ANALYZING TREE DISTRIBUTION AND ABUNDANCE IN YUKON-CHARLEY RIVERS NATIONAL PRESERVE: DEVELOPING GEOSTATISTICAL BAYESIAN MODELS WITH COUNT DATA

By

Samantha Winder, B.A.

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Statistics

University of Alaska Fairbanks

May 2018

APPROVED:

Margaret Short, Committee Chair
Carl Roland, Committee Member
Scott Goddard, Committee Member
Julie McIntyre, Committee Member
Leah Berman, Chair
Department of Mathematics and Statistics

Abstract

Species distribution models (SDMs) describe the relationship between where a species occurs and underlying environmental conditions. For this project, I created SDMs for the five tree species that occur in Yukon-Charley Rivers National Preserve (YUCH) in order to gain insight into which environmental covariates are important for each species, and what effect each environmental condition has on that species' expected occurrence or abundance. I discuss some of the issues involved in creating SDMs, including whether or not to incorporate spatially explicit error terms, and if so, how to do so with generalized linear models (GLMs, which have discrete responses). I ran a total of 10 distinct geostatistical SDMs using Markov Chain Monte Carlo (Bayesian methods), and discuss the results here. I also compare these results from YUCH with results from a similar analysis conducted in Denali National Park and Preserve (DNPP).

Contents

List of Figures

List of Tables

1. Introduction

Species distribution models (SDMs) are an increasingly popular tool that ecologists use to answer questions about where species exist and thrive. These models typically attempt to relate species distribution or abundance to underlying environmental variables. Once created, SDMs can be useful for a number of applications, from predicting the impacts of future climate, land use, or other environmental changes, to helping inform conservation planning and selecting locations for reserves (Guisan and Thuiller 2005). While these models include spatial information by their very nature, it has recently become more common for modelers to incorporate spatially explicit error terms as a way to quantify potentially unmeasured covariates (Lichstein et al. 2002; Latimer et al. 2006; Ver Hoef et al. 2001).

In 2013, Roland *et al.* published a series of SDMs created from tree distribution data in Denali National Park and Preserve (DNPP; Roland, Schmidt, and Nicklen 2013). They examined the distribution and abundance patterns of the six tree species that occur in DNPP, utilizing an extensive dataset that included nearly 1000 plots covering a 12,800 km^2 study area. These data were collected as part of the Alaska Region Inventory and Monitoring Program by the Central Alaska Network (CAKN) vegetation team (as described in Roland et al. 2004). In 2016, the team completed baseline vegetation sampling in Yukon-Charley Rivers National Preserve (YUCH), thereby finished the second stage of the program.

For this project, I created similar distribution models to describe how the same tree species react to various environmental covariates in YUCH, as opposed to DNPP. This was of interest because YUCH is generally warmer and more forested than DNPP, and fires have historically been more common. Thus, the ability to compare and contrast models resulting from identical field inventory data from these two areas holds the potential to yield important ecological insights into the factors governing tree distribution in interior Alaska at an extensive scale. For example, the conditions in YUCH resemble those predicted for DNPP under a warming climate (Shulski and Wendler 2007). Therefore, understanding the environmental conditions that are most important in driving tree species distribution in YUCH may give us insight into future drivers of tree distribution in Denali.

In this paper, I will describe the steps required to build spatially explicit species distribution models for the YUCH trees, discussing broader implications along the way. I will begin by describing the environmental predictors and the response variables we used to create our models (§2.1), then discuss the form of the statistical models we chose (§2.2.1). I will describe the variable selection techniques we employed to select the most important environmental predictors for each species (§2.2.2, Appendix A), and discuss our process for adding spatially explicit error terms in a Bayesian framework, as well as some pitfalls we encountered along the way (§2.3, Appendix B). I will also discuss options and limitations of model fit metrics for these complicated models (§2.4). Having laid out these steps, I will present the results of our models (§3), and

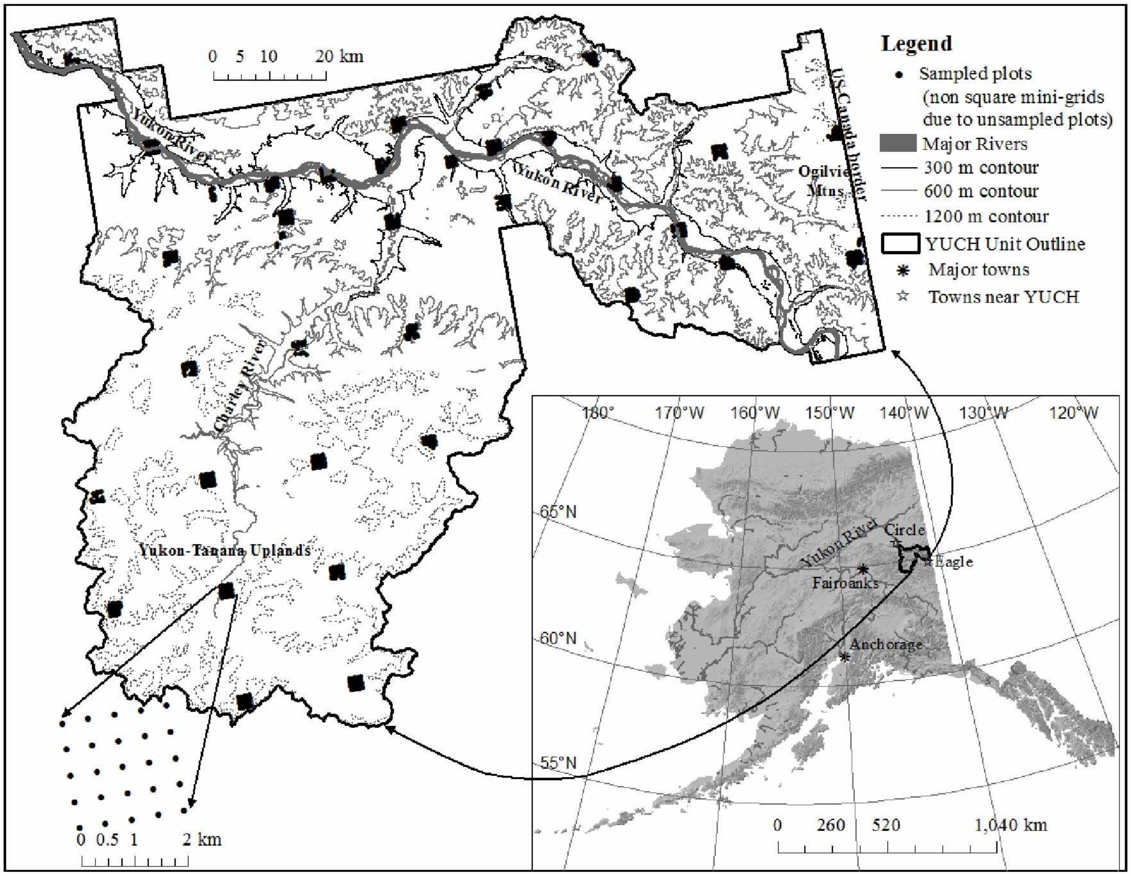

Figure 1. The location of Yukon-Charley Rivers National Preserve (YUCH) within Alaska, USA, including a map of YUCH with a schematic diagram of the sampling design showing the location of each sampled mini-grid in YUCH, and a diagram of the layout of a mini-grid (five rows of five plots spaced 500 m apart). Figure created by NPS.

briefly discuss a few ecological implications. Finally, I conclude with a discussion of the relative merits and difficulties of adding spatially correlated effects to this particular set of SDMs (§4.3).

2. Methods

2.1. **Data Description**

2.1.1. *Sampling design*

Data were collected by Carl Roland and the Central Alaska Network (CAKN) vegetation team from the National Park Service (NPS). Sampling was conducted following a two-stage systematic grid sampling design (shown in Fig. 1; Roland et al. 2004; MacCluskie et al. 2005). The team first generated a macro-grid at 10 km intervals that was laid over the entire park area using a random starting point. Anchored at each of these 10 km grid points was a secondary mini-grid consisting of 25 plots in a five by five grid, with each plot spaced 500 meters apart. Within a six-km buffer along the Yukon River, each of these mini-grids was sampled. Due to cost considerations, outside of the Yukon River corridor, the spacing between mini-grids was increased to 20 km (as in DNPP). At each of the 25 points within a mini-grid, a circular plot with a radius of 8 m was established (≈ 200 m^2). At each plot, the team collected a wide variety of vegetative and soil characteristics.

2.1.2. *Tree data*

Five of the six tree species that grow in interior Alaska can be found in YUCH, namely *Picea glauca* (white spruce; PICGLA), *Picea mariana* (black spruce; PICMAR), *Betula neoalaskana* (Alaska birch; BETNEO), *Populus tremuloides* (quaking aspen; POPTRE), and *Populus balsamifera* (balsam poplar; POPBAL). At each plot, both presence/absence (occurrence) and basal area (abundance) were measured for each of these species, according to the following protocol. If a live individual of any size was noted anywhere in the plot, that species was marked as present. For all individuals ≥ 1.37 m tall, the field team recorded diameter at breast height (dbh; 1.37 m), species, and condition class (live/dead). Collectively, these measurements allowed us to quantify the occurrence (presence/absence), density (stems per hectare), and abundance (basal area [BA]; m^2 of bole per hectare at breast height) of each tree species.

2.1.3. *Predictor variables*

We chose 15 environmental variables that we expected to explain patterns in distribution and abundance of the five species considered (Table 1; see Appendix A for a discussion about how we chose these particular 15 variables). Plot location and elevation were determined using a GPS in the field and corrected with Pathfinder Office software. Slope angle was measured in degrees using a clinometer. Estimates of annual solar radiation were made using the Solar Analyst tool in ArcGIS 10.0 (Dubayah and Rich 1995), which incorporates a number of distinct characteristics including slope angle, aspect, latitude, sun angle, and surrounding topography to determine insolation receipts of a topographic surface (Rich et al. 1994).

Site moisture (xeric, mesic, or subhygric) is an index based on plot slope angle, plot slope shape (convex, concave, or other), plot drainage class, soil texture, and whether or not frozen soil was detected within 1 m of the surface at the time of sampling (following Johnstone, Hollingsworth, and Chapin 2008, but condensing 6 categories into wet [subhygric], moist [mesic], or dry [xeric]). A site was classified as being a thawed alluvial terrace if the plot occurred on river-deposited alluvium and contained no evidence of near-surface permanent frozen ground.

Active layer depth was measured at each corner of four quadrats within the plot using a 1 m soil probe. This resulted in 16 measurements of soil depth per plot. Because rocks or other isolated restrictive features may be encountered within the soil column (resulting in artificially shallow depth measurements), we used the mean of the deepest probe depths recorded from each quadrat as the active layer depth for the plot.

Soil data were collected for each plot at four points 1 m beyond the plot perimeter in each cardinal direction. At each point, the vegetation team exposed a small soil pit (30-40 cm), where they measured depth of the

TABLE 1. Environmental factors for plots sampled in Yukon-Charley Rivers National Preserve (YUCH), Alaska, USA, that were used to develop spatial models of tree species occurrence and abundance.

Variable	Description	Units	Mean	Median	Minimum	Maximum
Topographic factors						
Elevation	Elevation at plot center	m	681	554	197	1735
Slope	Slope angle	degrees	14	11	0.5	45
Annual solar radiation	Solar radiation	WH/m^2	6.2×10^5	6.2×10^5	2.1×10^5	8.8×10^5
Site moisture	Site wetness	3 classes	NA	NA	NA	NA
Thawed alluvial terrace	Plot located on river terrace without permafrost	binary	0.07	0	0	1
Edaphic factors						
Active layer depth	Mean of 4 deepest probe depths	cm	54.8	50	0	130
Live mat depth	Depth of live mat (e.g. moss)	cm	3.6	3	0	12.5
SOL	Depth of soil organic layer	cm	12	10.25	0	30
pH	Reaction of the soil sample	pH	5.3	5.1	3.6	7.9
Gravel	Percent of soil sample > 2 mm	%	12.5	5.8	0	77.6
Total carbon	Percent carbon	%	15.1	8.2	0.2	53.1
Frozen soil	Presence of growing season shallow frozen soil	binary	0.42	0	0	1
Mineral cover	Percent of ground surface occupied by mineral material	%	10.35	0	0	100
Soil map factors						
Permafrost status	Soil unit permafrost status	3 classes	NA	NA	NA	NA
Fire factors						
Fire age	Age of fire affecting plot	3 classes	NA	NA	NA	NA
Biotic factors						
Dispersal deficit (PICGLA)	Estimated seed availability	binary	0.1	0	0	1
Broadleaf BA	Combined basal area of broadleaf trees	m^2/ha	1.8	0	0	28.5
PICGLA BA	Basal area of PICGLA	m^2/ha	2.4	0	0	58.9
PICMAR BA	Basal area of PICMAR	m^2/ha	1.8	0	0	29.2

litter, living, and soil organic (duff) layers, and averaged them for the plot. Soil samples were processed at the University of Alaska Fairbanks Agriculture and Forestry Experiment Station Soils Laboratory (Palmer, Alaska, USA), where the percent gravel (fragments > 2 mm), carbon content, and soil pH were measured. Percent mineral was measured along two transects that bisected the plots, and includes bare ground, rock, and gravel.

Permafrost status was quantified in two ways. First, each plot was categorized as continuous, sporadic, or discontinuous permafrost using the spatial data layer of the Soil Survey for Yukon-Charley Rivers National Preserve by Nathan Parry, 2013 USDA-NRCS. Secondly, a binary "growing season shallow frozen soil" (GsSFS) variable was created as described in Nicklen et al. 2016. A plot was classified as having growing season shallow frozen soil if the average of the 16 soil depth measurements within each plot was less than 50 cm and ice was encountered in one of the four soil pits. The plot was also classified as having GsSFS if the four soil temperatures taken at the plot were all below 1 °C. GsFSF is indicative of cooler sites, and encodes information about frozen ground at a much finer resolution than the permafrost status variable drawn from the soil survey maps.

The fire history of each plot was classified using two methods: (1) using the Alaska Fire Service spatial layer which contains all mapped fire perimeters from 1940 to 2010 (see Kasischke et al. 2010) and (2) through direct evidence of recent fires when sampling. From these sources, each site was classified into one of three categories: no evidence of fire in the past $\sim$ 100 years (unburned), plots affected by recent fires since 1982 (recent burn), and plots affected by fires before 1982, but after 1940 (old burn).

For *Picea glauca* occupancy, each of the 34 mini-grid study areas was evaluated to assess whether there was the potential for a lack of available seed sources to affect establishment dynamics. A mini-grid was classified as being located in an area subject to potential for dispersal deficit for *Picea glauca* (i.e. having a paucity of local seed sources) by a combination of examining high resolution aerial photography and direct experience hiking throughout mini-grid areas. Three out of 34 2×2 km square polygons did not have visible stands of coniferous trees and thus were designated as potential dispersal deficit areas for this species. Three additional lowland mini-grids that were most affected by recent and severe fire over a majority of their area were also designated as being in dispersal deficit zones due to the inferred paucity of live seed sources in these areas relative to other areas of the landscape.

Finally, we included *P. glauca* basal area, *P. mariana* basal area, and a combined broadleaf basal area as potential predictors, including only the two out of three that were not the model response (e.g. when modeling occurrence of *P. glauca*, we considered the effects of *P. mariana* and broadleaf basal area). This was done as a way of attempting to indirectly incorporate information about species interactions.

Before fitting the models, we standardized all the continuous variables by subtracting the sample mean and dividing by the sample standard deviation. This caused each variable to have a mean of 0 and a standard deviation of 1, thereby allowing us to draw comparisons between their coefficient estimates.

2.2. **Model Selection**

2.2.1. *Two stage process*

We conducted the analysis in two stages. We first created site occupancy models for each species at all 693 plots, using as a response presence or absence of the species. A species was counted as absent if no evidence of it was found in the plot. If a live tree, sapling, or seedling was present in the plot, then the species was counted as present.

After modeling species occurrence, we also modeled species abundance, where present, by looking only at the subset of plots for which the species was present. (We were able to create these abundance models for every species but *Populus balsamifera*, which was only present at 46 sites.) Abundance, which describes how common a species is, is often represented using individual counts. However, there are many other measures that describe different aspects of abundance, particularly for plants (Wilson 2011), and the choice of abundance measure can lead to differing conclusions about the data (Anderson, Chiarucci, and Williamson

2012). While we did count the number of stems present in our plots for each species, this is not a reliable count of individuals as several of the species in our study area reproduce through root sprouting. In addition, stem count does not always correspond well with biomass because a large number of stems may represent many small or many large individuals. Instead, plot basal area is a better proxy for tree biomass than stem count, as it directly incorporates information about the size of the trees. For these reasons, we used basal area as our response variable in our abundance models.

Basal area was calculated from living trees greater than 1.37 m high by summing the "area" of each individual tree at breast height, as determined by the diameter at breast height measurement. This measure can be thought of as the total area of the plot which would be covered by trees if looking at a 1.37 m high cross section. Because basal area is not measured at the ground, it misses seedlings and saplings. Therefore, it is possible (and, in fact, was common) for a species to be present in a plot but to have a basal area (and thus, abundance score) of zero. While basal area is a continuous variable (m^2/ha), the large number of zeros in this dataset proved a modeling difficulty. Rather than fitting a zero-inflated model to these data, we rounded the basal area measurements up to the nearest integer value (see Min and Agresti 2002 discussion about ordinal threshold models). Doing so allowed us to model abundance as a discrete variable, in fitting with the common ecological literature on abundance modeling (e.g. Vincent and Haworth 1983, White and Bennetts 1996, Potts and Elith 2006, Joseph et al. 2009, Roland, Schmidt, and Nicklen 2013).

Since presence/absence and basal area encode different information about the species (a tree may be able to exist, but not thrive, in certain conditions), we expected that a single species may respond to the same environmental covariates in different ways when looking at occurrence vs. abundance. Thus, we used the same set of covariates to predict both occurrence and abundance in the respective models (see Zuur et al. 2009). In total, we created ten models for this project. These were occupancy and abundance models for *Picea glauca*, *Picea mariana*, *Betula neoalaskana*, and *Populus tremuloides*, an occupancy model for *Populus balsamifera*, and a combined basal area abundance model. The response for this final model was total basal area in a plot (all species combined), using the 572 plots where at least one tree was present.

2.2.2. *Species-specific variable selection*

In addition to the main effects of 15 environmental variables that were considered for inclusion in all models (Table 1), we incorporated several secondary effects. First, we included quadratic terms for elevation and slope, as there is a strong theoretical basis for expecting that there could be "best" values of each of these variables (rather than a simple linear relationship between abundance and elevation, for example). Secondly, we included an interaction term between elevation and solar radiation. We included this term in order to test the hypothesis that high levels of solar radiation could allow certain species to grow at higher elevations than they would typically be able to.

Once we had compiled the list of 18 (including the two quadratic terms and the interaction term) potential explanatory variables, we used those and the other species' basal areas to fit a full (frequentist) generalized linear model for each species using `R` (R Core Team 2017). For occupancy models, these were binomial with a logit link (using the `glm` function in base `R`), while for the abundance models we fit negative binomial models with a log link using the `MASS` package (Venables and Ripley 2002). These models can be described as follows:

(1) Occupancy models

$$Y_i \sim \text{Bernoulli}(p_i), \qquad i = 1, ..., 693 \text{ sites}$$

$$\text{logit}(p_i) = \boldsymbol{x_i'\beta}$$

Where Y_i is a binary variable describing presence or absence of the species at site i, p_i is probability of occurrence at site i, $\boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, ..., x_{ip}), i = 1, ..., N$, which represents the values of the covariates at the i^{th} site and $\boldsymbol{\beta}' = (\beta_0, \beta_1, ..., \beta_p)$ is a vector of coefficients.

The logit link (also known as the log odds) is a function that converts the linear response into values ranging between 0 and 1. It can be solved as follows:

$$p_i = \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}}$$

(2) Abundance models

$$Y_i \sim \text{Negative Binomial}(\mu_i, \theta) \qquad i = 1, ..., N \text{ occupied sites}$$

$$log(\mu_i) = \boldsymbol{x_i'\beta}$$

Where Y_i is basal area (m^2/ha), μ_i is the mean response at site i, $\boldsymbol{x_i'}$ and $\boldsymbol{\beta}$ are as described above. Here we use a log link, which can be solved as follows:

$$\mu_i = e^{\boldsymbol{x_i'\beta}}$$

We first modeled abundance as a Poisson random variable. However, the Poisson distribution assumes that the mean and variance are equal, a condition that was not met in these datasets. Instead, the data appeared to be "overdispersed" (meaning that the variance is greater than the mean), likely due to the large number of zeros discussed in §2.2.1. This was determined by calculating the ratio of the residual deviance to the residual degrees of freedom. If the Poisson model is appropriate, the residual deviance has an approximate χ^2 distribution with $(n - p)$ degrees of freedom, where p is the number of unknown parameters in the fitted model (Agresti 2013). Since the expected value of a χ^2 distribution is its degrees of freedom, it follows that the residual deviance should be approximately equal to its degrees of freedom if the model is appropriate, leading to a ratio of approximately 1, with larger values indicating overdispersion. The ratios calculated for the models were much greater than 1, so we chose to use the negative binomial distribution, which includes a dispersion term. In the parameterization used here, $E(Y) = \mu$ and $Var(Y) = \mu + \mu^2/\theta$.

Initially, we hoped to use Bayesian variable selection methods to choose the best approximating models for each species. However, due to the complexity of our models, this was determined to be too computationally difficult and time intensive. Instead, we used frequentist model selection techniques as follows. We first fit full models for each species, which included all possible predictors from Table 1. From the full models, we used backwards and forwards stepwise selection to find the best combination of explanatory variables, using AIC as our measure of model fit. Stepwise selection is a process in which each variable is added or dropped in turn, AIC is calculated for each resulting model, and the new model with the lowest AIC is selected. This process is repeated until a combination of variables is found which results in the lowest AIC. Our best approximating models are shown in Table 4 and Table 5.

2.3. **Adding Spatial Effects**

While our best approximating models appeared to fit relatively well, we were concerned that we could still be missing an important predictor. Specifically, as there was evidence of spatially correlated residuals in several of the models, we chose to incorporate spatially explicit error terms into the best approximating models chosen above, as described below.

Since our plots are effectively points (on the scale of the entire park) that are georeferenced using latitude and longitude, our data are considered to be point-referenced, or geostatistical. A common way to model geostatistical data is by assuming a correlation structure between points that decays continuously as a function of distance (Finley, Banerjee, and Carlin 2007). We assumed that these spatially correlated errors followed a multivariate normal distribution (thus creating Gaussian spatial process models), with a mean of 0 and an exponential covariance function. This structure includes two spatial terms: σ^2, which represents site specific variance, and ϕ, which describes the distance over which the spatial correlation decays as follows: $\phi \approx 3/d_0$, where d_0 is effective distance. Once two sites are separated by d_0, they are expected not to exhibit any residual spatial correlation. Here we are assuming isotropy, which means that spatial correlation is only dependent on the distance between two sites, without a directional component. For geostatistical GLMs, the spatially correlated errors are added to the "mean structure", which is the linear part of the model, before the link function is applied.

Now our models were:

(1) Occupancy models with spatial effects

$$\begin{aligned}
Y(\boldsymbol{s}_i) &\sim \text{Bernoulli}(p(\boldsymbol{s}_i)) \\
logit(p(\boldsymbol{s}_i)) &= \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + w(\boldsymbol{s}_i) \\
w(\boldsymbol{s}_i) &\sim MVN(\boldsymbol{0}, k(\phi)) \\
k(\phi)_{ij} &= \sigma^2(e^{-\phi\|\boldsymbol{s}_i - \boldsymbol{s}_j\|})
\end{aligned}$$

(2) Abundance models with spatial effects

$$\begin{aligned}
Y(\boldsymbol{s}_i) &\sim \text{Poisson}(\lambda(\boldsymbol{s}_i)) \\
log(\lambda(\boldsymbol{s}_i)) &= \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + w(\boldsymbol{s}_i) \\
w(\boldsymbol{s}_i) &\sim MVN(\boldsymbol{0}, k(\phi)) \\
k(\phi)_{ij} &= \sigma^2(e^{-\phi\|\boldsymbol{s}_i - \boldsymbol{s}_j\|})
\end{aligned}$$

2.3.1. *Model fitting*

These spatial glms are fairly complex, and tend to be quite difficult to fit using maximum likelihood estimation (frequentist statistics). So we created Bayesian models using vague priors which we expected to be essentially equivalent to the frequentist models described above. This allowed us to utilize Bayesian methods, which are better suited to fit such complex models. Specifically, we used the `spBayes` function in `R` (Finley, Banerjee, and Carlin 2007), which fits spatial glms using the Metropolis-Hastings algorithm and includes an adaptive MCMC (AMCMC) algorithm that updates tuning parameters as it runs. The AMCMC algorithm used in `spBayes` preserves ergodicity, and returns valid and effective results (Rosenthal 2007).

All models were run for 500,000 iterations. Convergence was evaluated visually from the trace plots, as well as through the time-series standard error statistic (an estimate of Monte Carlo error; values lower than approximately 1% of the posterior mean suggest convergence). The first 150,000 samples were thrown out as burn-in for the abundance models, and 40,000 samples were used as burn-in for the occupancy models. We then used a thinning factor of 100 in order to reduce autocorrelation between iterations and to reduce the computational memory required to store and work with the MCMC samples. We calculated 95% credible intervals for each parameter estimate by dropping the bottom 2.5% and top 2.5% of the values (equal-tailed intervals).

2.3.2. *Prior selection*

Initially, we selected vague priors for all of our parameters, including the spatial terms. For the fixed effect coefficients (the β's), we assumed Normal distributions centered at 0 with standard deviations of 5. For ϕ, we assumed a uniform distribution bounded by the upper and lower expected effective distance (d_0). The lower

d_0 was the minimum distance between two plots in our dataset, and the upper d_0 was 3/4 of the maximum distance between two plots. While these values varied for the abundance models (since these models were built using a subset of the plots), for the occupancy models which included all 693 plots the lower d_0 was 85 m and the upper d_0 was 116,968 m. Specifically, as $\phi \approx 3/d_0$, we assumed that ϕ followed a uniform distribution between (3/(upper d_0), 3/(lower d_0)). For the abundance models, we assumed a vague Inverse Gamma distribution for σ^2. Initially, we used this same prior for the occupancy models, but due to issues with confounding between the σ^2 term and the β's in the occupancy models (see Appendix B for a discussion about this), we ultimately fixed $\sigma^2 = 1$ for these five models.

The priors we used in the final models are as follows:

(1) Occupancy models

$$\begin{aligned} \beta_k &\overset{\text{iid}}{\sim} N(0,25) \qquad\qquad k = 1, ..., P \text{ parameters} \\ \phi &\sim \text{Unif}\Big(\frac{3}{116968}, \frac{3}{85}\Big) \\ \sigma^2 &= 1 \end{aligned}$$

(2) Abundance models

$$\begin{aligned} \beta_k &\overset{\text{iid}}{\sim} N(0,25) \qquad\qquad k = 1, ..., P \text{ parameters} \\ \phi &\sim \text{Unif}\Big(\frac{3}{\text{upper } d_0}, \frac{3}{\text{lower } d_0}\Big) \\ \sigma^2 &\sim \text{IG}(2.1, 0.5) \end{aligned}$$

2.3.3. *Final models*

Thus, our full posterior distributions were as follows:

(1) Occupancy models

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

$$\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2, \phi\}$$

$$\text{Where} \qquad L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{s}_i)^{Y_i}(1 - p(\boldsymbol{s}_i))^{1-Y_i}$$

$$logit(p(\boldsymbol{s}_i)) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + w(\boldsymbol{s}_i)$$

$$\text{and} \qquad \pi(\boldsymbol{\theta}) = \prod_{k=1}^{P} \phi(\beta_k; 0, 25) \times I(\phi \in \frac{3}{116968}, \frac{3}{85})$$

(2) Abundance models

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

$$\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2, \phi\}$$

Where $$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{e^{-\lambda(\boldsymbol{s}_i)}\lambda(\boldsymbol{s}_i)^{Y_i}}{Y_i!}$$

$$log(\lambda(\boldsymbol{s}_i)) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + w(\boldsymbol{s}_i)$$

and $$\pi(\boldsymbol{\theta}) = \prod_{k=1}^{P} \phi(\beta_k; 0, 25) \times I(\phi \in \frac{3}{\text{upper } d_0}, \frac{3}{\text{lower } d_0}) \times IG(\sigma^2; 2.1, 0.5)$$

Careful readers may notice that in the variable selection phase we modeled abundance assuming that it followed a negative binomial distribution, but switched to a Poisson distribution when adding the spatial component. This is because the geostatistical Poisson model is more straightforward, and ends up being very similar to a negative binomial model with a spatial term (Jay M. Ver Hoef, personal communication 7/18/17). See Appendix C for a more in depth comparison of the two models, as well as a discussion and code for building a geostatistical negative binomial model in `STAN` (Stan Development Team 2016).

2.4. Assessing model fit

We calculated the deviance information criterion (DIC) for each model in order to assess whether the addition of spatial effects improved the models enough to justify the added complexity. DIC is a "somewhat Bayesian version of AIC" (Gelman et al. 2014), which is calculated as

$$DIC = \bar{D} + p_{DIC}$$

where $\bar{D}$ is the posterior mean of the deviance, $D(\boldsymbol{\theta})$, and p_{DIC} is the effective number of parameters. Here, $D(\boldsymbol{\theta})$ is defined as $-2\log L(\boldsymbol{\theta})$. $\bar{D}$ will be small if the model fits well, while p_{DIC} penalizes complex models. Thus, as with AIC, small values of DIC are preferred. It is possible for DIC values to be negative (as was the case for our abundance models). In this case, the more negative values indicate better fit. We ran a Bayesian version of every model with and without spatial effects for each species, and calculated DIC for each. In every case, the models including spatial effects had lower DIC values and were judged to be better.

Generalized linear models don't have a convenient measure of model fit such as R^2. Instead, there have been a number of pseudo R^2 statistics proposed, but none of them have all of the desirable qualities of the linear R^2 (Cameron and Windmeijer 1996). Recently, Tjur (2009) proposed a new statistic to measure how well logistic models fit, which he calls D, the coefficient of discrimination (to avoid future confusion, I will denote it D_{COD} here). D_{COD} is pleasantly intuitive, in that it provides a measure of how well the model's fitted

TABLE 2. Tjur's D_{COD} values for occupancy models before and after the addition of spatial effects. Values closer to 1 indicate a better fit, and can be interpreted somewhat similarly to R^2 values. Species codes are as follows: PICGLA, *Picea glauca*; PICMAR, *Picea mariana*; BETNEO, *Betula neoalaskana*; POPBAL, *Populus balsamifera*; POPTRE, *Populus tremuloides*.

	Conifer		Broadleaf		
Occupancy Model	PICGLA	PICMAR	BETNEO	POPBAL	POPTRE
Nonspatial	0.42	0.69	0.63	0.33	0.51
Spatial	0.54	0.75	0.71	0.42	0.59

values match the observed values. It is calculated as

$$D_{\text{COD}} = \widehat{\pi_1} - \widehat{\pi_2}$$

where $\widehat{\pi_1}$ and $\widehat{\pi_2}$ denote the average fitted values for successes and failures, respectively. Therefore, if the model perfectly predicts successes, the first term will be 1, and if it perfectly predicts failures, the second term will be 0, leading to $D_{\text{COD}} = 1$. If the model has no predictive power (the results could be generated by chance), then $D_{\text{COD}} = 0$. We used this statistic to assess model fit for the occupancy models, and found that all five models fit the data better after the inclusion of spatial effects (Table 2).

Unfortunately, D_{COD} is only defined for logistic models. While some pseudo R^2 terms have been defined for Poisson regression, they are difficult to extend to spatially explicit Poisson GLMs such as the abundance models presented in this paper. Instead, we present the proportional reduction in deviance (confusingly also called D, here referred to as D_{PRD}) of the abundance models *before* adding spatial effects. This statistic, discussed in Zheng and Agresti 2000, can be calculated as

$$D_{\text{PRD}} = \frac{\text{Null Deviance} - \text{Residual Deviance}}{\text{Null Deviance}}$$

where

$$\text{Null Deviance} = -2(\log L(M_{\text{null}}) - \log L(M_{\text{saturated}}))$$

$$\text{Residual Deviance} = -2(\log L(M_{\text{fitted}}) - \log L(M_{\text{saturated}}))$$

These are the deviance terms that `R` provides in the summary of models fitted using the `glm` function.

3. Results

3.1. General forest characteristics

Occupancy and abundance patterns varied substantially among the five species (Table 3). Overall, 82.5% of the 693 plots included in the analysis contained at least one individual of the five tree species that occur in the area, and the mean combined basal area in all the plots was 5.97 m^2/ha. *Picea glauca, P. mariana,* and *Betula neoalaskana* were the most common species, occurring in 48%, 52%, and 46% of the plots respectively.

Table 3. Tree data summaries, explaining how frequent each species was in our 693 plots (frequency describes percent of plots that were occupied), and describing the basal area data for each species.

		Basal area, BA (m^2/ha)			
Species	Frequency (%)	Mean	SD	CV	Maximum
Conifer					
PICGLA	48.2	2.4	6.24	2.6	58.9
PICMAR	52.4	1.77	4.14	2.33	29.2
Broadleaf					
BETNEO	45.6	1.42	3.9	2.75	28.2
POPBAL	6.6	0.08	0.79	9.8	14.3
POPTRE	13.9	0.3	1.79	6.06	26.7
All species	82.5	5.97	8.91	1.49	59.2

P. glauca was the most abundant species in our plots, with a mean overall basal area of 3.4 m^2/ha, and a maximum basal area of 58.9 m^2/ha in one plot. *P. mariana* and *B. neoalaskana* had fairly similar abundance measures (mean basal area 1.77 and 1.42, respectively, with maximum values of approximately 29 m^2/ha). *Populus balsamifera* was both the least common (6.6% occupancy) and the least abundant (mean basal area 0.08, maximum 14.3) species in our study.

3.2. **Occupancy models**

3.2.1. *Coefficient estimates*

Our final occupancy models are shown in Table 4. Slope^2, site moisture - subhygric, live mat depth, and gravel were not significant ($\alpha = .05$) for any of the occupancy models, so are not included.

All tree species preferred lower elevations (generally less than 800 m), though *Picea glauca* predicted occupancy was relatively high over a greater range of elevation values (up to 1200 m, Fig. 2). We found that the interaction between elevation and solar radiation was significant in the occupancy models for both *P. glauca* and *B. neoalaskana.* For both species, higher values of solar radiation resulted in a high probability of occurrence over a larger range of elevations than was predicted in sites with lower values of solar radiation. *P. mariana* and *Populus tremuloides* had opposite responses to solar radiation, in that *Picea mariana* was more likely to occur in sites with low solar radiation, while *Populus tremuloides* showed a strong preference for sites with high solar radiation. Solar radiation was not included in our final occupancy model for *Populus balsamifera.*

The two coniferous species, *P. glauca* and *P. mariana*, differed substantially in their habitat preferences. *P. mariana* preferred lower, cooler sites with lower solar radiation, while *P. glauca* occurred more frequently in sunny, warm sites over a greater range of elevations (Fig. 3).

Table 4. Occupancy models. Direction of significant relationships between environmental covariates and occupancy probability in the best approximating models, including adjustments made after adding spatial random effects, for five tree species in YUCH. Square brackets indicate that the variable was significant before the addition of spatial random effects, but that the 95% credible interval included 0 after incorporating spatial effects. Effective distance is the median of the posterior distribution of d_0, and is in square brackets if the 95% credible interval included distances ≤ 500, the distance between regularly spaced plots.

	Conifer		Broadleaf		
Covariate	PICGLA	PICMAR	BETNEO	POPBAL	POPTRE
Tjur's D_{COD} (spatial models)	0.54	0.75	0.71	0.42	0.59
Topographic factors					
Elevation		-	-	-	-
Elevation^2	-	-	-		-
Slope	-				+
Annual solar radiation		-			+
Elevation × solar radiation	+		+	NA	NA
Site moisture - xeric	+				+
Thawed alluvial terrace		[-]	[-]		
Edaphic factors					
Active layer depth		-			
SOL	-	+		-	
pH	+	-		+	
Total carbon			-		
Frozen soil	-	+	-		
Mineral cover	-	-	-		
Soil map factors					
Permafrost - Discontinuous	[+]	-			
Permafrost - Sporadic	+	-			
Fire factors					
Recent Burn	-		+		+
Old Burn	-	+	+	+	+
Biotic factors					
Dispersal deficit (PICGLA)	-	NA	NA	NA	NA
Broadleaf BA	+	-	NA	NA	NA
PICGLA BA	NA	-			
PICMAR BA		NA	+	-	
Effective distance (m)	5340	6263	[4926]	[236]	[174]

Occurrence of *P. glauca* was negatively correlated with both recent and old burns, however every other species was found more often in plots that had burned (effects were significant on plots with old burns, but not always on the plots that had burned more recently, see Table 4 and Fig. 4). *P. mariana*, in particular, was far more likely to occur in sites which had experienced a burn before 1982. *Populus tremuloides* and *Betula neoalaskana* were both most likely to occur in sites which had burned recently.

Permafrost (as determined from the soil survey map) was only a significant predictor for the coniferous species, though whether or not the plot contained growing season shallow frozen soil was significant for

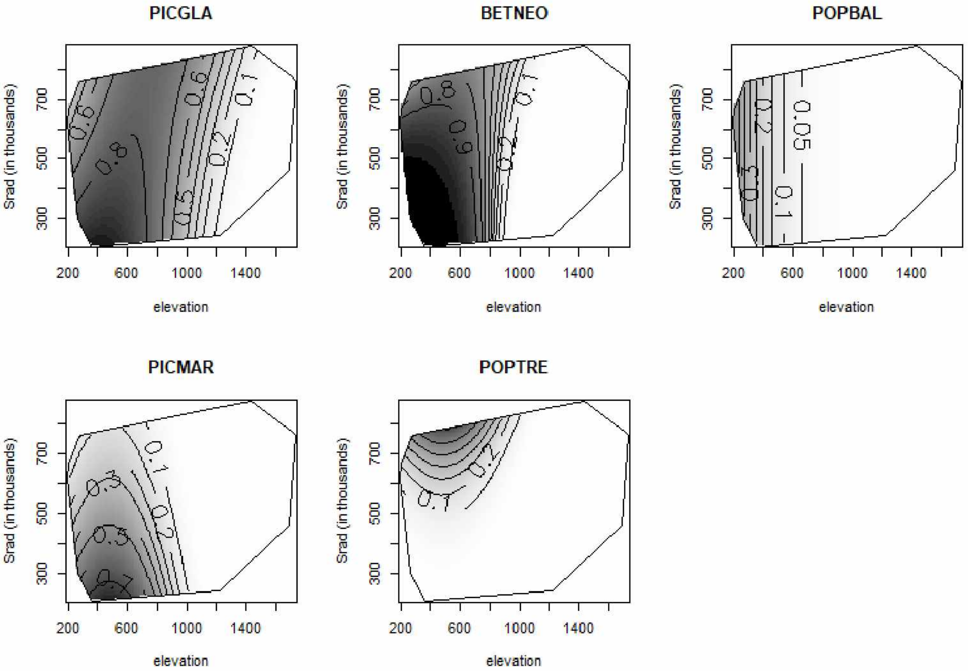

Figure 2. Contours showing occupancy probability across observed values of elevation and solar radiation for each species. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species, except as follows: Broadleaf species are modeled assuming an old burn, and *P. balsamifera* (POPBAL) is modeled at high pH (mean + 2 SD).

both the conifers and *Betula neoalaskana* (Table 4). Presence of frozen soil or permafrost had a negative effect on *P. glauca* and *B. neoalaskana* occurrence, but a positive effect on *P. mariana*. The models did not identify either permafrost or growing season frozen soil to be significant predictors for the other broadleaf species.

We found that increased broadleaf basal area was associated with an increased probability of *Picea glauca* occurrence, and a negative probability of *P. mariana* occurrence. However, *Betula neoalaskana* was more likely in plots with high *Picea mariana* basal area. This is likely due to shared responses to the covariates and successional dynamics rather than being a strictly causal result.

Soil pH was found to be a significant predictor of occupancy for *Picea glauca*, *P. mariana*, and *Populus balsamifera* (Fig. 5). The effect on *P. balsamifera* was positive, but quite small, while *Picea glauca* preferred sites with higher soil pH, and *P. mariana* preferred sites with lower soil pH.

Soil organic layer (SOL) depth was found to be significant for the same species that responded to soil pH, and the pattern appears to be the opposite (Fig. 6). Specifically, *P. glauca* preferred sites with shallow SOL, while *P. mariana* was found more often in sites with deep SOL.

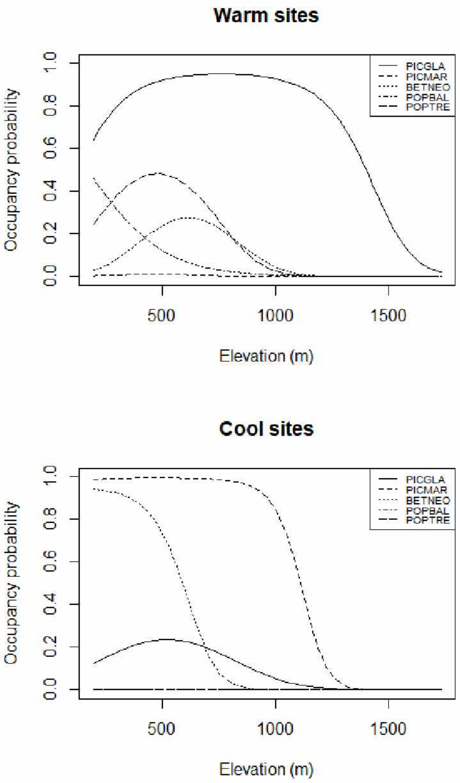

FIGURE 3. The predicted occupancy probability of each species across the range of observed elevations at warm sites (high annual radiation [mean + 2 SD], deep soil depth [mean + 2 SD], shallow soil organic layer [2 cm], high soil pH [mean + 2 SD], and no frozen soil) and at cool sites (low annual radiation [mean - 2 SD], shallow soil depth [mean - 2 SD], deep soil organic layer [30 cm], low soil pH [mean - 2 SD], and frozen soil). All other variables were held to mean values.

3.2.2. *Spatial Effects*

Both the calculated DIC values and Tjur's D_{COD} values supported the inclusion of spatial effects in every occupancy model, suggesting that including a spatially correlated error term improved the models. Including this error term did not have a large effect on our parameter estimates, with only a few becoming insignificant after inclusion of the spatial effects (these parameters are indicated by the square brackets in Table 4). Having fixed $\sigma^2 = 1$, ϕ and its counterpart d_0 are the only interpretable spatial output.

The range parameter, ϕ, is meaningful primarily for its relationship to effective distance ($d_0 \approx 3/\phi$), the distance after which locations no longer exhibit spatial correlation. Despite spatial models being preferred in every case, d_0 was only found to be significant (95% credible interval does not include 500 m, the minimum distance between regularly spaced plots) for the conifers (Table 4). The estimated effective distance for both *Populus* species was less than 500 m, indicating that there was no spatial autocorrelation present in these two models (Fig. 7). For the coniferous species, the estimated effective distance was just over 5 km (95% credible interval [2.4 km, 15.7 km]) for *Picea glauca*, and just over 6 km (1.5 km, 23.2 km) for *Picea mariana*.

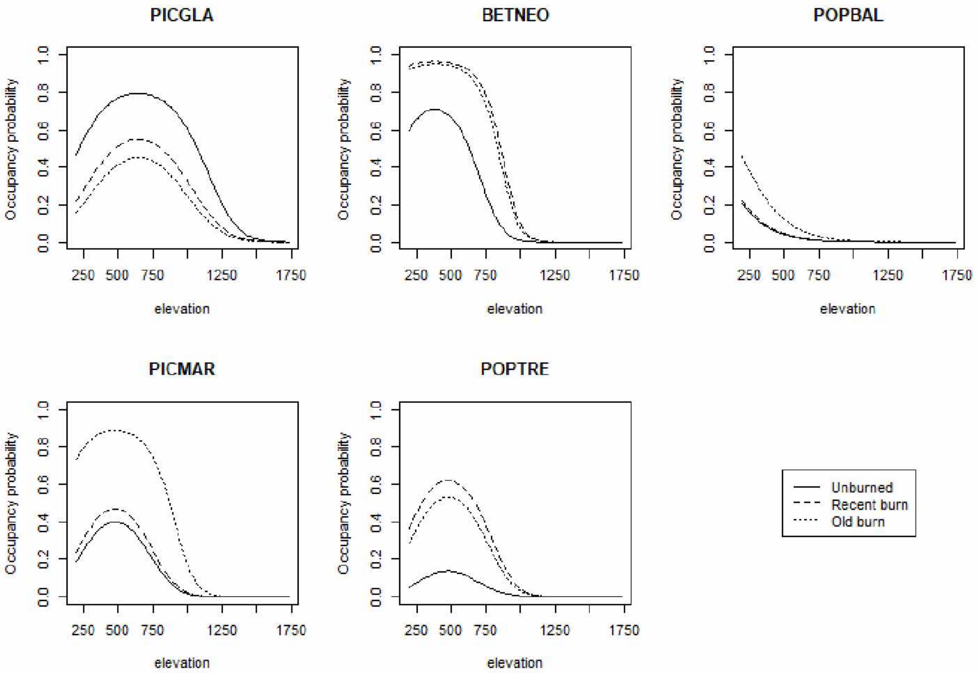

FIGURE 4. Occupancy probability of each species across the range of observed elevations at sites that were unburned, recently burned, or burned prior to 1982. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species, except as follows: *P. tremuloides* (POPTRE) is modeled at xeric sites and *P. balsamifera* (POPBAL) is modeled at high pH (mean + 2 SD).

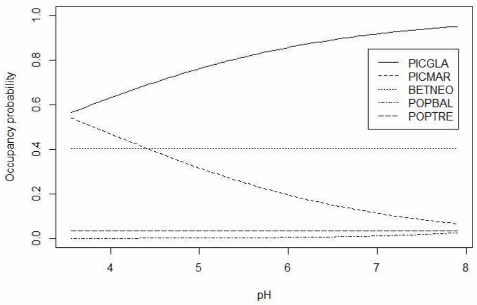

FIGURE 5. Occupancy probability of each species across the range of observed soil pH. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species.

Betula neoalaskana effective distance appears to be quite similar to the conifers, despite that fact that its wider credible interval did include 500 m (0.2 km, 24.9 km). The distance between mini-grids in the Yukon River corridor is 8 km (10 km between the centers of each mini-grid, so 8 km from the easternmost plot in one grid to the westernmost in a neighboring grid), while outside this corridor the distance between mini-grids is 18 km (20 km between the centers of the mini-grids). So for each of these species, it appears that any

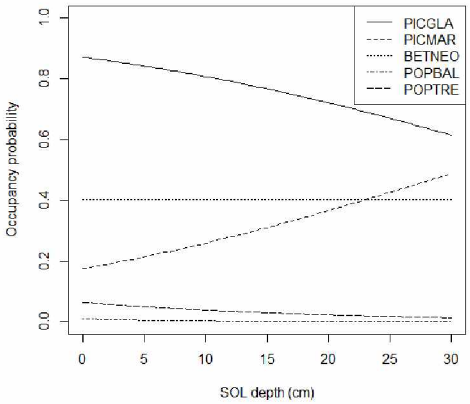

Figure 6. Occupancy probability of each species across the range of soil organic layer (SOL) depths. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species.

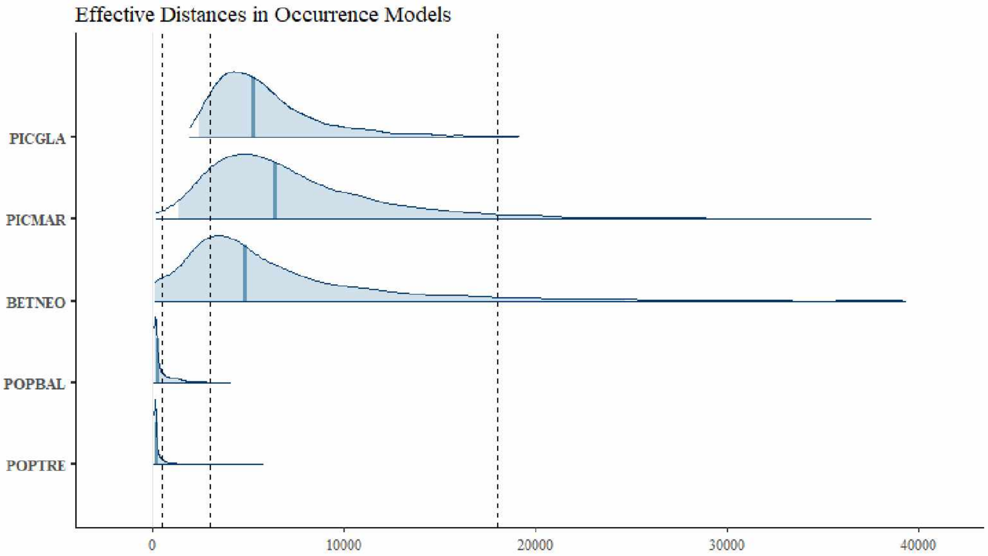

Figure 7. Posterior distributions of d_0, the effective distance parameter in the spatial occupancy models. Vertical lines are at 500 m, 3000 m, and 18000 m, which represent the minimum distance between regularly spaced plots, the maximum distance across a minigrid, and the typical distance between two minigrids, respectively.

residual spatial effects are limited to the scale of the mini-grid, and could potentially be modeled more easily by incorporating a mini-grid level random effect in a generalized linear mixed model (GLMM).

3.3. **Abundance models**

3.3.1. *Coefficient estimates*

Our final abundance models are shown in Table 5. Sporadic permafrost, gravel, and *Picea glauca* dispersal deficit were not significant ($\alpha = .05$) for any of the abundance models, so are not included in this table.

Table 5. Abundance models. Direction of significant relationships between environmental covariates and abundance (BA) in the best approximating models including adjustments made after fitting spatial random effects models for four tree species, as well as all species combined, in YUCH. Square brackets indicate that the variable was significant before the addition of spatial random effects, but that the 95% credible interval included 0 after incorporating spatial effects. Effective distance is the median of the posterior distribution of d_0, and is in square brackets if the 95% credible interval included distances $\leq$ 500, the distance between regularly spaced plots.

	Conifer		Broadleaf		
Covariate	PICGLA (n = 334)	PICMAR (363)	BETNEO (316)	POPTRE (96)	All species (572)
Proportional reduction in deviance (D_{PRD}; nonspatial models)	0.52	0.44	0.52	0.59	0.54
Topographic factors					
Elevation	-	[-]		[+]	-
Elevation^2	-	-			-
Slope	+	+			+
Slope^2	-	-	-		-
Annual solar radiation	+	+	-		
Elevation × solar radiation	+				+
Site moisture - subhygric	[-]	[-]	-		-
Site moisture - xeric	[-]			+	
Thawed alluvial terrace	+				+
Edaphic factors					
Active layer depth	+				+
Live mat depth	+	[+]	-	-	
SOL		+	[-]		
pH	+		-	+	[+]
Total carbon		-			-
Frozen soil	-				[-]
Mineral cover	-		-		-
Soil map factors					
Permafrost - Discontinuous		[-]			[-]
Fire factors					
Recent Burn	-	-	-		-
Old Burn	-	[-]	+	+	[-]
Biotic factors					
Broadleaf BA		-	NA	NA	NA
PICGLA BA	NA	-			NA
PICMAR BA	-	NA			NA
Effective distance (m)	[1092]	2510	[1246]	[320]	1656

The interaction term between elevation and solar radiation was significant (95% credible interval did not include zero) for *Picea glauca* abundance and overall tree abundance. In high elevation plots, *P. glauca* and overall tree abundance were higher, on average, in areas with increasing solar radiation receipts (Fig. 8). *Picea mariana* abundance was positively associated with solar radiation, while *Betula neoalaskana* abundance was negatively associated with solar radiation. While *Populus tremuloides* occupancy was positively influenced

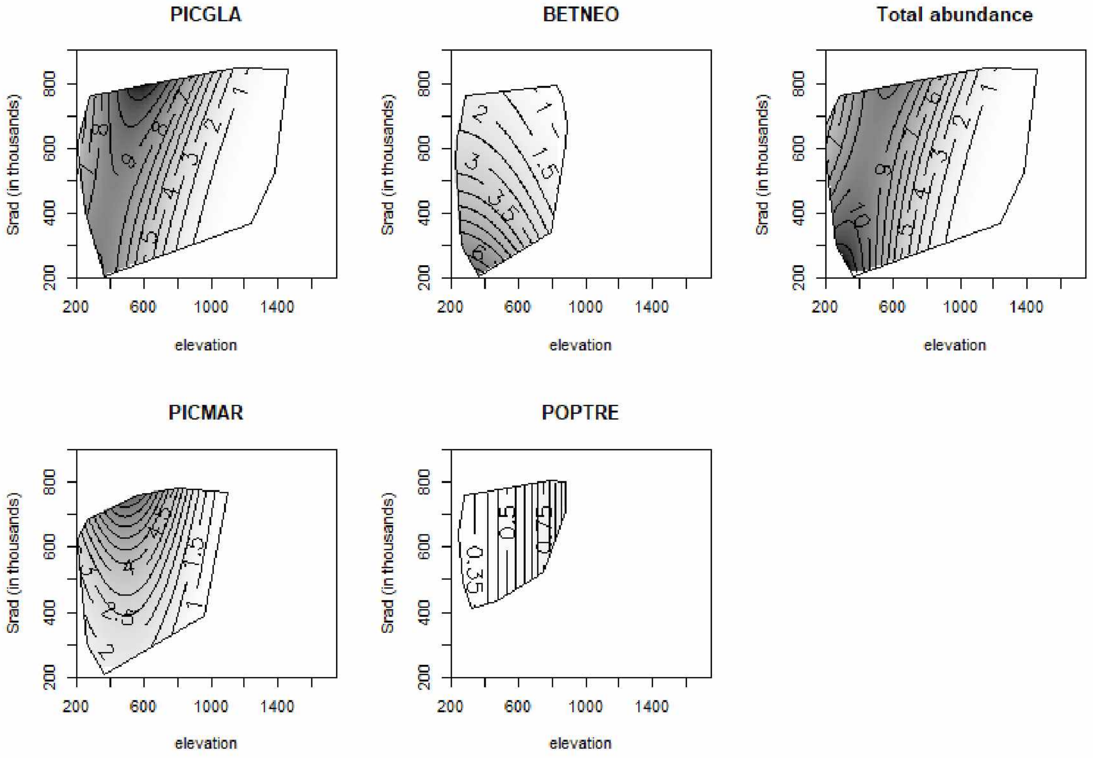

Figure 8. Contours showing predicted abundance (BA) across observed values of elevation and solar radiation for each species. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species, except as follows: *P. tremuloides* (POPTRE) is modeled assuming an old burn. Contours only show observed combinations of elevation and solar radiation for each species.

by increasing solar radiation, abundance of this species did not show a significant response to solar radiation within the sites where it occurred.

Peak overall tree abundance in YUCH occurs at approximately 500 m with slopes of just over 20 degrees, a trend that is shared by both conifer species (Fig. 9). *Betula neoalaskana* prefers lower elevations and slightly shallower slopes, with peak abundance occurring below 400 m and between 10 and 20 degree slopes. Slope was not included in our final model for *Populus tremuloides* abundance, though this species did exhibit a slightly positive relationship with elevation.

Picea glaua predicted abundance is much higher than any other species at warm sites (Fig. 10). At cool sites, *Betula neoalaskana* abundance is predicted to be higher than other species, likely because individual *B. neoalaskana* trees tend to grow much larger than *Picea mariana* individuals, the other species which commonly occurs in cool sites. Total tree abundance is much higher at warm sites, across the range of observed elevations (Fig. 10). Interestingly, the peak abundance appears at higher elevations (approximately

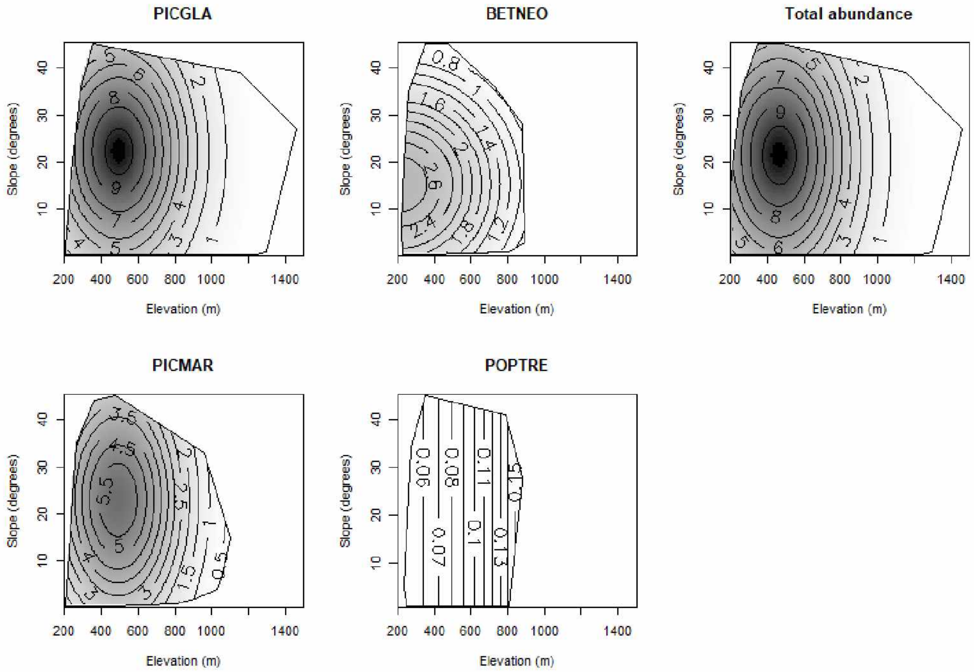

Figure 9. Contours showing predicted abundance (BA) across observed values of elevation and plot slope for each species. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species.

600 m) in warm sites than it does at cool sites (approximately 400 m), likely due to *Picea mariana* and *Betula neoalaskana*'s preference for lower elevation sites.

For all species, abundance was negatively correlated with recent burns (Fig. 11). However, old burn sites had higher abundance of the broadleaf species than unburned sites. *Picea glauca* was much less abundant in sites with any type of burn history, while *P. mariana* abundance was almost the same at old burn sites and unburned sites. The competing responses to fire of conifers and broadleaf species meant that overall tree abundance was approximately equal at unburned and old burn sites, with a sharp decrease at recently burned sites.

Soil pH was a significant predictor of abundance for every species but *Picea mariana* (Fig. 12). Increasing soil pH had a positive effect on overall tree abundance, *P. glauca* abundance, and *Populus tremuloides* abundance. *Betula neoalaskana* abundance, on the other hand, was negatively correlated with soil pH.

3.3.2. *Spatial effects*

In the abundance models, d_0 was significant only for *Picea mariana* (median 2510 m, 95% credible interval [1.4 km, 4.7 km]) and the all species abundance measure (median 1656 m, 95% credible interval [1.1 km, 2.4 km]). However, both *P. glauca* and *B. neoalaskana* also had quite a bit of posterior mass between 500 m and 3000 m. The location of these effective distances suggests that there may be within-mini-grid spatial

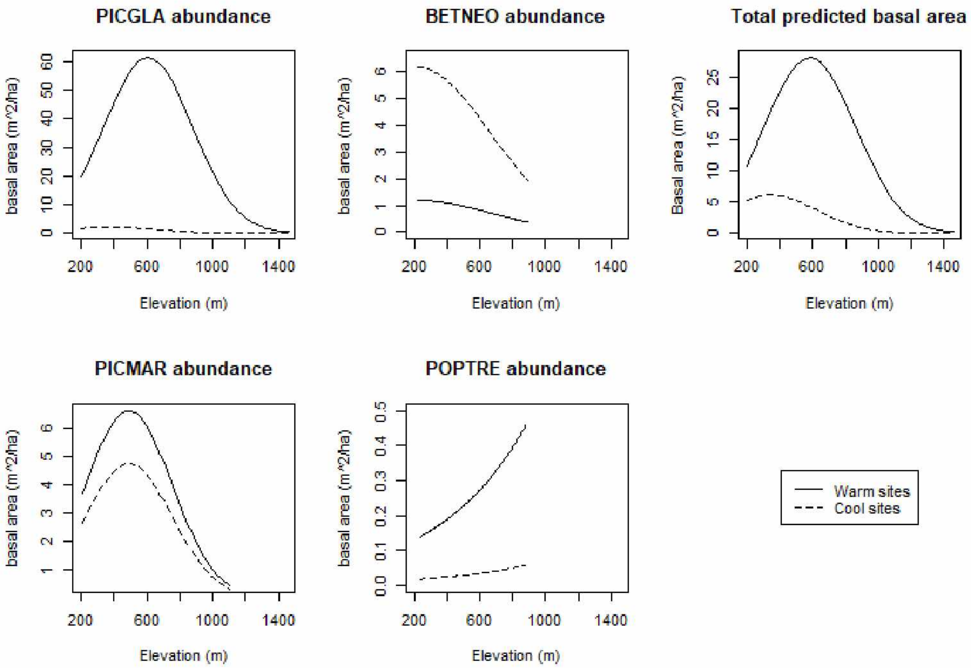

FIGURE 10. Predicted abundance (BA) across observed values of elevation at warm sites (high annual radiation [mean + 2 SD], deep soil depth [mean + 2 SD], shallow soil organic layer [mean - 1 SD], high soil pH [mean + 2 SD], no frozen soil, thawed alluvial terraces and at cool sites (low annual radiation [mean - 2 SD], shallow soil depth [mean - 2 SD], deep soil organic layer [mean + 2 SD], low soil pH [mean - 2 SD], not a thawed alluvial terrace, and frozen soil). All other variables were held to mean values. Lines are constrained to the elevations at which species were observed.

effects for these three species and the overall abundance measure. Only *Populus tremuloides* showed strong evidence of no spatial effect, with the majority of its posterior mass below the 500 m threshold. As with the occurrence models, these results suggest that spatial correlation could be more easily incorporated through the addition of mini-grid level random effects in a GLMM.

4. Discussion

4.1. **Ecological discussion**

In this paper we presented geostatistical species distribution models for the five tree species that occur in Yukon-Charley Rivers National Preserve (YUCH). These models were built using data from a 10,220 km^2 area in interior Alaska, meaning that they incorporate a wide range of ecological conditions in an area that is generally warmer and more forested than Denali National Park (DNPP). Generally, we found agreement between the models for DNPP (Roland, Schmidt, and Nicklen 2013) and for YUCH, validating these statistical techniques and affirming that both sets of models represent true underlying phenomena.

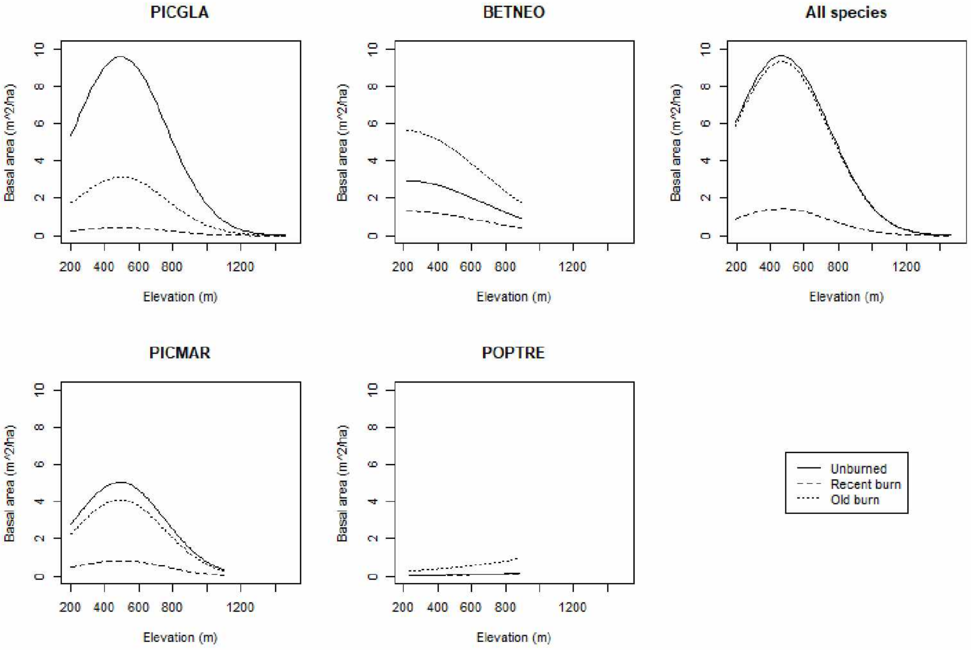

FIGURE 11. Predicted abundance (BA) across observed values of elevation under different fire histories. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species. Lines are constrained to the elevations at which species were observed.

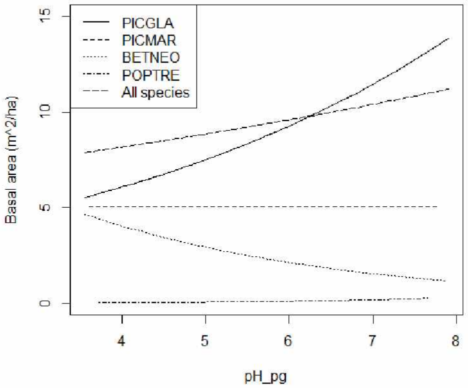

FIGURE 12. Predicted abundance of each species across the range of observed soil pH. Because the model parameters vary for each species, graphs are based on the mean parameter values from the best approximating spatial model for each species

However, there were important differences. Fire played a larger role in YUCH, with burn history being a significant predictor in every model. This is important because fire frequency and size are increasing in interior Alaska (Kasischke et al. 2010). We found that both *Picea glauca* occurrence and abundance were negatively correlated with recent and old burns, suggesting that an increase in fires may lead to a decrease in *P. glauca* in YUCH (Fig. 4, Fig. 11). *P. mariana*, on the other hand, was much more likely to occur in sites

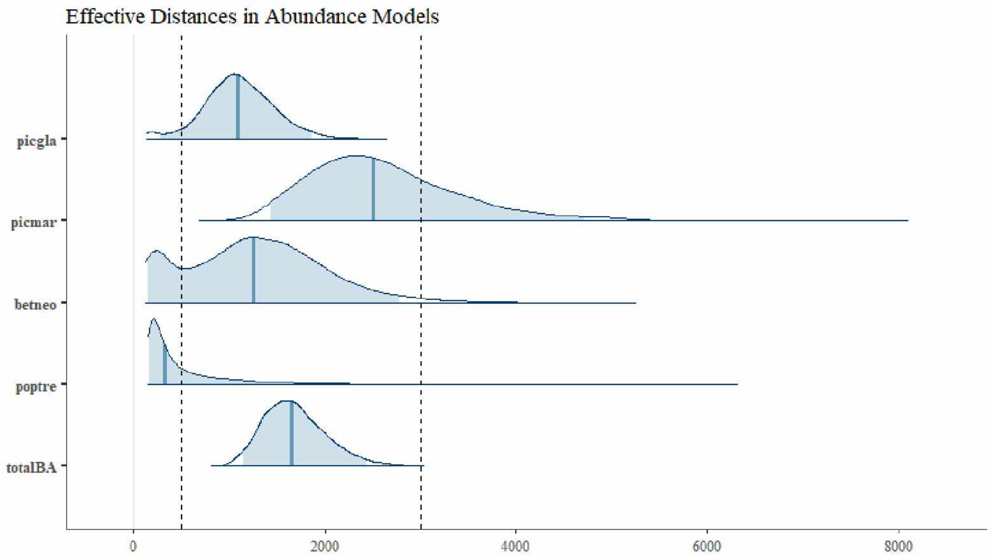

Figure 13. Posterior distributions of d_0, the effective distance parameter in the spatial abundance models. Vertical lines are at 500 m and 3000 m, which represent the minimum distance between regularly spaced plots and the maximum distance across a mini-grid. 18000 m, the distance between two mini-grids, is off the scale of this plot.

that had experienced a burn before 1982, with approximately equal predicted abundance in these sites as in unburned sites. This result is not unexpected, as *P. mariana* is a semi-serotinous species. The broadleaf species were all more likely to occur in burned sites, with *B. neoalaskana* and *P. tremuloides* showing a slight preference for sites which had burned recently, while *P. balsamifera* was more likely to occur in old burn sites (Fig. 4). Predicted abundance of both *B. neoalaskana* and *P. tremuloides* was highest at sites which had burned before 1982. Overall, this suggests that YUCH may be subject to a conversion from conifer (or at least, *P. glauca*) dominated forests to broadleaf-forest ecosystems, as has been observed recently throughout interior Alaska (Johnstone et al. 2010).

The occurrence models showed stronger evidence of spatial effects, over greater distances (higher d_0 values), than the abundance models. This result makes sense, as the drivers of occupancy are more likely to occur at a larger scale. Specifically, if a tree doesn't grow in one plot, its neighbors within a few kilometers may also be subject to similar conditions driving absence, such as issues with dispersal or being an alpine site which doesn't support trees. However, the drivers affecting abundance tend to be more local. There is little reason to expect that because a tree is abundant in one site, it would also be abundant in a site that is several kilometers distant.

Overall, the models presented in this paper give a comprehensive overview of the important drivers of tree species occurrence and abundance in YUCH. In addition to allowing comparisons with DNPP, these models provide a valuable baseline from which to examine changes in the coming decades.

4.2. **Incorporation of spatial effects**

Adding spatial effects to the models proved to be fairly difficult, due to the complications discussed in Appendix B and Appendix C. While including spatial effects was supported by the AIC and Tjur's D_{COD} values, this spatial correlation may be sufficiently explained using generalized linear mixed models (GLMMs) by adding a mini-grid level random effect. In particular, our effective distance (d_0) values demonstrate that spatial correlation between plots tends to be essentially gone at distances greater than 18 km for occupancy models (Fig. 7) and greater than 5 km for abundance models (Fig. 13). In both cases, this indicates that there is no significant spatial correlation between neighboring mini-grids. Therefore, allowing each mini-grid to have its own random effect term may effectively model these patterns. The benefit of using GLMMs instead of geostatistical GLMs is mainly in computational efficiency, as GLMMs can be fit using frequentist methods much more quickly than the Bayesian geostatistical GLMs presented here.

4.3. **Model building steps and difficulties**

In this paper we presented the process and results from creating 10 geostatistical species distribution models using count data. While models such as these exist in the literature, they are not straight forward to build and required a great number of steps and several interesting theoretical decisions.

The first decision, as in any model building process, was to choose a statistical model. Since we were interested in learning about two distinct responses, both where a species can grow (occupancy) and where it thrives (abundance), we chose not to use zero-inflated models. This was in large part because of the identifiability issues which prohibit using the same set of covariates in the two halves of the zero-inflated model (Zuur et al. 2009), and thus preclude any opportunity to learn about how a single environmental variable may affect occurrence and abundance differently. For example, *Betula neoalaskana* was more likely to occur in sites which had recently burned, but when it occurred in these sites the predicted abundance was low. In this case, occurrence was associated positively with recent burns, while abundance exhibited a negative association. This relationship would have been impossible to identify had we chosen to use zero-inflated models, rather than following the two-step model approach which we ultimately chose (§2.2.1).

Early in the process we also hoped to develop a joint species distribution model (JSDM), which would have incorporated all 5 of our species into a single model that could account for species' interactions as well as their responses to environmental variables (Clark et al. 2014; Kissling et al. 2012; Pollock et al. 2014; Thorson et al. 2015; Warton et al. 2015). However, these models proved excessively difficult to build, and this idea must be relegated to future work.

Having decided to create two distinct models for each species, we still needed to decide on a statistical form. For the occupancy models, this was straightforward. Since presence/absence is a binary response, we modeled it assuming a binomial distribution (§2.2.2). However, the abundance models proved more difficult, since we were using basal area, a continuous variable with a great number of zeros, as our response. After talk of several more complicated models, we settled on the ordinal regression model presented in Min and Agresti 2002, rounded our basal area data up to the nearest integer, and treated it as count data from then on. However, we were still left with a decision between assuming a Poisson distribution or a negative binomial distribution. As discussed in §2.2.2 and Appendix C, we ultimately used the negative binomial distribution for the nonspatial models before switching to the Poisson distribution for the spatial models.

Once we had settled on statistical models, the next step was to determine which of many potential covariates to include. After narrowing our list to only those variables which we had theoretical basis to believe belonged (Appendix A), we chose to use frequentist variable selection techniques to choose our best approximating nonspatial models (§2.2.2). While many Bayesian variable selection criteria have been proposed (see, e.g. Hooten and Hobbs 2015), none of them are able to compete with the computational speed and ease of implementing stepwise selection using AIC. The relative complexity of our models (choosing between 18 potential covariates for each of 10 distinct models) made this hybrid half-frequentist, half-Bayesian modeling approach our most feasible option.

Having selected these nonspatial models, the next step was to add spatially correlated errors. Adding spatial effects to a categorical model is more complicated than adding spatial effects to a simple linear model, due to the link function. This requires that spatial effects be added to the mean structure, which leads to potential complications. In particular, we found evidence of confounding between the site specific variance term, σ^2, and the coefficient estimates in our occupancy models (Appendix B). We suspect that this is a common issue in binomial models with either spatially correlated errors or a random effect term (GLMMs), but that it is often missed. Since the confounding leads to inflated coefficient estimates, it can lead to misleading results. After a great deal of problem solving, during which we attempted building GLMMs but found the same issue, we ultimately chose to set $\sigma^2 = 1$, as site specific variance is unidentifiable for binary data anyway (Hadfield, personal communication, 8/4/17). We believe that this topic deserves further research since we were unable to find anything in the literature addressing it, even though it has potentially far-reaching consequences.

A final difficulty lay in finding measures of model fit (§2.4), a challenging question when dealing with complicated models. We settled on using DIC as a simple measure of whether the added complexity of spatial effects was justified by the improvement in the models, and found that it was for every species. However, DIC values are only useful relative to one another, and don't contain any information about the overall fit. In order to gauge model fit we chose to use Tjur's D_{COD} for the occupancy models and proportional reduction in deviance (D_{PRD}) for the abundance models, though there are many other options which have various attractive and unattractive qualities (Cameron and Windmeijer 1996). Unfortunately, we were unable to find any measure of model fit which could be calculated for both the frequentist nonspatial abundance models and the Bayesian spatial abundance models in order to compare them.

In the end, this project presents a road map to others contemplating building spatially explicit species distribution models using count data. We present here the steps, some of the questions to consider, a few pitfalls, and their solutions. In the process, we describe distribution and abundance patterns of the five tree species that occur in Yukon-Charley Rivers National Preserve, patterns which lend insight into a unique region of Alaska and may provide some idea of what the Preserve may look like in the coming decades. Overall, species distribution models are a powerful tool with which to describe species' habitat preferences, and they deserve a place in any quantitative ecologist's toolbox.

Acknowledgements

I would like to thank my advisor, Margaret Short, for help and advice throughout the modeling and writing process. I would also like to thank Carl Roland for providing the data, mentorship, and an internship to conduct this research. Funding was provided by the Geoscientists in the Parks (GIP) internship program, and data was generously provided by the National Park Service. Thank you as well to my committee, for help and constructive feedback.

References

Agresti, Alan (2013). *Categorical Data Analysis*. 3rd ed. Hoboken, New Jersey: Wiley.

Anderson, Barbara J., Alessandro Chiarucci, and Mark Williamson (2012). "How differences in plant abundance measures produce different species-abundance distributions". In: *Methods in Ecology and Evolution* 3.5, pp. 783–786. ISSN: 2041-210X.

Bolker, Ben (2015). *GLMM worked examples*. GLMM worked examples from Ecological Statistics: Contemporary Theory and Applications. URL: `https://ms.mcmaster.ca/~bolker/R/misc/foxchapter/bolker_chap.html` (visited on 01/17/2018).

Cameron, A. Colin and Frank A.G. Windmeijer (1996). "R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization on JSTOR". In: *Journal of Business & Economic Statistics* 14.2, pp. 209–220.

Clark, James S. et al. (2014). "More than the sum of the parts: forest climate response from joint species distribution modesl". In: *Ecological Applications* 24.5, pp. 990–999. ISSN: 1051-0761.

Dubayah, Ralph and Paul M. Rich (1995). "Topographic solar radiation models for GIS". In: *International Journal of Geographical Information Systems* 9.4, pp. 405–419. ISSN: 0269-3798.

Finley, Andrew O., Sudipto Banerjee, and Bradley P. Carlin (2007). "spBayes: An R Package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models". In: *Journal of Statistical Software* 19.4, pp. 1–24. ISSN: 1548-7660.

Gelman, Andrew et al. (2014). *Bayesian Data Analysis*. 3rd ed. Texts in Statistical Science. CRC Press.

Guisan, Antoine and Wilfried Thuiller (2005). “Predicting species distribution: offering more than simple habitat models”. In: *Ecology Letters* 8.9, pp. 993–1009. ISSN: 1461-023X, 1461-0248.

Hooten, M. B. and N. T. Hobbs (2015). “A guide to Bayesian model selection for ecologists”. In: *Ecological Monographs* 85.1, pp. 3–28. ISSN: 1557-7015.

Johnstone, Jill F., Teresa N. Hollingsworth, and F. Stuart Chapin III (2008). *A key for predicting postfire successional trajectories in black spruce stands of interior Alaska.* General Technical Report PNW-GTR-767. Pacific Northwest Research Station: USDA Forest Service.

Johnstone, Jill F. et al. (2010). “Fire, climate change, and forest resilience in interior Alaska”. In: *Canadian Journal of Forest Research* 40.7, pp. 1302–1312. ISSN: 0045-5067, 1208-6037.

Joseph, Liana N. et al. (2009). “Modeling abundance using N-mixture models: the importance of considering ecological mechanisms”. In: *Ecological Applications: A Publication of the Ecological Society of America* 19.3, pp. 631–642. ISSN: 1051-0761.

Kasischke, Eric S. et al. (2010). “Alaskas changing fire regime implications for the vulnerability of its boreal forests”. In: *Canadian Journal of Forest Research* 40.7, pp. 1313–1324. ISSN: 0045-5067.

Kissling, W. D. et al. (2012). “Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents: Modelling multispecies interactions”. In: *Journal of Biogeography* 39.12, pp. 2163–2178. ISSN: 03050270.

Latimer, Andrew M. et al. (2006). “Building statistical models to analyze species distributions”. In: *Ecological Applications* 16.1, pp. 33–50.

Lichstein, Jeremy W. et al. (2002). “Spatial Autocorrelation and Autoregressive Models in Ecology”. In: *Ecological Monographs* 72.3, pp. 445–463. ISSN: 1557-7015.

MacCluskie, Maggie et al. (2005). *Central Alaska Network: Vital Signs Monitoring Plan.*

Min, Yongyi and Alan Agresti (2002). “Modeling nonnegative data with clumping at zero: a survey”. In: *Journal of the Iranian Statistical Society* 1.1, pp. 7–33.

Nicklen, E. Fleur et al. (2016). “Local site conditions drive climate-growth responses of *Picea mariana* and *Picea glauca* in interior Alaska”. In: *Ecosphere* 7.10, e01507. ISSN: 21508925.

Pollock, Laura J. et al. (2014). “Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)”. In: *Methods in Ecology and Evolution* 5.5. Ed. by Jana McPherson, pp. 397–406. ISSN: 2041210X.

Potts, Joanne M. and Jane Elith (2006). “Comparing species abundance models”. In: *Ecological Modelling.* Predicting Species Distributions 199.2, pp. 153–163. ISSN: 0304-3800.

R Core Team (2017). *R: A language and environment for statistical computing.* Version 3.4.0. Vienna, Austria.

Rich, P. et al. (1994). “Using viewshed models to calculate intercepted solar radiation: applications in ecology. American Society for Photogrammetry and Remote Sensing Technical Papers”. In: *American Society of Photogrammetry and Remote Sensing*, pp. 524–529.

Roland, Carl A., Joshua H. Schmidt, and E. Fleur Nicklen (2013). “Landscape-scale patterns in tree occupancy and abundance in subarctic Alaska”. In: *Ecological Monographs* 83.1, pp. 19–48.

Roland, Carl A. et al. (2004). *Monitoring vegetation structure and composition at multiple scales in the Central Alaska Network.* NPS Technical Report CAKN-001. Fairbanks, Alaska, USA: National Park Service.

Rosenthal, Jeffrey S. (2007). “AMCMC: An R interface for adaptive MCMC”. In: *Computational Statistics & Data Analysis* 51.12, pp. 5467–5470.

Shulski, Martha and Gerd Wendler (2007). *Climate of Alaska.* Fairbanks, AK: University of Alaska Press. 208 pp. ISBN: 978-1-60223-007-1.

Stan Development Team (2016). *RStan: the R interface to Stan.* R package version 2.14.1.

Thorson, James T. et al. (2015). “Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range”. In: *Methods in Ecology and Evolution* 6.6. Ed. by David Warton, pp. 627–637. ISSN: 2041210X.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S.* Fourth. New York: Springer.

Ver Hoef, Jay M. et al. (2001). “Uncertainty and Spatial Linear Models for Ecological Data”. In: *Spatial Uncertainty in Ecology.* New York: Springer, pp. 214–237.

Vincent, P. J. and J. M. Haworth (1983). “Poisson Regression Models of Species Abundance”. In: *Journal of Biogeography* 10.2, pp. 153–160. ISSN: 0305-0270.

Warton, David I. et al. (2015). “So Many Variables: Joint Modeling in Community Ecology”. In: *Trends in Ecology & Evolution* 30.12, pp. 766–779. ISSN: 01695347.

White, Gary C. and Robert E. Bennetts (1996). “Analysis of Frequency Count Data Using the Negative Binomial Distribution”. In: *Ecology* 77.8, pp. 2549–2557. ISSN: 0012-9658.

Wilson, J. Bastow (2011). “Cover plus: ways of measuring plant canopies and the terms used for them”. In: *Journal of Vegetation Science* 22.2, pp. 197–206. ISSN: 1100-9233.

Zheng, Beiyao and Alan Agresti (2000). “Summarizing the predictive power of a generalized linear model”. In: *Statistics in medicine* 19.13, pp. 1771–1781.

Zuur, Alain et al. (2009). *Mixed Effects Models and Extensions in Ecology with R.* 1st ed. Statistics for Biology and Health. New York: Springer-Verlag New York.

Appendix A. Narrowing the list of potential predictors

Initially, we considered nearly 40 potential environmental covariates. However, several of these variables are highly correlated, either because they are directly calculated from one another (for example aspect and solar radiation) or because they encode approximately the same information about environmental conditions (e.g. temperature tends to decrease as elevation increases, total annual solar radiation has a very close relationship with summer solar radiation, etc.). Fifteen of the original potential predictors were dropped after determining, based on expert opinion, that other variables better reflected the environmental conditions we cared about. There were still several highly related variables, and with no clear theoretical reason to prefer one over the other we compared AIC values of frequentist models including one then the other variable. The variable whose inclusion resulted in a lower AIC value was kept, while the other was dropped. At the end of these processes, we had 15 variables that remained for consideration for every model (Table 1).

Appendix B. Issues with confounding

After fitting our occupancy models using vague priors as described in §2.3.2, we noticed an unusual and unexpected result. Our σ^2 term, which describes site-specific variance, had become extremely large in all of our occupancy models (taking values between 40 and 90). This is much larger than expected for logistic regression, as parameter estimates greater than approximately ±10 generally indicate a problem (Bolker 2015). Furthermore, in the models that had these large values of σ^2, our βs (the estimates of the parameter coefficients), had taken on far more extreme values than in the models without the spatial effects.

Ultimately, we discovered a discussion of a very similar problem (Bolker 2015), in which the author was fitting binary generalized linear mixed models (logistic GLMMs). These mixed models are very similar to our spatial models, with the only difference being that in the GLMM the error terms are assumed to be independent, rather than drawn from a spatially correlated distribution. However, the GLMMs still contain a σ^2 term that describes site-specific (aka observation-level) variance. Bolker describes how, because the model is

$$\operatorname{logit}(p_i) = m_i + \epsilon_i$$

where m_i in our case includes all our covariates, the expected value of p_i is the average of logistic$(m_i + \epsilon_i)$, which is a nonlinear average. So, the mean of p_i is not equal to logistic(m_i), as described by Jensen's inequality. As a result, the estimates of m_i and σ^2 are confounded.

This explains our strangely large values of σ^2 and of the βs, because as σ^2 crept upwards in the MCMC iterations, the β estimates crept steadily outward (towards more extreme values) to compensate. Having identified the problem, we followed advice from Jarrod Hadfield (discussed in Bolker's worked example, and confirmed via personal communication, 8/4/17) to fix σ^2 at 1. Since observation level variance is unidentifiable for binary data (Hadfield, personal communication, 8/4/17), this should not result in a loss of information.

Interestingly, we did not experience similar issues with fitting the abundance models, something which could be investigated further in future work.

Appendix C. Negative binomial vs. Poisson distribution

The negative binomial distribution was chosen for the variable selection phase because of its ability to account for overdispersion (the larger number of zeros than would be expected under a Poisson distribution). However, no existing software packages have built a function to evaluate a negative binomial spatial model, as `spBayes` and geoRglm both allow for only binomial and Poisson distributions. This is likely due to the fact that a spatially explicit negative binomial model ends up being very similar to a spatially explicit Poisson model, as explained below.

A careful investigation into the derivation of the negative binomial model (and its possible interpretation as a Poisson-Gamma model), showed that it can be derived by adding a random intercept term to the Poisson mean structure. In the case of the negative binomial distribution, this random intercept is modeled as a Gamma distributed random variable. Creating a geostatistical model is essentially the process of adding a spatially explicit error term to a non-spatial model. In the case of categorical models, this term is, by necessity, added to the systematic component of the model, before the link function is applied. Therefore, adding a spatially correlated error term to a negative binomial distribution is, in some ways, duplicating the extra-mean variance term that is already included in the negative binomial from adding an independent error term to the Poisson distribution. There are potentially issues with identifiability at this point, as well as the creation of a particularly complicated categorical model. While there are minor differences in the structure of the mean-variance relationship between a negative binomial that includes a spatially explicit error term and the Poisson model with a spatial term, the two models end up being fairly similar. These similarities, and the idea that adding a spatial term to the Poisson model may seem simpler to understand and code, may explain the fact that the currently developed software packages do not include spatially explicit negative binomial distributions (Ver Hoef, personal communication 7/18/17).

While I ultimately chose to model the abundance data using a spatially explicit Poisson model, I first built a spatially explicit negative binomial model in STAN for *Picea glauca*, in order to test whether the two models gave me approximately equivalent results. The STAN model ran much more slowly than the spBayes model, but the β estimates converged rather quickly. However, there were lingering problems in the convergence of the spatial components, which perhaps could be fixed with tuning. Nonetheless, the β estimates were quite close to the estimates given by the spatial Poisson model, and I decided to model abundance using the Poisson model for simplicity's sake.

C.1. **STAN code for a spatially explicit negative binomial model**

```
// Pic gla abundance spatial model.

// negative binomial parameterized as eta (log(mu)) and dispersion (phi)
// note that phi in stan is equivalent to theta in glm.nb, and named as such here
// see p286 in stan-reference-2.4.0.pdf

data {
  int<lower=1> N; // rows of data
  int<lower=1> P; // # predictors

  matrix[N,N] dist;   //distance matrix
  int<lower = 0> BA[N]; // neg binom response
  matrix[N,P] X; // predictors
}

parameters {
  // spatial parameters
  real<lower=0> sigma2;
  real<lower=3.0/117000, upper = 3.0/200> phi;
  vector[N] w;    // spatial random effects term

  // covariate parameters
  real<lower=0> theta;
  vector[P] beta;
}

model {
  matrix[N,N] K_phi; // variance covariance matrix for gaussian process
  vector[N] zero_vec; // zero mean for gp
  zero_vec = rep_vector(0, N);

  // vectorized
  K_phi = sigma2 * exp( -phi * dist);

  // priors
  sigma2 ~ inv_gamma(2.1, 10);
  theta ~ normal(0,5);
  beta ~ normal(0,5);

  // data model
  w ~ multi_normal(zero_vec,K_phi);
  BA ~ neg_binomial_2_log(X * beta + w, theta);
}
```