEPARTMENT OF MATHEMATICAL SCIENCES

TRANSITION PROBABILITIES:
Whitebark Pine and Blister Rust in the Greater Yellowstone Ecosystem

June 30, 2017

Author: Michaela Powell
Advisor: Steve Cherry

Transition Probabilities

Whitebark Pine and Blister Rust in the Greater Yellowstone Ecosystem

Michaela Powell

June 30, 2017

Abstract

Whitebark Pine (*Pinus albicaulis*) plays an integral role in the Greater Yellowstone Ecosystem (GYE). It has been often referred to as a keystone species due to the extensive changes that the ecosystem would undergo if the species were to be lost. Attributed to white pine blister rust (*Cronartium ribicola*) infection and mountain pine beetles (*Dendroctonus ponderosae*), the decline of the whitebark population in the GYE over the last few decades led to the establishment of the Interagency Whitebark Pine Monitoring Program in 2004. Using data collected by the program from 2008 through 2016, we focused on white pine blister rust infections, using mixed logistic regression to model the probability of transitioning from no infection to a canopy infection (Model 1) and modeling the probability of transitioning from a canopy infection to a bole infection (Model 2). Explanatory variables included elevation, aspect, tree size, and a time variable - the length of time the tree remained uninfected (Model 1) or the length of time the tree remained canopy infected (Model 2). We found that the length of time the tree remined uninfected and tree size were the most useful predictors in the probability of transitioning from no infection to a canopy infection; while the length of time the tree remained canopy infected and elevation were the most useful predictors in the probability of transitioning from a canopy infection to a bole infection.

Contents

Introduction **3**

Methods **4**
Data Collection 4
Variables of Interest 4
Generalized Linear Mixed Models 12

Results **13**
Model 1: No infection to a canopy infection 13
Model 2: Canopy infection to a bole infection 16

Discussion **19**

Appendix **21**
Diagnostics for Model 1 21
Diagnostics for Model 2 21

References **27**

Introduction

Whitebark pine (*Pinus albicaulis*) are conifers native to the subalpine zone[1] of the Pacific Northwest[2] and northern Rocky Mountains. Often gnarled and twisted from the extreme environments in which they grow, they are low in commercial value; however, their ecological importance is invaluable. Growing on steep terrain and in poor soils, these trees facilitate snow retention, necessary to the survival of many of the species throughout their ecosystem, and create microhabitats for the establishment of plant species less adapted to such environments. Their pine cone seeds serve as a high-energy food source for species such as the Clark's nutcrackers (*Nucifraga columbiana*), American red squirrels (*Tamiasciurus hudsonicus*), and North American brown bears (*Ursus arctos ssp.*). Within the Greater Yellowstone Ecosystem[3], the cone production of the whitebark pines serves as the primary indicator of the brown bear population's survivorship and reproductivity each year (Greater Yellowstone Whitebark Pine Monitoring Working Group 2011). In the past few decades, the whitebark pine population has suffered drastic declines across its natural range, attributed to reasons such as climate change, wildfires, mountain pine beetle (*Dendroctonus ponderosae*) infestations, and, the focus of this analysis, white pine blister rust.

White pine blister rust is caused by a non-native fungal pathogen (*Cronartium robicola*) originating in Asia. After being established in Europe in the 18th century via the planting of very susceptible American white pines, the disease spread to North America in the early 1900s when seedlings grown in European nurseries were imported into the eastern United States (Maloy 2001). In 1926, blister rust infection was first discovered on whitebark pine; by 1945, it had spread across the GYE, reaching northwestern Montana in 1927, southern Oregon in 1929, Glacier National Park in 1939, and southern Idaho in 1945 (Graphics, n.d.).

The infection requires to host species, bouncing between the whitebark pine and, most commonly, a gooseberry plant (*Ribes ssp.*). The concern, however, lies with the whitebark, because while the "rust is shed from the gooseberry plant when the plant naturally drops its leaves in the autumn... one successful rust infection of a pine can persist and expand for years", with the potential to eventually kill the tree (Graphics, n.d.). The infection can occur in the canopy of the tree or the bole of the tree[4], with the infection in the bole being considered more severe because of the higher threat of mortality. Due to the ecological importance of the species, described above, the spread of the infection has evoked grave concern. In this analysis, we wish to determine the predictors most valuable in the modeling transition probabilities of the blister rust infection. We will consider two of the possible transitions - the first from no infection to a canopy infection and the second from a canopy infection to a bole infection.

[1]The subalpine zone refers to the biotic zone immediately below tree line in the PNW and Rocky Mountains.

[2]The Pacific Northwest a geographic region whose boundaries are defined by the Pacific Ocean on the West and the Rocky Mountains on the East.

[3]The Greater Yellowstone Ecosystem (GYE) is 34,375 square miles spanning across northwestern Wyoming, southwestern Montana, and eastern Idaho. It is "one of the largest nearly intact temperate-zone ecosystems on Earth" ("Greater Yellowstone Ecosystem," n.d.).

[4]The bole of the tree is where the trunk splits into the branches that form the canopy of the tree.

Methods

Data Collection

The Greater Yellowstone Whitebark Pine Monitoring Working Group (GYWPMWG), comprised of affiliates of the U.S. Forest Service (USFS), the Bureau of Land Management (BLM), the National Park Service (NPS), the U.S. Geological Survey (USGS), and Montana State University (MSU), was established as a long-term monitoring program of Whitebark Pine in the Greater Yellowstone Ecosystem. From 2004 to 2007 the different organization began their collaborative work, establishing the sampling units for future field surveys.

A two-stage cluster design was used. In the first-stage, a simple random sample of 150 "whitebark pine dominated stands of approximately 2.5 hectares or larger" was taken from the population of 10,770 stands identified in the GYE. In the second-stage, a sample of ten-by-fifty meter transects was taken from within the randomly sampled stands. To sample the transects, a simple random sample of five points was assigned to each of the sampled whitebark pine stands. The first point of the sample was used as the "targeted mid-point" of the transect to be sampled. A vector originating from the point was then randomly selected to determine the ten-by-fifty meter transect. If no whitebark were captured by the transect, the next sampled point was used as the "targeted mid-point", and so on. During the four-year establishment period, 176 transects were sampled and all whitebark pine trees 1.4 meters or taller with a diameter at breast height[5] (DBH) of at least one centimeter were permanently marked (Greater Yellowstone Whitebark Pine Monitoring Working Group 2016). For 124 of the stands only a single transect was sampled, however in the remaining 26 stands a second transect was sampled to facilitate the exploration of within stand variability.

Visiting transects is logistically challenging and blister rust infection spreads very slowly so a "rotating panel" sampling structure was used; the sampled transects were randomly assigned to one of four panels. Each year field surveys were conducted on predetermined panels, with all four panels being surveyed by the end of a four-year time step. The survey conducted was either a "full survey", in which both blister rust and life status were observed, or a "status survey", in which signs of mountain pine beetle and life status were observed. Our analysis focused on the data from only the full surveys. Field surveys were conducted in 2016, the first year of T3, however we will not include these observations as the data for the full four-year time step are not available. Figure 1 summarizes the schedule of the surveys.

Variables of Interest

There is subjectivity in the identification of blister rust infection, so the status of blister rust (in either the canopy or the bole) was recorded as "definitely infected", "probably infected", and "uninfected". These were then translated into a binary variable by merging "definitely infected" and "probably infected" into the single category: "infected" (Greater Yellowstone

[5]Diameter at breast height is used as a measure of the tree's age, with breast height in this study being defined as 1.4 meters from the ground.

Panel	Transects Per Panel	T0	T1				T2			
		2004 - 2007	2008	2009	2010	2011	2012	2013	2014	2015
1	43	Establishment Period: Surveys over all 176 transects	FULL		STATUS		FULL			
2	45			FULL		STATUS		FULL		
3	44		STATUS		FULL				FULL	
4	44			STATUS		FULL		FULL		FULL

Figure 1: Visitation schedule for rotating panel sampling design.

Whitebark Pine Monitoring Working Group 2011). The result was a binary response variable (TransitionCanopy) for Model 1, and a second binary response variable (TransitionBole) for Model 2.

Discussions with agency personnel identified three variables of interest for modeling both types of transition probabilities: DBH, Elevation, and Aspect.

- DBH is the diameter at breast height of the tree measured in centimeters.
- Elevation is the elevation of the tree measured in meters.
- Aspect is the cardinal direction that the slope surrounding the tree is facing, or, equivalently, the cardinal direction that water would flow from the tree, measured in radians.

Two temporal explanatory variables were defined: the first measuring the number of years that the tree was infection free prior to transitioning to a canopy infection (TimeUninfected) and the second measuring the number of years that a tree had a canopy infection before transitioning to a bole infection (TimeCanopyInfected).

TransitionCanopy and TimeUninfected were constructed simultaneously to be used in the first model, and TransitionBole and TimeCanopyInfected were constructed simultaneously to be used in the second model. This simultaneous construction is necessary, because number of observations differs depending upon the model being considered.

To construct each pair of variables, four cases had to be considered - observations with no NAs[6], observations with NAs in T0, observations with NAs in T1, and observations with NAs in T2. Observations with NAs in more than one time-step were not considered because there is not enough information to construct either pair of variables. Figures 2, 3, 4, and 5

[6]NAs exist in the data due to reasons such as, but not limited to, a tree having a DBH of less than one centimeter during the establishment period, the tree being missed in a field survey, or the tree dying.

describe the construction of TransitionCanopy and TimeUninfected under each of the four cases, and, similarly, Figures 6, 7, 8, and 9 describe the construction of TransitionBole and TimeCanopyInfected under each of the four cases. For the transition variables, a 0 indicates no transition, a 1 indicates a transition, and a -1 indicates that the observation was removed from the dataset for that particular model. For the time to transition variables the numerical value indicates the number of years that the tree remained uninfected (Model 1) or canopy infected (Model 2), with a -1 also indicating that the observation was removed from the dataset for that particular model. In Figure 6, there are two weights of lines with values for the variables in either standard-font or bold-font to correspond to the differently weighted lines. These represent two different paths that met the criteria for the TransitionBole and TimeCanopyInfected variables in the case of no NAs. After the construction of both pairs of variables, 2,805 observations remained for Model 1 and 421 observations remained for Model 2.

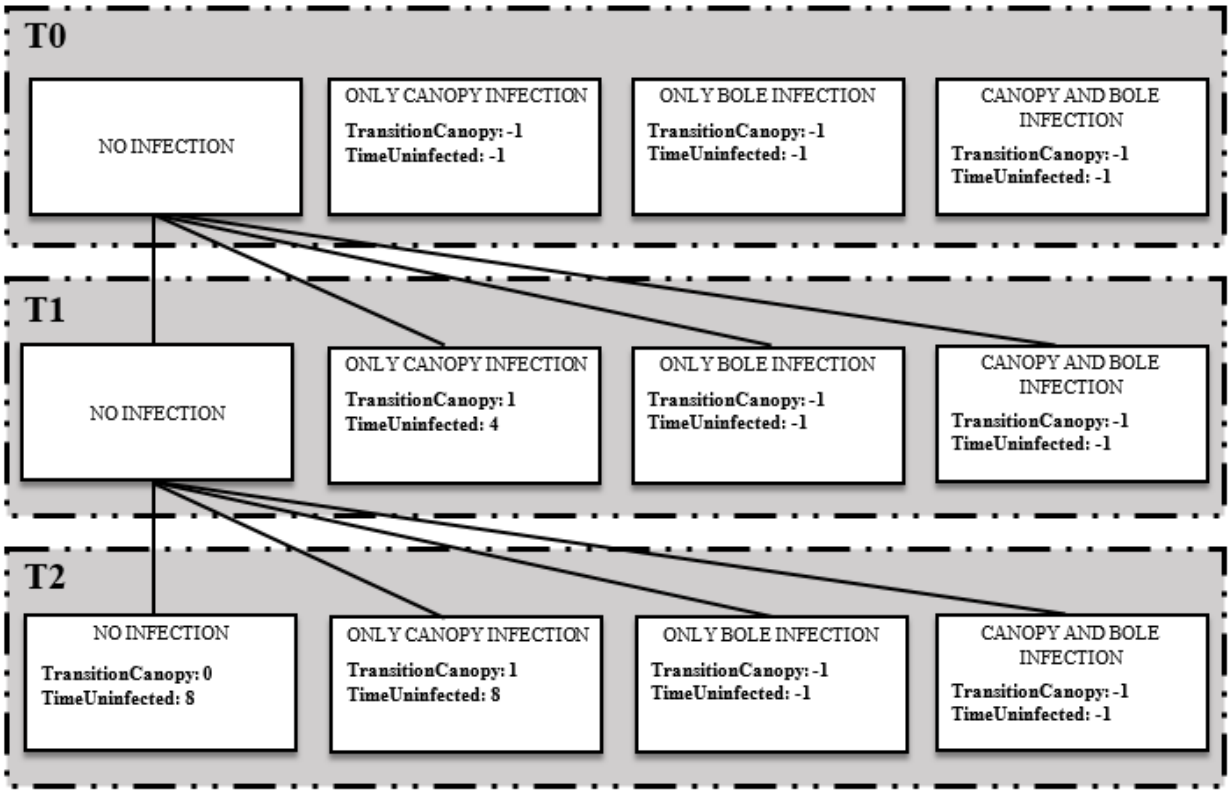

Figure 2: Construction of TransitionCanopy and TimeUninfected for observations with no NAs.

Aspect is a circular variable and standard statistical summary measures do not apply. To illustrate this, consider a case when there are two observed directions, 5 degrees and 355 degrees. A standard summary statistic would be the mean, in this case 180 degrees; however, this is clearly a poor summary measure, as the two directions, in practice are only 10 degrees apart. For this reason, we cannot simply include aspect as a linear predictor in our models.

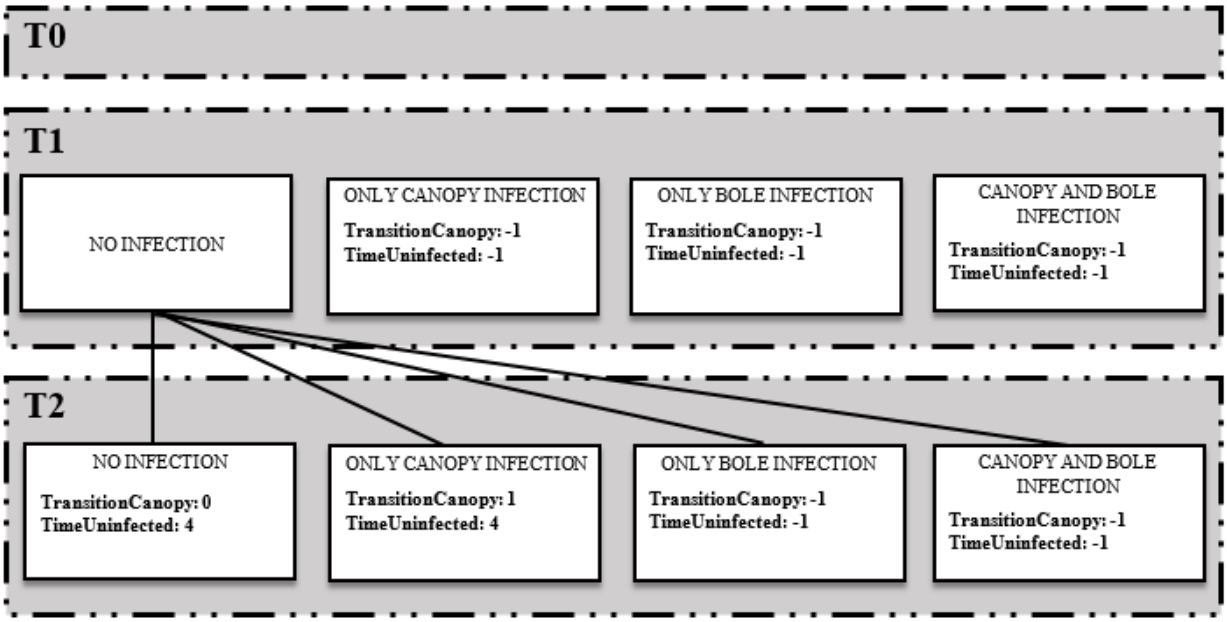

Figure 3: Construction of TransitionCanopy and TimeUninfected for observations with NAs in T0.

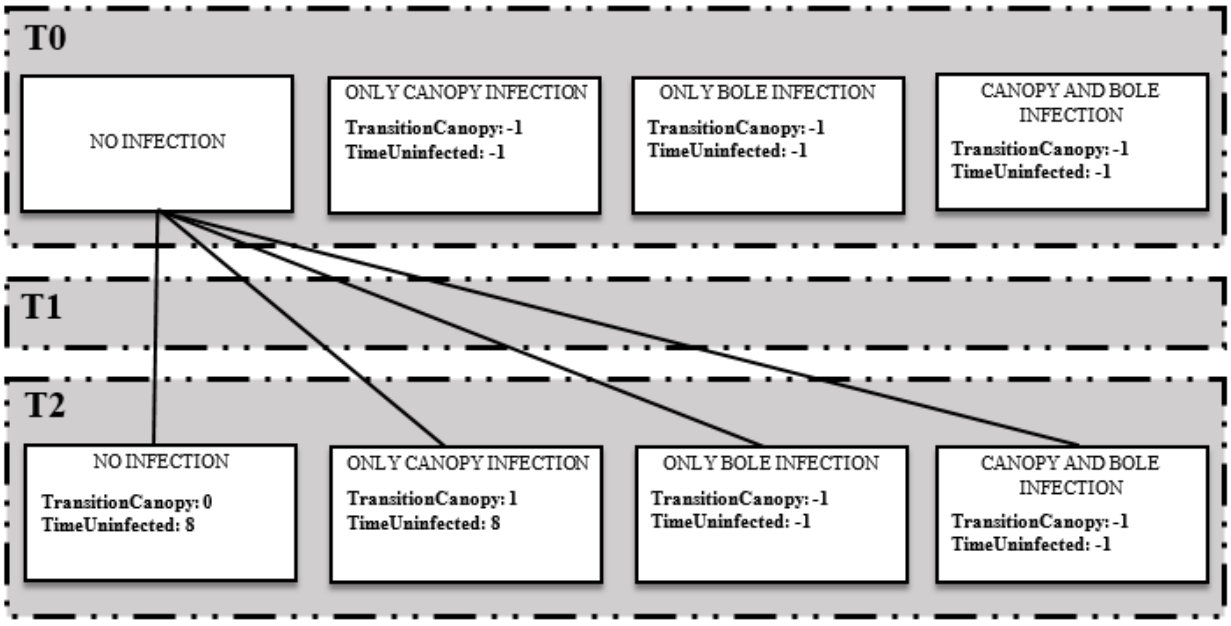

Figure 4: Construction of TransitionCanopy and TimeUninfected for observations with NAs in T1.

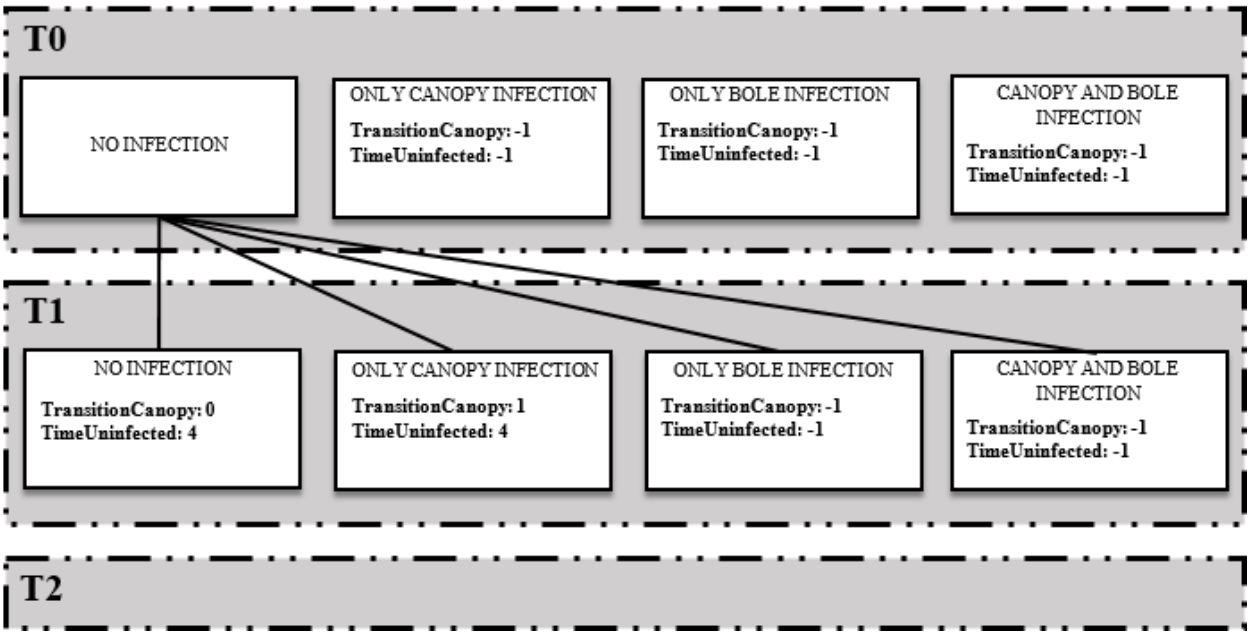

Figure 5: Construction of TransitionCanopy and TimeUninfected for observations with NAs in T2.

A well-accepted solution (Fisher 1993) is to instead include both the sine and cosine of aspect as predictors in the model. This can be explained using the relationship between the polar coordinate system and the rectangular coordinate system. Recall that:

- The polar coordinate system describes a point on a plane with (r, α), where r is the distance to the origin and α is the angle from the origin.
- The rectangular coordinate system describes a point on a plane with (x, y), where x is the horizontal distance from the origin and y is the vertical distance from the origin.

Consider Figure 10 (Al-Daffaie and Khan 2017).

We see that we can easily move from polar coordinates to rectangular coordinates (and vice versa) using:

$$x = rcos(\alpha)$$

$$y = rsin(\alpha)$$

In the context of circular variables, magnitude is not of interest, so we let $r = 1$, and we have:

$$x = cos(\alpha)$$

$$y = sin(\alpha)$$

Therefore, we can describe the circular angle measurement with the two continuous variables formed by the translation from the polar to rectangular-coordinate system (Al-Daffaie and

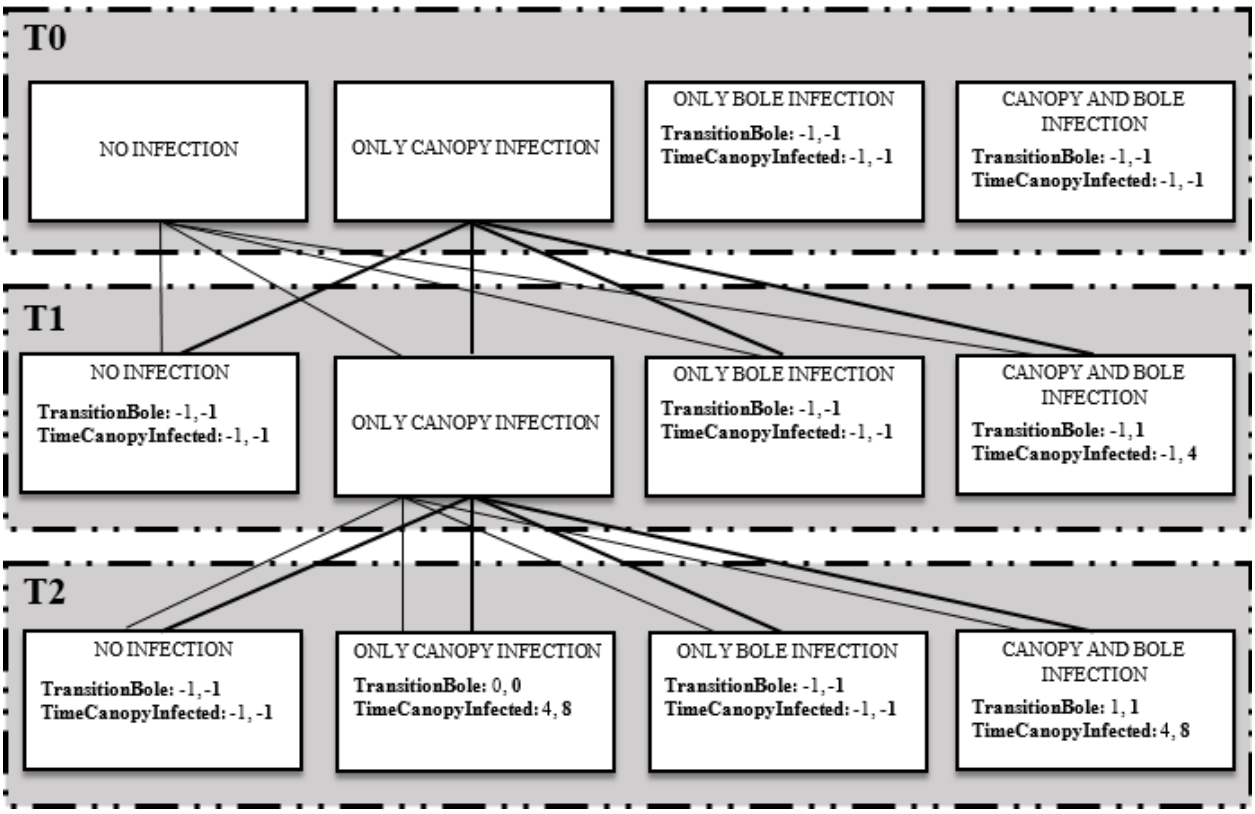

Figure 6: Construction of TransitionBole and TimeCanopyInfected for observations with no NAs.

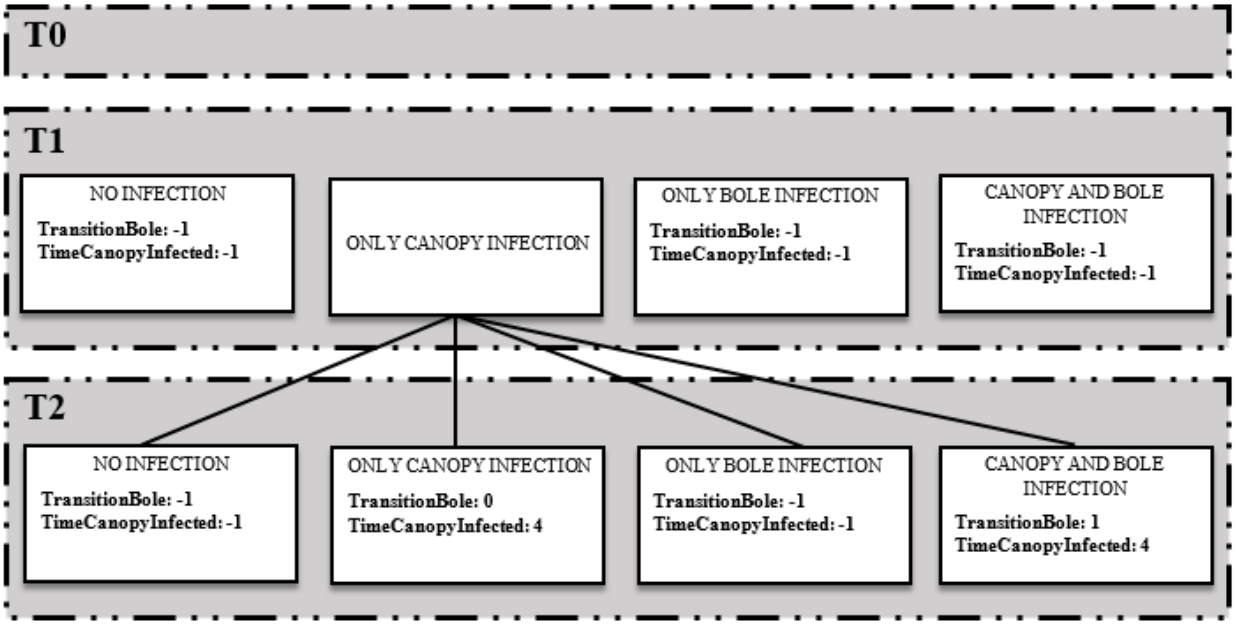

Figure 7: Construction of TransitionBole and TimeCanopyInfected for observations with NAs in T0.

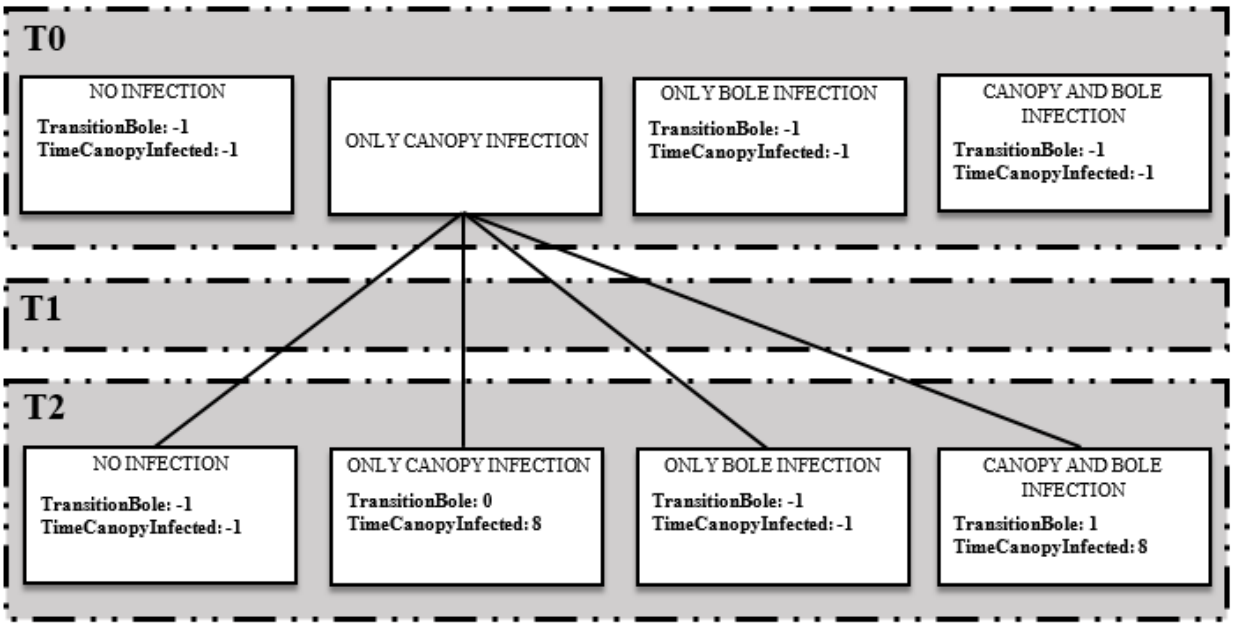

Figure 8: Construction of TransitionBole and TimeCanopyInfected for observations with NAs in T1.

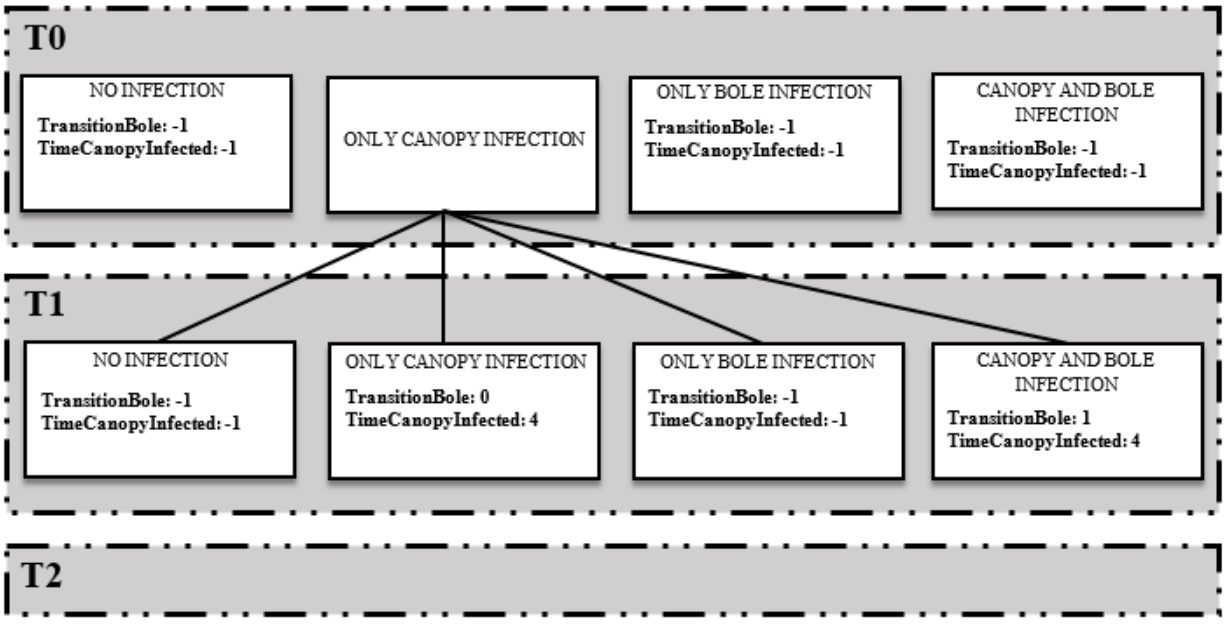

Figure 9: Construction of TransitionBole and TimeCanopyInfected for observations with NAs in T2.

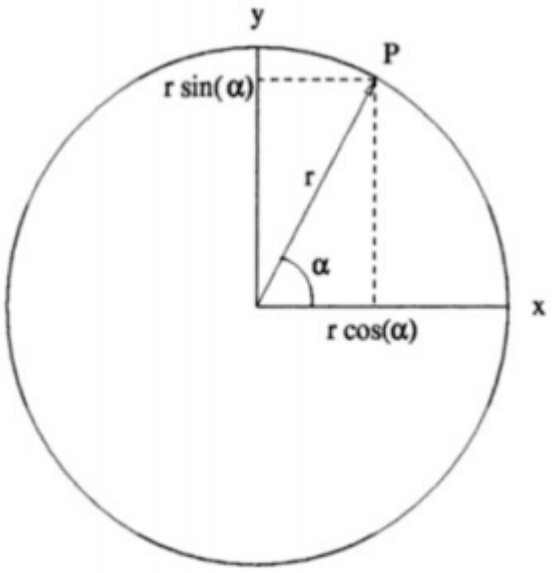

Figure 10: Relationship between polar and rectangular coordinate systems.

Khan 2017). It is important to recognize that the angle, α, is described by both $x = sin(\alpha)$ *and* $y = cos(\alpha)$, so, although they will be fit as two separate predictors, when it comes to model selection they must be added/removed as a pair.

Generalized Linear Mixed Models

An "ordinary linear regression model uses linearity to describe the relationship between the mean of the response variable and a set of explanatory variables" when the response is assumed to be normally distributed. In the context of modeling transition probabilities, the response is binary - indicating transition or no transition - and follows a binomial distribution. Thus, we turn to generalized linear models which "extend standard linear regression models to encompass non-normal response distributions" (Agresti 2015), specifically members of the exponential dispersion family of which the binomial distribution belongs.

A generalized linear model is comprised of three components: the random component, the linear predictor, and the link function.

- The *random component* defines the response and its distribution:

$$\boldsymbol{y} = (y_1, ..., y_n)^\intercal$$

- The *linear predictor* describes the linear relationship of the predictors:

$$\boldsymbol{X\beta} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

- The *link function*, g, relates the mean of the response to the linear predictor:

$$g(E[Y]) = \boldsymbol{X\beta}$$

Like an ordinary linear regression model, in the context of GLMs it is still assumed that the observations are independent. Considering the data being analyzed, this assumption is not reasonably satisfied due to the cluster design of the sampling units. Recall that the individual whitebark pine trees from which the variables were measured are clustered by the randomly sampled transects. Thus it unreasonable to assume that all of the trees are independent, because some trees are related by transect.

To account for this correlation among the responses, we can model the transects through the inclusion of a random effect (Agresti 2015). A generalized linear model with a random effect (or random effects) is called a generalized linear mixed model (GLMM). This model is defined by the same three components of a GLM; however, they are modified to accommodate the inclusion of the random effect.

- The *random component* defines the response and its distribution conditional on the random effects $\boldsymbol{u}$:

$$\boldsymbol{y} = (y_1, ..., y_n)^\intercal$$

- The *linear predictor* describes the linear relationship of the fixed and random effects:

$$\boldsymbol{X\beta} + \boldsymbol{Zu}$$

- The *link function*, g, relates the mean of the response conditional on the random effects to the linear predictor:

$$g(E[Y]) = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

It is assumed that the observations, are independent conditional on the random effects, which is reasonable to assume for the data analyzed. It is also assumed that the random effects are normally distributed with a mean of zero and constant variance.

Having determined that a GLMM is appropriate, a link function appropriate to the distribution of the response must be chosen. In this analysis, the response for both transitions of interest are binary, $y_i|u_i \sim Binomial(\pi_{ij})$. Two link functions are typically used for binary response variables: the logit and probit. We will use the logit link as it is more appropriate for our inferential purposes, however "in practice, probit and logistic regression models provide similar fits" (Agresti 2007, 72). Using the logit link function, $log(\frac{\pi_{ij}}{1-\pi_{ij}})$, we can refer to our GLMM as a logistic mixed model.

Results

Model 1: No infection to a canopy infection

The transition from no infection to a canopy infection was considered first. For this model the response was the canopy transition binary variable, and the set of predictors considered was length of time the tree was uninfected in years, elevation in kilometers[7], DBH in centimeters, the sine of the aspect variable in radians, and the cosine of the aspect variable in radians. When determining all possible linear subsets of the main effects of the predictors, there are only four predictors in practice as the two aspect variables must be added and removed as a pair. This equates to only fifteen linear subsets, so the logistic mixed model for each subset was considered. Only the main effects of the predictors were considered in the models, because there were no apriori expectations of interactions or quadratic effects.

Using the glmer() function in the lme4 package (Bates et al. 2015), R studio (RStudio Team 2015) was used to fit each of the candidate models.

The Akaike Information Criterion (AIC) was of primary consideration in the selection of the top candidate models and the selection of the final model used to make inferences. This criterion evaluates the fit of a model by comparing simpler models to the most complex or

[7]Upon fitting the most saturated model, it was discovered that the glmer() function was unable to fit a model including elevation as recorded in meters. The measurements of this variable are so large in comparison to those of the other variables, identifiability issues arose when fitting the model. As a result, the variable was converted to kilometers.

full model in the set of candidate models, with lower AIC indicating a better fit. However, this criterion must be used with caution as the "best" model in the set based upon AIC may fit poorly if all of the models in the set fit poorly. The ΔAIC is often considered in the place of the AIC as it is easier to understand in the context of how models compare across a set. It is calculated by subtracting the smallest AIC from the set of candidate models from the AIC of the candidate model in consideration.

model	time uninfected	elevation	DBH	sin(aspect)	cos(aspect)	ΔAIC_i	w_i
A	✓		✓			0.00	0.5352
B	✓	✓	✓			1.04	0.3181

Table 1: Summary of top candidate models for the probability of transitioning from no infection to a canopy infection.

The last column of Table 1 includes the model weights, which can be understood as "the weight of evidence in favor of model I being the actual... best model for the situation at hand given that one of the R models must be the... best of that set of R models" (Burnham and Anderson 2002). We choose to focus only on these top two candidate models because just over 85% of the weight is held by these models, out of the entire set of fifteen models[8].

The assumptions were checked and found to be reasonably satisfied for both models; the details of which can be found in the Appendix. Table 2 reports the parameter estimates of each model followed by the associated standard errors in parentheses. We see that the parameter estimates for the variables common to both models are similar in both sign and magnitude with identical standard errors. σ, the estimated variance of the random transect effect, is also similar, differing by only 0.0030 between the two models.

parameter	Model A	Model B
intercept	2.9082 (0.3159)	5.4391 (2.6159)
time uninfected	−1.0126 (0.0547)	−1.0107 (0.0547)
elevation		−0.9077 (0.9294)
DBH	0.0721 (0.0085)	0.0728 (0.0085)
sin(Aspect)		
cos(Aspect)		

[8]The next best candidate model had a weight of 0.0894.

parameter	Model A	Model B
σ	1.3150	1.3120

Table 2: Summary of parameter estimates and standard errors for Model A and Model B.

Before focusing on the "best" model to make inferences, we explore the predictive power of the top two candidate models by testing their ability to predict the observed transitions (or lack of) in the original data. With the models predicting probabilities of transitioning (values ranging anywhere between 0 and 1), a threshold must be chosen as to how high a predicted probability must be for that tree to be considered as having transitioned. We will use the threshold as indicated by the data[9], calculated to be 0.1141, and consider the sensitivity[10] and specificity[11] of the two models. Table 3 summarizes the results. With the same sensitivity and nearly the same specificity, the difference in the predictive power of these models is negligible.

model	sensitivity	specificity
A	0.8531	0.8571
B	0.8531	0.8567

Table 3: Sensitivities and specificities of top candidate models for the probability of transitioning from no infection to a canopy infection.

Using Model A (the "best" model as judged by AIC), we will make inferences about the effect of the variables on the odds of transitioning. With the probability of success defined as π, the odds of success are $\frac{\pi}{1-\pi}$ (Agresti 2007, 28).

Our theoretical model is:

[9]We will determine the proportion of trees that did transition from no infection to a canopy infection and use this proportion as the threshold. Any predictions with a probability of transitioning less than this threshold will be considered a failure (the tree did not transition) and any probability of transitioning greater than this threshold will be considered a success (the tree did transition).

[10]Sensitivity is the true positive rate - of the trees that did transition, the proportion of those trees that were predicted to have transitioned.

[11]Specificity is the true negative rate - of the trees that did not transition, the proportion of those trees that were predicted to have transitioned.

$$\log(\frac{\pi_{ij}}{1-\pi_{ij}}) = (\beta_0 + u_{ij}) + \beta_1 TimeUninfected_{ij} + \beta_2 DBH_{ij}$$

We see that exponentiating both sides gives us an equation relating the parameter estimates to the odds (Agresti 2007, 104):

$$\frac{\pi_{ij}}{1-\pi_{ij}} = e^{(\beta_0+u_{ij})}(e^{\beta_1})^{TimeUninfected_{ij}}(e^{\beta_2})^{DBH_{ij}}$$

Therefore, the exponentiated parameter estimates can be interpreted as a multiplicative change in the odds specific to the cluster u_j, conditional on the other variables in the model. Table 4 summarizes the exponentiated estimates and approximate 95% confidence intervals for the fixed effects of Model A. The units of the variables are years and centimeters, respectively.

variable	lowerbound	point estimate	upperbound
time uninfected	0.1956	0.3633	0.6747
DBH	1.0570	1.0748	1.0928

Table 4: Exponentiated approximate 95% confidence intervals and parameter estimates for the fixed effects.

It is very important to remember that these estimates are transect specific, as indicated in the following interpretations of the confidence intervals.

1. We are 95% confident that within a given transect each additional year that a tree remains uninfected is associated with between a 32.53% and 80.44% decrease in the likelihood that the tree will transition from having no infection to having a canopy infection, after accounting for the DBH of the tree.
2. We are 95% confident that within a given transect, each additional centimeter in the DBH of a tree is associated with between a 5.70% and 9.28% increase in the likelihood that the tree will transition from having no infection to having a canopy infection, after accounting for the length of time the tree has remained uninfected.

Model 2: Canopy infection to a bole infection

The transition from a canopy infection to a bole infection was considered next. For this model the response was the bole transition binary variable, and the set of predictors considered was length of time the tree was canopy infected in years, elevation in kilometers, DBH in centimeters, the sine of the aspect variable in radians, and the cosine of the aspect variable in

radians. Again, there are only four predictors in practice, equating to fifteen linear subsets, so the logistic mixed model for each subset was considered. Only the main effects were included for the same reason as explained for Model 1.

Using the same packages and functions in R studio (RStudio Team 2015) as with the first model, each of the candidate models was fit. Similar to Model 1, the AIC was of primary consideration in the selection of the top candidate models and the selection of the final model used to make inferences. Table 5 summarizes the top three models ordered from lowest to highest ΔAIC.

model	time canopy infected	elevation	DBH	sin(aspect)	cos(aspect)	ΔAIC_i	w_i
A	✓	✓				0.00	0.5415
B	✓	✓	✓			1.88	0.2117
C	✓	✓		✓	✓	3.20	0.1094

Table 5: Summary of best candidate models for the probability of transitioning from no infection to a canopy infection.

We choose to focus on the top three candidate models because, like the top candidate models in Model 1, just over 85% of the weight is held by these models, out of the entire set of fifteen models[12].

The assumptions were checked and found to be reasonably satisfied for both models; the details of which can be found in the Appendix. Table 6 reports the parameter estimates of each model followed by the associated standard errors in parentheses. As we saw in Model 1, the parameter estimates for the variables common to all three models are similar in both sign and magnitude with similar associated standard errors. σ, the estimated variance of the random transect effect, is also similar across the three models.

parameter	Model A	Model B	Model C
intercept	10.2222 (3.5590)	10.1250 (3.5863)	10.6597 (3.5807)
time canopy infected	−0.3834 (0.07625)	−0.3845 (0.0765)	−0.3824 (0.0760)
elevation	−3.1447 (1.2782)	−3.1401 (1.2849)	−3.2800 (1.2821)
DBH		0.0047 (0.0134)	
sin(aspect)			0.0652 (0.2528)

[12]The next best candidate model had a weight of 0.0594.

parameter	Model A	Model B	Model C
cos(aspect)			0.2622 (0.2996)
σ	1.2180	1.229	1.193

Table 6: Summary of parameter estimates and standard errors for Model A, Model B, and Model C.

We again explore the predictive power of the top candidate models by testing their ability to predict the observed transitions (or lack of) in the original data. We will again use the threshold as indicated by the data, calculated to be 0.3919 for the Model 2 data, and consider the sensitivity and specificity of the two models. Table 7 summarizes the results. All three models have nearly the same sensitivity and specificity, indicating comparable predictive powers.

model	sensitivity	specificity
A	0.8424	0.7227
B	0.8428	0.7344
C	0.8303	0.7305

Table 7: Sensitivities and specificities of top candidate models for the probability of transitioning from a canopy infection to a bole infection.

Using Model A (the "best" model as judged by AIC), we will make inferences, again, about the effect of the variables on the odds of transitioning following the same procedure described for Model 1. Table 8 summarizes the exponentiated estimates and approximate 95% confidence intervals for the fixed effects of Model A. The units of the variables are years and 100 meters[13], respectively.

[13]A one kilometer change in elevation is not meaningful in this ecological context, so the estimates for elevation were re-scaled to be in the units of 100 meters.

variable	lowerbound	point estimate	upperbound
time canopy infected	0.5864	0.6810	0.7908
elevation	0.5684	0.7300	0.9381

Table 8: Exponentiated approximate 95% confidence intervals and parameter estimates for the fixed effects.

Again, it is important to remember that these estimates are transect specific, as indicated in the following interpretations of the confidence intervals.

1. We are 95% confident that within a given transect each additional year that a tree remains canopy infected is associated with between a 20.92% and 41.36% decrease in the likelihood that the tree will transition from having a canopy infection to having a bole infection, after accounting for the elevation of the tree.
2. We are 95% confident that within a given transect, each additional 100 meters in the elevation of a tree is associated with between a 6.19% and 43.16% increase in the likelihood that the tree will transition from having a canopy infection to having a bole infection, after accounting for the length of time the tree has remained canopy infected.

Discussion

While two "final" models were chosen for inferential purposes, the other top candidate models tell similar stories and could justifiably be used for inference as well. For each of the two transitions modeled, there is a consistency in the variables deemed important. For Model 1 (the transition from no infection to a canopy infection) the length of time the tree remained uninfected and DBH of the tree appeared in both of the top candidate models; and, for Model 2 (the transition from a canopy infection to a bole infection) the length of time the tree remained canopy infected and elevation appeared in all three of the top candidate models.

The final model chosen for inference for the transition from a canopy infection to a bole infection did not include DBH. However, the next best candidate model (in terms of ΔAIC) did contain this predictor. When this next best model is used to make inferences, an interesting result arises. Recall, it was found that, for the transition from no infection to a canopy infection, an increase in DBH was associated with an increase in the odds of transitioning. However, in the case of the transition from a canopy infection to a bole infection, an increase in DBH is associated with a decrease in the odds of transitioning. This inverse relationship seems counter-intuitive, however discussions with agency personnel confirmed that this is a well-hypothesized idea that has never been quantified. It is thought that a larger tree has a larger canopy, so it is more likely to catch the spores of the fungus in its foliage, resulting in a canopy infection; but, a larger tree also has more distance between its canopy and bole, so

the infection is less likely to transition from the canopy to the bole. The quantification of this hypothesis is an exciting outcome of this analysis.

There is one main area for improvement in this analysis, involving the two time-infected variables. These were fit as continuous variables, however due to the rotating-panel design of the surveys, they were more discrete than continuous (only taking on the values 0, 4, and 8). They were not included as such because identifiability issues arose when fitting the models. Therefore, caution should be taken when using the interpretation of the parameter estimates for these variable; a one year increase is not something that was ever recorded.

Finally, there are many options for future research. Perhaps the most interesting involves the subjectivity in the identification of blister rust infection. As described previously, the categories of "probably infected" and "definitely infected" were merged into the single binary variable "infected". If models were fit using only the "definitely infected" trees, would the results change? Such knowledge could shed light on how much importance should placed on the results of past research and greatly impact any future research on whitebark pine blister rust infection in the GYE.

Appendix

When using a logistic model with a random intercept, it is assumed that the observations are independent conditional on the random intercept and that the random intercepts are normally distributed with a mean of zero and constant variance. We have previously addressed the assumption of independence, and there is no viable way to assess assumption of equal variance; however, we will now address the assumption of normality and assess the general fit of the model through consideration of the residuals. This was done for each of the top candidate models.

Functions in the packages sjPlot (Lüdecke 2017b) and sjmisc (Lüdecke 2017a) were used to construct the Normal QQ plots used to assess the assumption of the normally distributed random intercepts. In this plot the random intercepts are plotted against the quantiles of the normal distribution. If the assumption is reasonably satisfied, the random intercepts will appear to follow the line formed by the quantiles of the normal distribution.

Functions built-in to R (R Core Team 2017) were used to construct the boxplots of the residuals. If the model fits reasonably well, the residuals will be small. A common rule-of-thumb is that residuals over three units from zero are considered to be large. Just as with ordinary linear regression, large residuals can be indicative of missing important covariates, missing higher-order terms of the included covariates, and/or missing important interactions between the included covariates.

Diagnostics for Model 1

The assumption of normality for Models A and B was checked using Figures 11 and 13, respectively. We see that this assumption does appear to be reasonably satisfied for both models. The fit of these models was considered using Figures 12 and 14, respectively. For both candidate models, we see that the residuals are very, very large, with nearly identical distributions. Future analysis will involve further model building to address this issue.

Diagnostics for Model 2

The assumption of normality for Models A, B, and C was checked using Figures 15, 17, and 19, respectively. We see that this assumption does appear to be reasonably satisfied for all three models. The fit of these models was considered using Figures 16, 18, and 20, respectively. For all three candidate models, we see that the residuals are small, again, with nearly identical distributions. As a result, future analysis will not involve further model building.

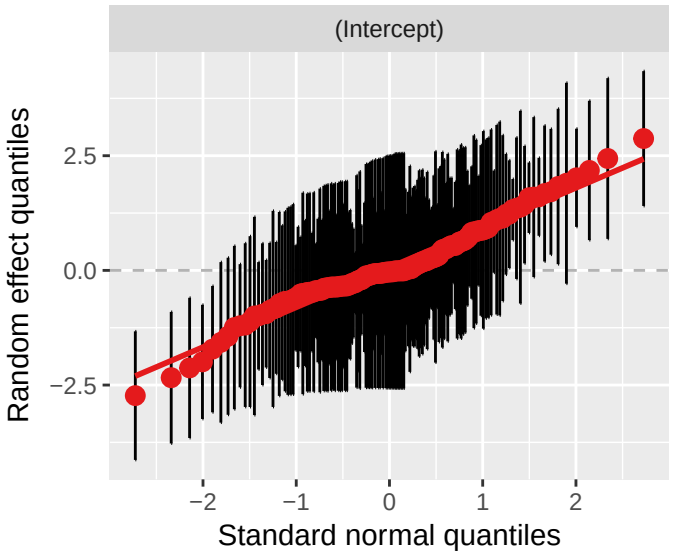

Figure 11: Normal QQ Plot used to assess the normality of the random intercepts in Model 1-A.

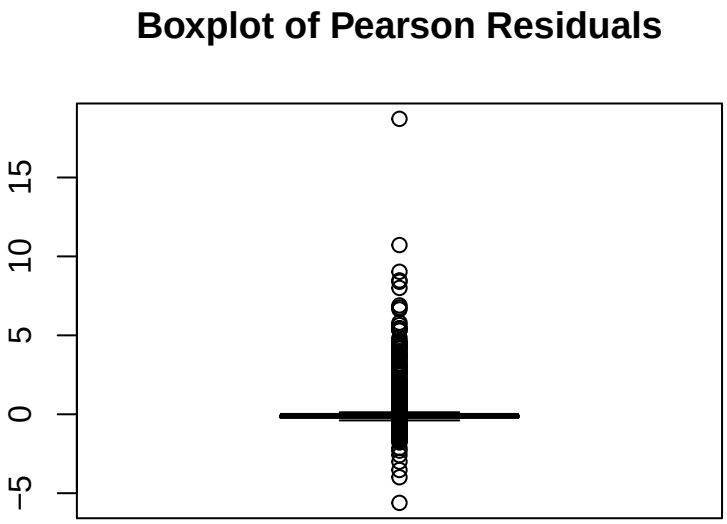

Figure 12: Boxplot of residuals for Model 1-A.

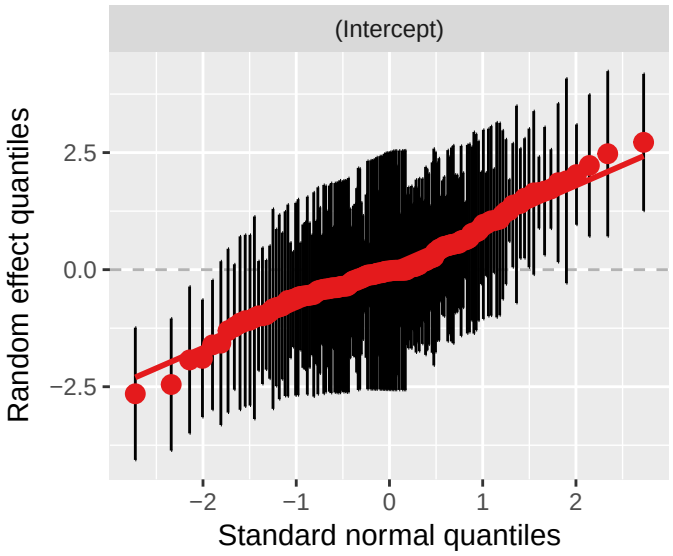

Figure 13: Normal QQ Plot used to assess the normality of the random intercepts in Model 1-B.

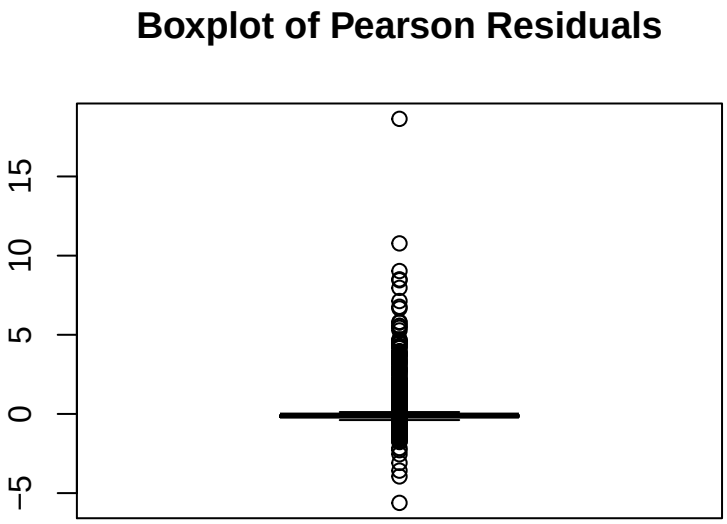

Figure 14: Boxplot of residuals for Model 1-B.

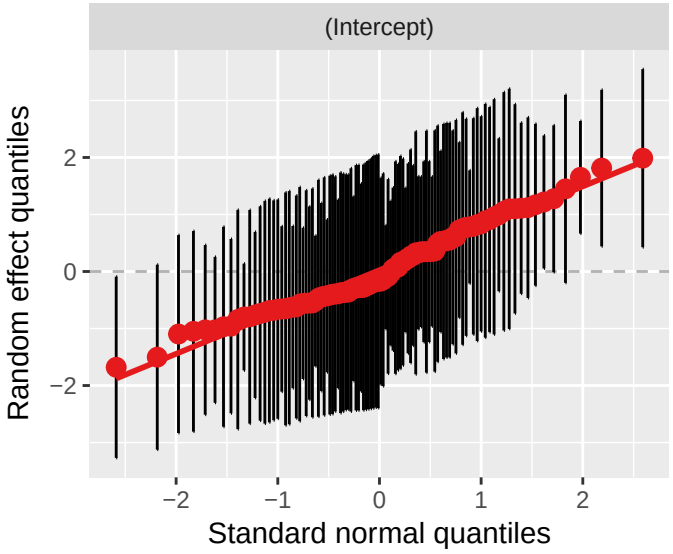

Figure 15: Normal QQ Plot used to assess the normality of the random intercepts in Model 2-A.

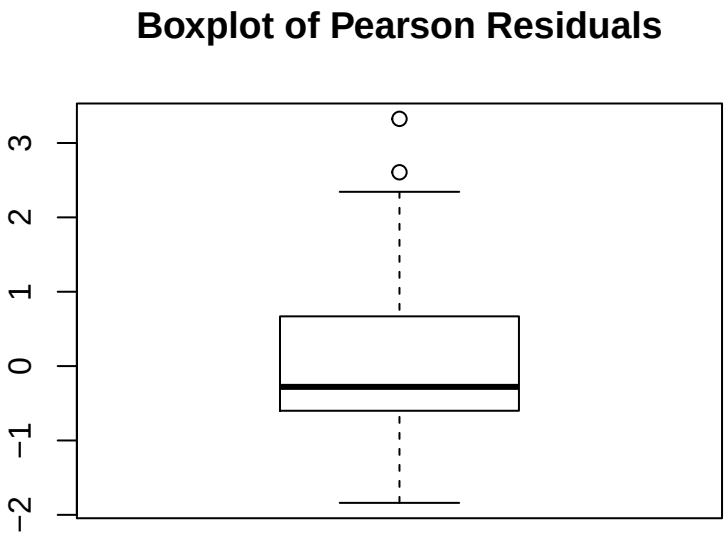

Figure 16: Boxplot of residuals for Model 2-A.

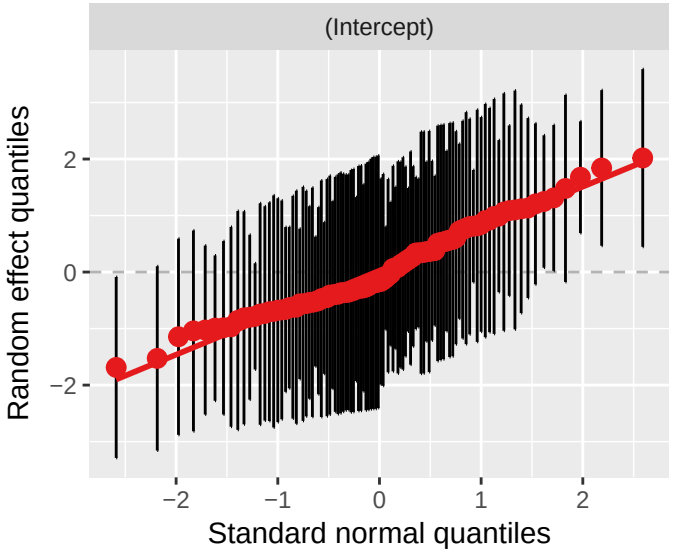

Figure 17: Normal QQ Plot used to assess the normality of the random intercepts in Model 2-B.

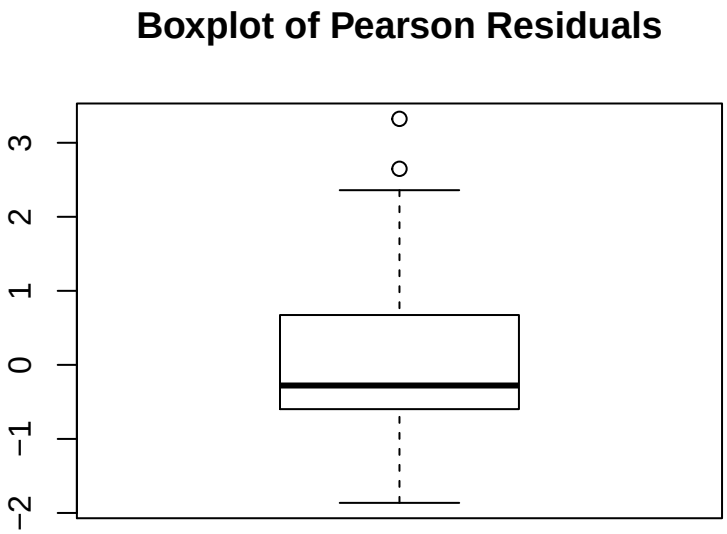

Figure 18: Boxplot of residuals for Model 2-B.

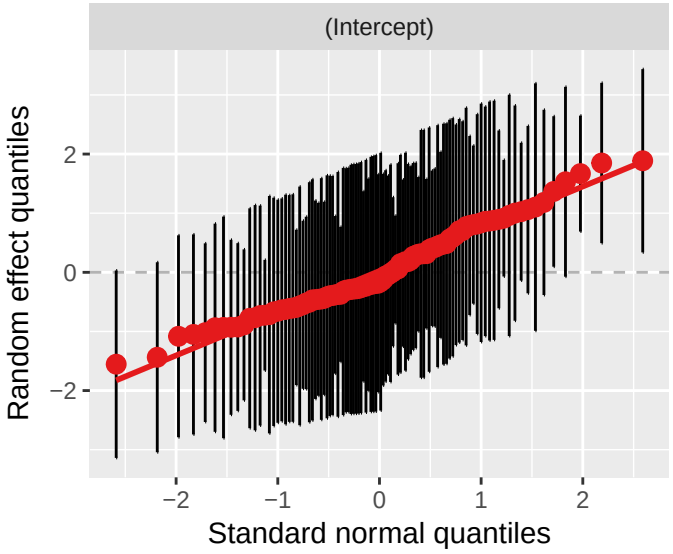

Figure 19: Normal QQ Plot used to assess the normality of the random intercepts in Model 2-C.

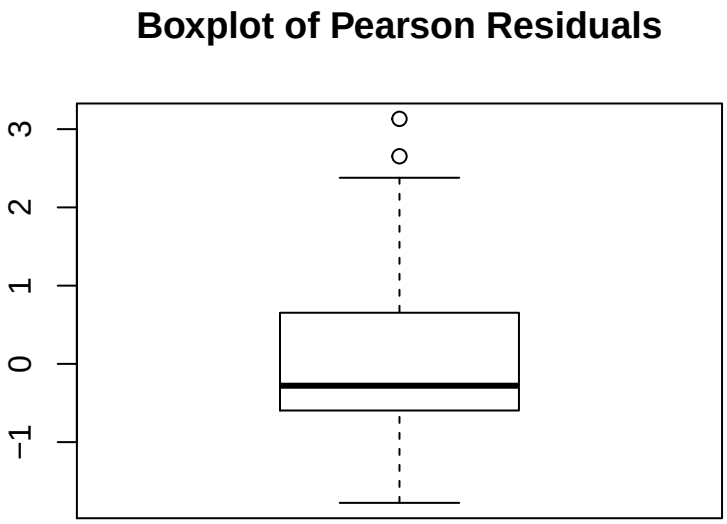

Figure 20: Boxplot of residuals for Model 2-C.

References

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*. John Wiley & Sons Inc.

———. 2015. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons Inc.

Al-Daffaie, Kadhem, and Shahjahan Khan. 2017. "Logistic Regression for Circular Data," May.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Fisher, N. I. 1993. *Statistical Analysis of Circular Data*. Cambridge University Press.

Graphics, Cotton Digital. n.d. "White Pine Blister Rust and Its Threat to to High Elevation White Pines." https://www.fs.fed.us/rm/highelevationwhitepines/Threats/blister-rust-threat.htm.

"Greater Yellowstone Ecosystem." n.d. *National Parks Service*. U.S. Department of the Interior. https://www.nps.gov/yell/learn/nature/ecosystem.htm.

Greater Yellowstone Whitebark Pine Monitoring Working Group. 2011. "Interagency Monitoring Protocol for the Greater Yellowstone Ecosystem, Version 1.1." Greater Yellowstone Coordinating Committee, Bozeman, MT.

———. 2016. "Monitoring Whitebark Pine in the Greater Yellowstone Ecosystem: 2015 Annual Report." *Natural Resource Report NPS/GRYN/NRR-2016/1146*. National Park Service, Fort Collins, Colorado.

Lüdecke, Daniel. 2017a. *Sjmisc: Miscellaneous Data Management Tools*. https://CRAN.R-project.org/package=sjmisc.

———. 2017b. *SjPlot: Data Visualization for Statistics in Social Science*. https://CRAN.R-project.org/package=sjPlot.

Maloy, O. C. 2001. "White Pine Blister Rust." *Plant Health Progress*, September. doi:10.1094/php-2001-0924-01-hm.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2015. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. http://www.rstudio.com/.