Bayesian Spatial Analysis of Hardwood Tree Counts in Forests via MCMC

Reihaneh Entezari[*1], Patrick E. Brown[†1,2], and Jeffrey S. Rosenthal[‡1]

[1]Department of Statistical Sciences, University of Toronto

[2]Centre for Global Health Research, St Micheal's Hospital

October 9, 2019

Abstract

In this paper, we use a Bayesian spatial model to spatially interpolate forest inventory data from the Timiskaming and Abitibi River forests in Ontario, Canada. We consider a Bayesian Generalized Linear Geostatistical Model and implement a Markov Chain Monte Carlo algorithm to sample from its posterior distribution. How spatial predictions for new sites in the forests change as the amount of training data is reduced is studied and compared with a Bayesian Logistic Regression model without a spatial effect. Finally, we discuss a stratified sampling approach for selecting subsets of data that allows for potential better predictions.

[*]entezari@utstat.utoronto.ca
[†]patrick.brown@utoronto.ca web: http://pbrown.ca
[‡]jeff@math.toronto.edu web: http://probability.ca/jeff/

1 Introduction

1.1 The forest inventory problem

The forest industry is a substantial part of the economies of many countries, and a "forest inventory" is an estimate of monetary value of the timber resources in a specific managed area. The value of a timber resource depends on different features of trees such as size, species, age, defects, etc. Tree species have different types of wood with different qualities, and hence influence the timber value.

Tree species have two main categories, hardwood (deciduous) trees and softwood (coniferous) trees, with hardwood trees generally having wider leaves that are lost annually, while softwood trees have smaller leaves and retain their leaves throughout the year. Hardwood trees provide much longer lasting wood compared to softwood trees, with slower growth rates which makes them more expensive compared to softwood. Hence, knowing the number of hardwood trees in a forest is valuable information. Collecting data on forests requires hiring workers to travel to different sites around the forests and measure the quantities needed, which can be costly and time consuming.

Remote sensing technologies can overcome this issue. Although they are cheap and efficient and can cover a wide range of geographical areas, they can suffer from lack of accuracy. Geostatistical models are powerful tools for analyzing and predicting such spatial data, and can be used to calibrate remotely sensed data (see Curran & Atkinson, 1998). Existing literatures by Giorgi et al. (2017); Shaby & Reich (2012); Abellan et al. (2007) are examples of the importance of statistical models for spatial analysis. The focus of this paper will also be to take advantage of statistical tools to predict the number of hardwood trees using geostatistical models that take into account the spatial factor.

1.2 Model-based geostatistics

In the past few decades, spatial statistics has become an established field of statistics with well developed models applied to many real-world problems. Conventional geostatistical models for Gaussian spatial data were first popularized by Matheron (1962) and later on built upon by Cressie (1993). The generalization of these models for non-Gaussian data were introduced by Diggle et al. (1998).

Let Y_i be the observed spatial data at location s_i, with arbitrary distribution f that has mean λ and possible additional parameters γ. Consider $X(s_i)$ as the covariates at location s_i. Modelling this data with the Generalized Linear Geostatistical Model (GLGM) described in Diggle et al. (1998) and Diggle & Ribeiro (2007), will be as following:

$$\begin{aligned} Y_i|U(s_i), \lambda(s_i), \gamma &\sim f[\lambda(s_i), \gamma] \\ g[\lambda(s_i)] &= \mu + \beta X(s_i) + U(s_i) \end{aligned} \tag{1}$$

where $g(.)$ is the link function (i.e. logit or log). Here $U(s)$ is a Gaussian random field U evaluated at location s, which is characterized by the joint multivariate normal distribution:

$$[U(s_1), ..., U(s_N)]^T \sim MVN(0, \Sigma)$$

where the elements of Σ are defined by a spatial correlation function ρ as

$$\Sigma_{ij} = \text{cov}[U(s_i), U(s_j)] = \sigma^2 \rho(||s_i - s_j||/\phi, \nu)$$

where ϕ is a range parameter and ν is a vector of other possible parameters. The range parameter ϕ controls the rate at which the correlation decreases with distance. There are many possible parametric functions for ρ, with Matérn correlation function (see Stein, 1999)

being the most commonly used. The Matérn correlation is defined as:

$$\rho(h;\phi,\kappa) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\Big(\frac{||h||}{\phi}\Big)^{\kappa} K_{\kappa}\Big(\frac{||h||}{\phi}\Big), \tag{2}$$

where $\Gamma(.)$ is the gamma function and $K_{\kappa}(.)$ is the modified Bessel function of the second kind of order $\kappa > 0$ (κ being a shape parameter). This function is particularly interesting, as it is flexible in the differentiability of the Gaussian process $U(s)$ by adjusting κ (Stein, 1999).

Bayesian inference is the dominant paradigm for use with GLGM's due to the difficulty in computing Maximum Likelihood Estimates. Although methods for Frequentist inference are now well developed and software available, the Bayesian treatment of parameter uncertainty remains a strong justification for Bayesian inference in many circumstances. Bayesian inference via Markov Chain Monte Carlo (MCMC) methods (Brooks et al., 2011; Craiu & Rosenthal, 2014) has many advantages as discussed in Diggle et al. (1998). The Integrated Nested Laplace Approximation (INLA) algorithm introduced by Rue et al. (2009), is an alternative to MCMC for Bayesian Inference on latent Gaussian models. In particular INLA makes numerical approximations to the marginal posterior distributions rather than the joint distributions which we will describe in detail in section 2.5. There are facilities in the R-INLA software for producing approximate joint posterior samples, but the properties of these samples have yet to be explored.

1.3 The motivating problem

In this paper, we will analyze the spatial hardwood tree count data collected from the Timiskaming & Abitibi River forests in Ontario, Canada. Our analysis is constructed in a Bayesian framework for a binomial geostatistical model to predict the proportion of hardwood trees from remotely sensed elevation and vegetation data. For posterior simulations, we implement an MCMC method using the Langevin-Hastings (see Roberts & Rosenthal,

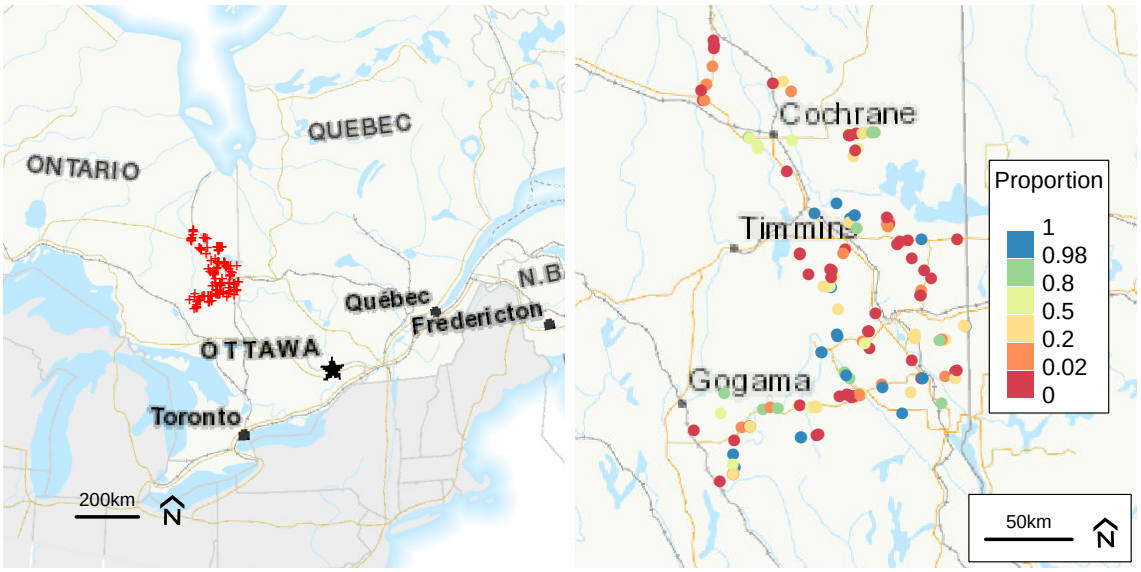
1998) and the Random-Walk Metropolis Hastings (see Roberts et al., 1997; Roberts & Rosenthal, 2001) algorithms. By reducing the amount of training data fitted to the model, and evaluating the out-of-sample predictive performance for the same validation set (which also accounts for model misspecification), we are able to mimic a scenario where fewer ground truth measurements are collected and assess the usefulness and accuracy of less costly forest inventories. We will show that with training data size as small as 10 spatial locations, despite the increase in uncertainty, the true number of hardwood trees lies within a 95% prediction interval. This conclusion is very valuable as it will significantly reduce costs of collecting ground truth data. We will also compare our results with a Bayesian Logistic Regression model where there is no spatial effect.

A secondary consideration is to evaluate the need for a well considered spatial sampling design, which can potentially increase the cost of data collection by requiring visites to inaccessible locations. Many existing papers (see Wang et al., 2012; Brus & De Gruijter, 1997) discuss the importance of design-based sampling for spatially correlated data in order to improve estimation of population parameters. Therefore we also explore a stratified sampling approach in choosing the training data that will show a potential improvement in the predictions.

The paper is organized as follows. The spatial data from the Timiskaming & Abitibi River Forests are described in section 2.1. Section 2 describes the geostatistical model used for our data and the MCMC algorithm applied to perform Bayesian Inference. In addition, we explain our stratified approach and describe the measurements we will use to compare and assess predictions. Section 3 discusses the numerical results from fitting the data, where comparisons are also made with the Bayesian Logistic Regression. At last, we summarize our results in Section 4. The Appendix includes details of the implementation of the MCMC algorithm, and results from different simulations are presented in the Supplemental Document.

2 Methods

2.1 Description of Data



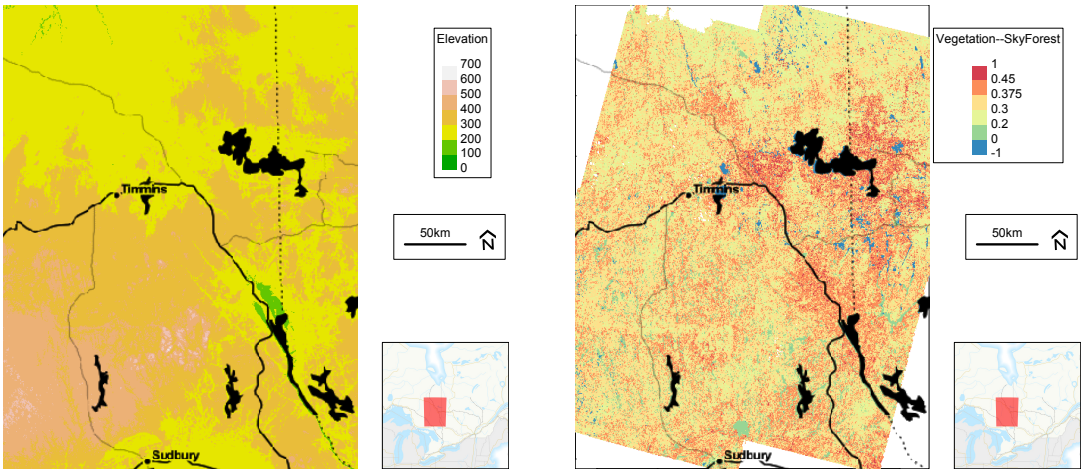
(a) Sample locations (b) Proportions of hardwood trees

Figure 1: Locations of 162 forest plots in the Timiskaming and Abitibi River Forests (background Natural Resources Canada).

The Timiskaming and Abitibi River forests are geographically located next to one-another in northern Ontario, Canada. The First Resource Management Group Inc. has provided detailed data from 162 individual forest plots inside these adjacent forests. Each forest plot is 11.28m in radius to provide a 400m^2 circular surface. The geographical locations of these 162 sites are shown in Figure 1.

The data from each site consists of information on the total number of trees, whether each tree is living or dead, and the species of each tree. Figure 1b shows the proportion of live trees which are hardwood from the 162 sites. As can be seen, many sites have no hardwood trees and such sites are scattered throughout the forests.

The remotely sensed data considered includes elevation values from satellite data provided by the SRTM program (Figure 2a). A measure of forest vegetation was provided by the First Resource Management Group Inc. using the proprietary remote sensing technol-



(a) Elevation (b) SkyForest^{TM} vegetation index

Figure 2: Elevation & Vegetation index around the Timiskaming and Abitibi River Forests (Background ©Stamen Design).

ogy "SkyForest^{TM}", which is shown in Figure 2b. This vegetation measure is predicted by SkyForest^{TM} across the forest landscape by selecting an arithmetic transformation of spectral bands (ATSB) from a candidate list of ATSBs. The ATSBs are constructed similarly to well known vegetation indices such as the Normalized Difference Vegetation Index (NDVI), with some of them being multi-temporal. It is thus expected that hardwood trees are located where this measure is high.

In the next section, we will describe the geostatistical model for our dataset, along with the steps taken to perform a Bayesian analysis.

2.2 Logistic Regression

Before describing the full geostatistical model for our data, a simple Logistic Regression model with binomial response will be outlined. Consider Y_i to be the count of hardwood trees in forest plot i, and write $Y_i \sim \text{Binom}(n_i, p_i)$, where n_i is the total number of live trees at site i (s_i) and p_i is the probability of a tree in plot i being hardwood. Elevation and

the SkyForest^{TM} index are covariates in the model. The SkyForest^{TM} covariate is treated as a linear effect with change point at 0.3 (approximately its average value), giving some additional flexibility to this covariate in the regression model. The elevation values are also centered at the average value of about 320. For computational reasons, we normalize the covariates by dividing by the standard deviation. The model is:

$$\begin{aligned} Y_i &\sim \text{Binom}(n_i, p_i) \quad i = 1, ..., 162 \\ \log\left(\frac{p_i}{1-p_i}\right) &= X(s_i)\beta \end{aligned} \tag{3}$$

Writing $A(s)$ as the SRTM-measured altitude at location s and $V(s)$ as the SkyForest^{TM} vegetation index, the normalized vector of covariates $X(s)$ is constructed by:

$$\begin{aligned} X_1(s) &= 1 \\ X_2(s) &= \frac{A(s) - 320}{50} \\ X_3(s) &= \frac{\min(V(s) - 0.3, 0)}{0.05} \\ X_4(s) &= \frac{\max(V(s) - 0.3, 0)}{0.05} \end{aligned}$$

2.3 The geostatistical model

Spatial dependence in the prevalence of hardwood trees should be expected as sites in the forests close to one another may benefit from the same soil, weather, etc, and hence may have similar tree types. Thus we expect a geographical effect to play an important role in explaining such data with a more sophisticated model such as the Generalized Linear Geostatistical Model (GLGM). A geostatistical model for our spatial data will have an extra

spatial term $U(s)$ and an independent term Z compared to the model in (3), resulting:

$$\begin{aligned} Y_i &\sim \text{Binom}(n_i, p_i) \quad i = 1, ..., 162 \\ \log\left(\frac{p_i}{1-p_i}\right) &= t_i = X(s_i)\beta + U(s_i) + Z_i \end{aligned} \tag{4}$$

where Z_i's are mutually independent zero-mean Gaussian variables which are also independent of $U(s_i)$:

$$Z_i \overset{i.i.d.}{\sim} N(0, \tau^2),$$

$$U(s) \sim N(0, \sigma^2),$$

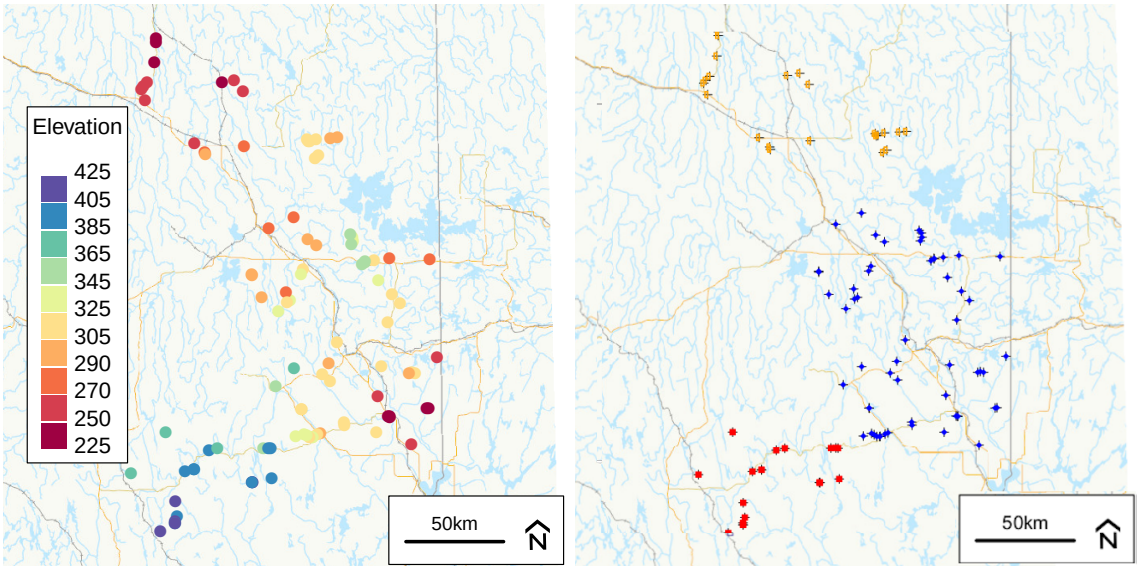
$$\text{cov}(U(s+h), U(s)) = \sigma^2 \rho(||h||; \phi, \kappa)$$

This model is equivalent to (1) where f is Binomial and g is a logit link function.

2.4 Out of sample predictions

For our analysis, we explore reducing the size of the training data fitted to the model, to observe and examine the trade-off between prediction accuracy and costs of collecting ground truth data. More specifically, if we were only given data from 25 or 10 plots on the ground, could useful predictions still be made? To answer this question, the 162 plots in the dataset were divided into 100 training and 62 validations sets. Keeping the 62 validation set fixed, we examine the performance of results generated by fitting 100, 25, and 10 training data to the model. For this purpose, we can do this by two different approaches, 1) choosing random subsets of data and 2) choosing stratified subsets of data. Since the spatial data is correlated, choosing the subset of data with a stratified approach should be expected to improve the results, as it can force the training plots to be as scattered as possible. Both elevation and vegetation covariates are taken into account for choosing the 25 and 10 dataset from the 100. Hence, we begin by looking at the elevation from all the 100 training data

(first simulation) as shown in Figure 3a.



(a) Elevation for 100 training data (b) Stratified regions

Figure 3: Plots of elevation from 100 training data, along with the plot of stratified regions.

The 100 plots are stratified into three groups based on their spatial locations as well as extreme elevation values (min, mid-point, max) (shown in Figure 3b). Keeping the proportion of the data from each strata constant, we systematically sample 25 plots from the 100 by sorting the vegetation index in each strata and taking every $j-th$ element depending on the number of data needed (similarly for the 10 data points from the 25). The Results section will explore how stratified sampling can (potentially) improve prediction accuracy with smaller training data fitted to the model, compared to random sampling.

2.5 Inference

We will apply a Bayesian approach to the model in (4), and this methodology will be referred to as the Bayesian Generalized Linear Geostatistical Model (BGLGM). Let $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \beta_3)$, $\boldsymbol{\theta}^T = (\sigma^2, \phi, \tau)$, and $\mathbf{t}^T = (t_1, ..., t_n)$ with $t_i = X(s_i)\beta + U(s_i) + Z_i$, be the three sets of parameters. We treat κ as fixed at 1.5, since it is not of direct interest and according to Zhang (2004), not all the parameters (σ^2, ϕ, and κ) are consistently estimable.

We define priors for each parameter as

$$\theta \sim \pi_1(.) \quad \& \quad \beta|\theta \sim \pi_2(.) = N(\mu, \sigma^2\Omega) \quad \& \quad t|\beta, \theta \sim \pi_3(.) = MVN(X\beta, \Sigma(\theta))$$

where the prior distribution $\pi_1(.)$ on θ is the product of the prior distributions on σ^2, ϕ, and τ. Thus the joint posterior distribution will be given as:

$$\pi(\beta, \theta, t|y) \propto \pi_1(\theta)\pi_2(\beta|\theta)\pi_3(t|\beta, \theta)f(y|t) \tag{5}$$

where $f(y|t) = \prod_{i=1}^{n} f_i(y_i|t_i)$ is the likelihood function. Here $\Sigma(\theta)$ is the covariance matrix with diagonal elements equal to $\sigma^2 + \tau^2$ and off-diagonal elements of $\sigma^2\rho(||s_i - s_j||; \phi, \kappa)$ where ρ is the Matérn correlation function. We consider *Exponential*($\lambda = 0.5$) priors for σ and τ, and a *Gamma*($\alpha' = 3, \beta' = 35$) prior for ϕ.

The INLA methodology from Rue et al. (2009) has become the dominant tool for inference with the BGLGM, and some explanation as to why INLA is not suitable for the forest inventory problem is warranted. The defining feature of INLA is it computes only marginal posterior distributions such as $\pi(\beta_p|y)$ and $\pi[U(s_i)|y]$ and linear combinations of the latent variables, but not the joint posterior distributions $\pi[\beta_p, U(s_i)|y]$. Inferring the prevalence of hardwood trees involves non-linear functions of the random effect $U(s)$ evaluated for all locations s, specifically sums of inverse logit transforms of the t_i. Obtaining posterior distributions of non-linear quantities is easily accomplished with MCMC algorithms, as MCMC outputs samples from the joint posterior distribution $\pi(t|y)$ and these samples can be used to compute any quantity of interest.

The main drawback of MCMC methods as compared to INLA is the former are computationally more intensive, and for this application computational concerns have proven to be minor. The number of ground truth plots s_i is small (10 to 100) and the intention for this methodology is to be applied to problems on the lower end of this range. Making spatial predictions over the entire region of interest is a high-dimensional problem, although this

can be done outside of the MCMC algorithm using a thinned set of posterior samples.

The recently developed **PrevMap** package (Giorgi & Diggle, 2017) uses a Hamiltonian Monte Carlo method to generate joint posterior draws, and has demonstrate impressive performance for a variety of practical spatial problems. Unfortunately the **PrevMap** package has struggled with our specific problem of forest inventories, particularly when the number of data points fitted were very small (as shown in section 3.1), and a customized MCMC algorithm was developed specifically for this application.

The bespoke MCMC method was implemented through the reparameterizations recommended by Christensen et al. (2006) that has shown to help facilitate the choice of proposal densities as well as reducing the correlation between variables that will significantly improve mixing and convergence of the MCMC algorithm.

Algorithm 1: MCMC algorithm

1. Initialize θ, β, and t
2. Transform to $\tilde{\theta}, \tilde{\beta}$, and $\tilde{t}$
3. Update $\tilde{\theta}_1, \tilde{\theta}_2$ and $\tilde{\theta}_3$ using a RWMH, each with standard deviation s_i calculated iteratively as:
$$s_i = s_{i-1} + c_1 i^{-c_2}(\alpha_i - 0.45)$$
where $c_1 > 0$ and $c_2 \in (0, 1]$ are constants, and α_i is the acceptance probability up to $i - th$ iteration with optimal acceptance probability of 0.45.
4. Update $\tilde{\beta}$ using a RWMH
5. Update $\tilde{t}$ with a Langevin-Hastings algorithm, i.e. $\tilde{t}' \sim \text{MVN}(\tilde{t} + 0.5h\nabla \log \pi(\tilde{t}), hI)$ where h is recommended to be $1.65^2/n^{1/3}$.
6. Repeat steps 3-5 until the desired number of samples are collected.
7. Transform samples of $\tilde{\theta}, \tilde{\beta}$, and $\tilde{t}$ back to θ, β, and t.

Denoting $\tilde{\theta}, \tilde{\beta}$, and $\tilde{t}$ as the transformed variables from (5) (the details of all calculations are included in the appendix), we have implemented a Metropolis-Hastings-within-Gibbs sampling method that updates each blocks of $\tilde{\theta}, \tilde{\beta}$, and $\tilde{t}$ at a time. However, for high-dimensional parameters, it is more suitable to use the Langevin-Hastings algorithm as they will have much faster convergence rates (Roberts & Rosenthal, 1998; Roberts & Tweedie, 1996; Møller et al., 1998). For our model and data, $\tilde{t}$ has the highest dimension, hence we

will use Langevin-Hastings algorithm to update $\tilde{t}$. For the remaining blocks we will use the Random-Walk Metropolis Hastings (RWMH) algorithm. The summary of the steps used to run the MCMC algorithm are shown in the diagram above.

2.6 Prediction & Assessment

After running our MCMC algorithm on the BGLGM, we will combine the posterior samples for each parameter to generate posterior distributions for hardwood probabilities at each of the 62 validation locations. We will then emphasize on assessing the predictions from the number of hardwood counts rather than proportions, since the observed proportions are often 0 or 1, while predictions are $0 < p < 1$. Below we describe the various assessments we have considered:

1. *Coverage Probability*: For each of the 62 validation points, we generate posterior samples of hardwood counts from the corresponding posterior probability samples, then examine whether the true hardwood count is inside the (say) 95% posterior interval. The coverage probability will be the proportion of 62 points that are inside their posterior intervals, i.e.:

$$\#(true\ hardwood\ count\ \in\ posterior\ interval\ of\ hardwood\ counts)/62$$

2. *Cross-entropy*: We will use a cross-entropy loss to measure the dissimilarity between observed and estimated hardwood probabilities from both BGLGM and Bayesian Logistic Regression (BayLog), as calculated by:

$$CrossEntropy = -\frac{1}{62}\sum_{j=1}^{62}\Big[y_j \log \hat{p}_j + (n_j - y_j)\log\left(1-\hat{p}_j\right)\Big]$$

where $\hat{p}_j$ is the mean posterior predicted hardwood probability in BayLog and BGLGM.

3. *RMSE (root mean squared error)*: We will also compare RMSE of hardwood probabilities from both BGLGM and BayLog (Bayesian Logistic Regression), calculated as:

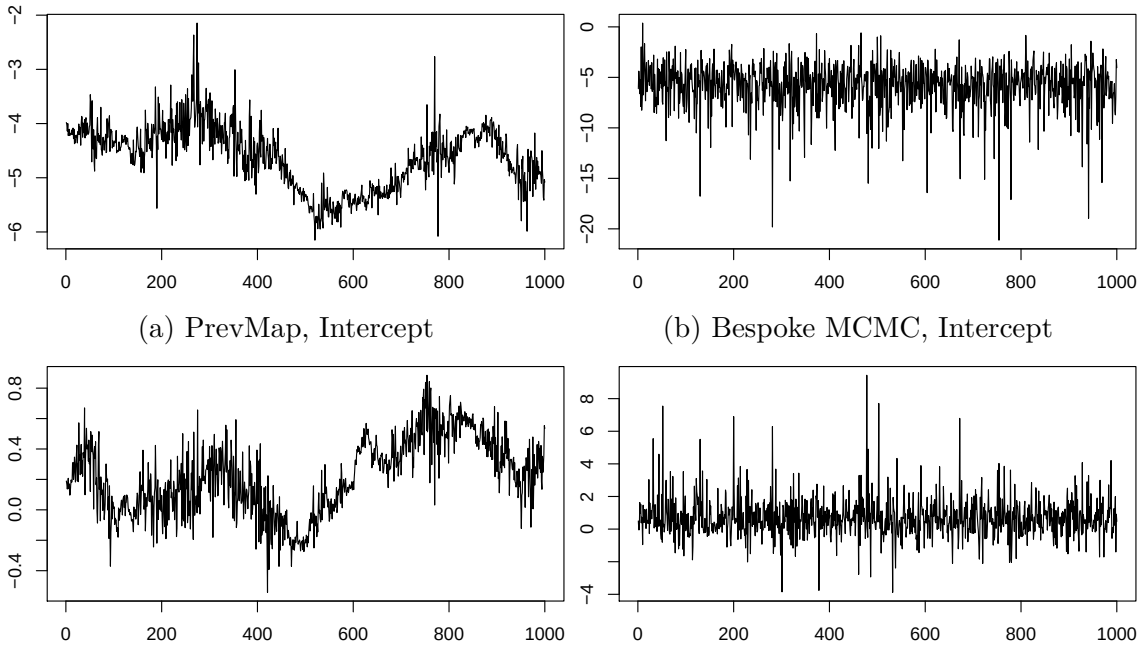
$$RMSE = \sqrt{\frac{1}{62}\sum_{j=1}^{62}(\hat{p}_j - p_j)^2}$$

where p_j is the true proportion of hardwoods in plot i (often 0 or 1) and $\hat{p}_j$ is the mean posterior predicted hardwood probability in BayLog and BGLGM.

4. *Total hardwood count distribution*: We also consider the distribution of the total number of hardwoods in *all* 62 validation sites and examine whether the true total hardwood counts is covered within the 95% posterior interval. Unlike the posterior distributions of hardwood counts in each of the 62 plot, the total count has a reasonably symmetric distribution. In addition, we have compared this to the corresponding distribution generated from BayLog.

3 Results

For the main analysis we have ran the MCMC algorithm for 2,000,000 iterations with 1,000,000 burnin and 100 thinning. Runs consist of fitting 100, 25, and 10 sites as training data, both via random and stratified sampling, with predictions made for the 62 validation data. We have repeated this procedure for five different simulations by randomly choosing five different validation sets of size 62.



(a) PrevMap, Intercept

(b) Bespoke MCMC, Intercept

(c) PrevMap, elevation

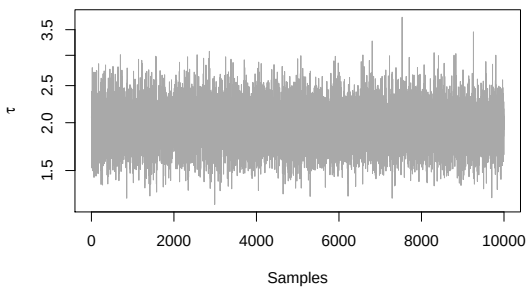
(d) Bespoke MCMC, elevation

Figure 4: Comparing trace plots of β_0 and β_1 from the bespoke MCMC implementation and the PrevMap package.

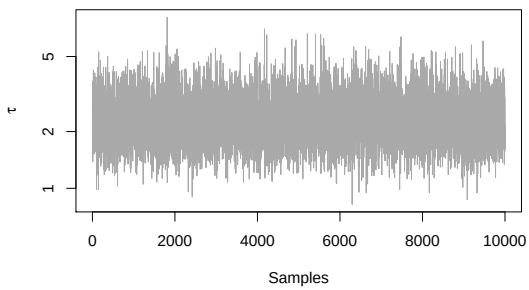
3.1 MCMC Convergence and Mixing

Figure 4 shows, as a simple initial evaluation of the methodology implemented, a comparison of trace plots from the bespoke MCMC (right panels) and the PrevMap package (let panels), using only 10 training sites. PrevMap trace plots exhibit strong autocorrelation and different values of PrevMap's tuning parameters did not result in improvement. There could well be some combination of tuning parameters which would rectify PrevMap, although it is notable that the purpose built MCMC was successful with minimal tuning.

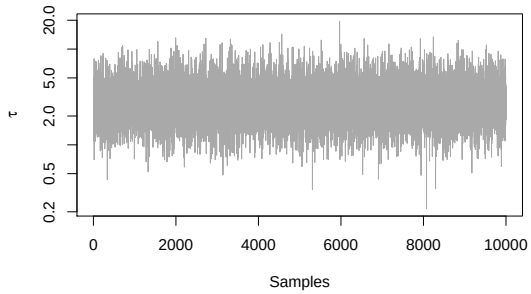
Figures 5a, 5b, and 5c are showing the MCMC trace plots from the bespoke MCMC for the τ parameter with 100, 25, and 10 data fitted to the model respectively. All trace plots show that the MCMC is mixing well and thus, the chains have converged. In addition, more variability in the trace plots is expected with less training data because the sample size is smaller. The remaining trace plots for other parameters as well as other simulations are included in the Supplemental Document.



(a) 100 training sites — τ



(b) 25 training sites — τ



(c) 10 training sites — τ

Figure 5: Trace plots of 10,000 MCMC posterior samples for τ (simulation 1).

For quantitively verifying this variability between different training data size, we have compared the numerical values of posterior mean, 2.5 %, and 97.5 % quantiles of all model parameters in Table 1. While the posterior means remain almost unchanged, the 95% posterior intervals for each model parameter (except ϕ), become wider with less training data fitted to the model, indicating more uncertainty in parameter estimation.

3.2 Parameter posteriors & spatial surfaces

The prior and posterior densities of model parameters from the first simulation are shown in Figures 6 and 7. From these figures we can ascertain that with fewer training data, posterior densities become wider and hence result in more uncertainty of predictions. The posterior distributions of σ suggest small spatial random effects for this dataset, as they have modes concentrated at smaller values. Posterior densities of ϕ are all similar and remain unchanged

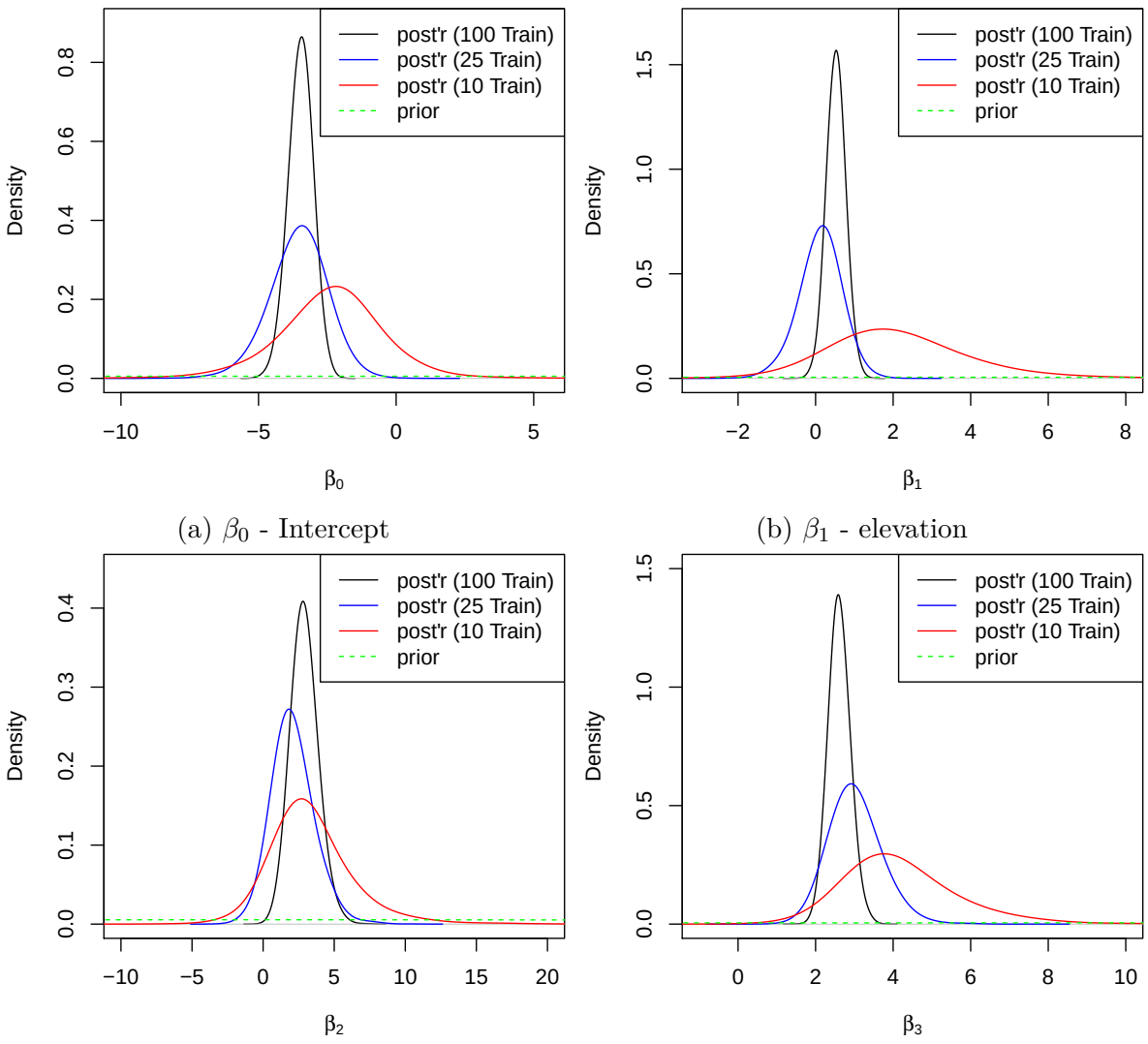
Parameters	# of training	Mean	2.5% quantile	97.5% quantile
Intercept - β_0	100	-3.47	-4.33	-2.65
	25	-3.54	-5.67	-1.66
	10	-2.37	-6.38	1.31
Elevation - β_1	100	0.53	0.07	0.99
	25	0.12	-1.03	1.13
	10	2.16	-0.96	6.38
SkyF<0.3 - β_2	100	2.89	1.19	4.87
	25	2.09	-0.50	5.26
	10	3.38	-1.39	10.42
SkyF>0.3 - β_3	100	2.61	2.10	3.17
	25	3.02	1.86	4.44
	10	4.20	1.85	7.70
Spatial sd - σ	100	0.04	0.02	0.11
	25	0.04	0.02	0.12
	10	0.06	0.02	0.17
Indep. sd - τ	100	1.98	1.52	2.55
	25	2.38	1.36	4.05
	10	3.09	1.04	7.23
Range(km) - ϕ	100	104.94	21.89	255.14
	25	105.42	22.00	252.93
	10	105.06	21.30	255.27

Table 1: Comparison of posterior mean, 2.5 %, and 97.5 % quantiles of model parameters, for different sizes of training data. These results are from only the first of five training samples.

for different training data, as small σ implies a weak spatial signal which provides little information on ϕ.

One surprising feature of Figure 7b, is the posterior density with 10 training data points does not resemble the prior. Even the smallest training dataset considered provides clear evidence that there is more variation in the observed counts than the covariates predict, which is manifest in the results as τ has a posterior distribution concentrated away from zero. There is also evidence that this extra variation is not spatially structured, since σ is clearly much smaller than τ.

The main goal is to predict the composition of trees at unmeasured sites in the forests



(a) β_0 - Intercept

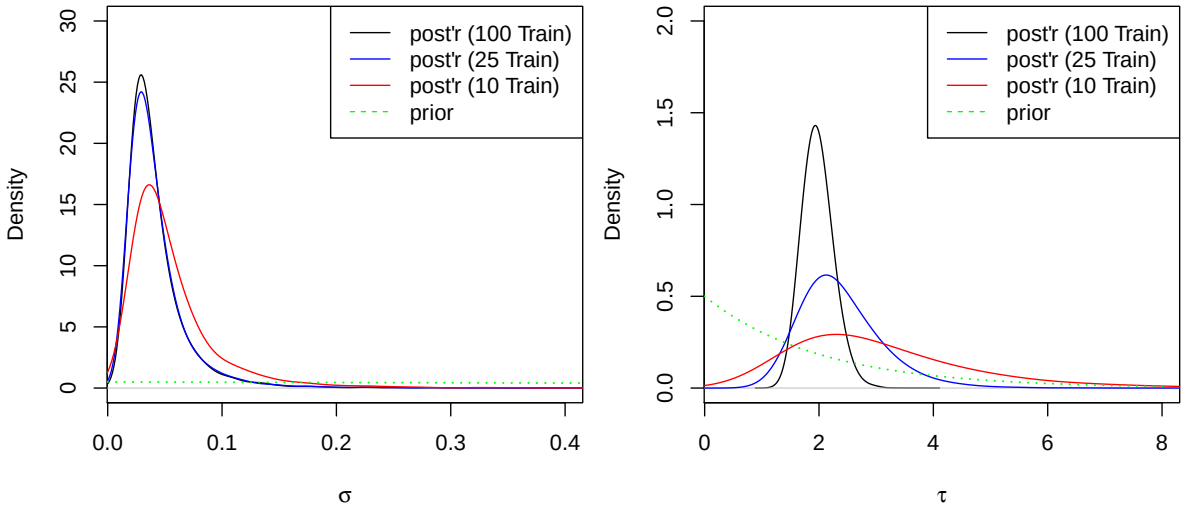
(b) β_1 - elevation

(c) β_2 - SkyForest^{TM} vegetation below 0.3

(d) β_3 - SkyForest^{TM} vegetation above 0.3

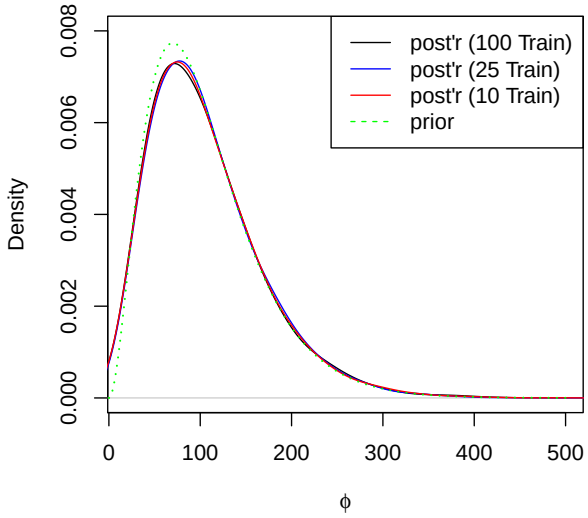
Figure 6: Prior and posterior distributions of parameters from the first simulation.

via simulating posterior samples of $U(g_\ell)$ for new locations $g_\ell : \ell = 1 \dots L$, conditional on MCMC posteriors $\{U(s_i) + Z(s_i) : i = 1 \dots n\}$. Considering a 100×100 grid with $L = 10,000$ cells inside the forests as our new locations, we can simulate $U(g_\ell)$ using the the RandomFields package and make predictions for hardwood probabilities $p(g_\ell)$ for each cell. The RandomFields package has very efficient algorithms for simulating from conditional distributions of spatial processes without using the full variance matrix. Thus assuming we have grid cells $g_1, ..., g_L$, we simulate $[U(g_1), ..., U(g_L)|Y]$ and independent $Z_1, ..., Z_L$ along with the use of other posterior samples to generate $[p(g_1), ..., p(g_L)|Y]$.



(a) σ - spatial sd

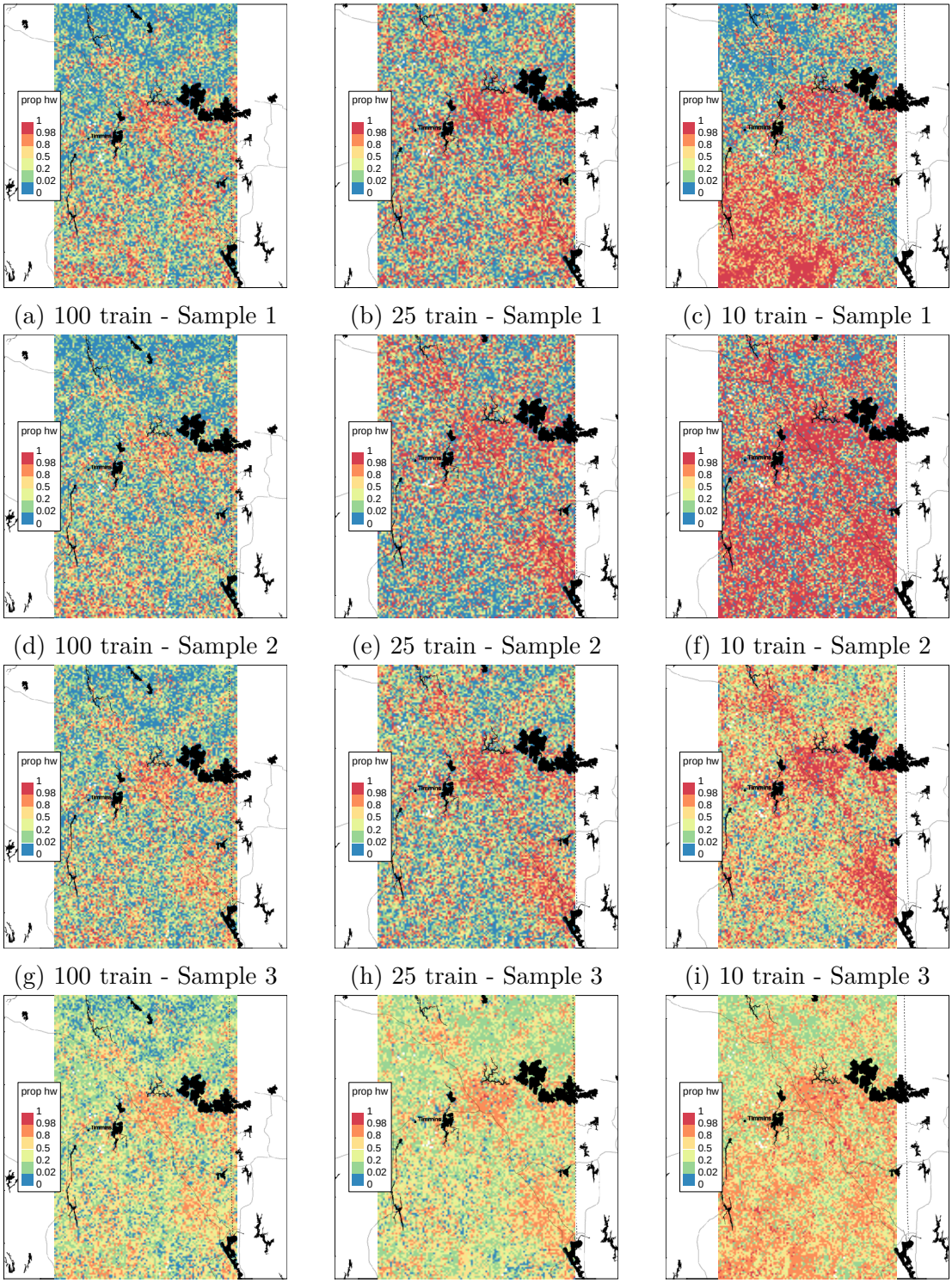
(b) τ - indep sd



(c) ϕ (in km) - range

Figure 7: Priors and posteriors from the first simulation.

Figure 8 shows images of three different posterior samples along with posterior means (in each column) generated from fitting different training data sizes. With fewer training data the posterior rasters appear to become smoother, possibly indicating less precise predictions.



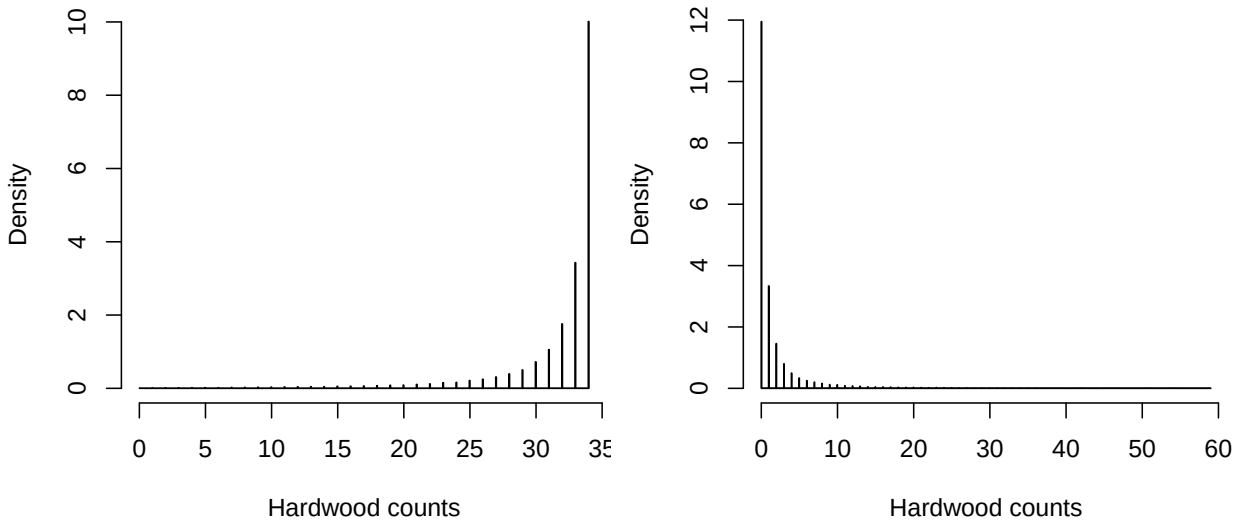
(a) 100 train - Sample 1 (b) 25 train - Sample 1 (c) 10 train - Sample 1

(d) 100 train - Sample 2 (e) 25 train - Sample 2 (f) 10 train - Sample 2

(g) 100 train - Sample 3 (h) 25 train - Sample 3 (i) 10 train - Sample 3

(j) 100 train - Sample mean (k) 25 train - Sample mean (l) 10 train - Sample mean

Figure 8: Three posterior samples of the hardwood proportion surface $p(s)$ along with their posterior means from different training data sizes (Background ©Stamen Design).



(a) All hardwoods, plot 37 (b) No hardwoods, plot 15

Figure 9: Posterior distributions of hardwood counts from two validation plots.

The 62 validation sites with their ground truth number of hardwood trees are used to evaluate predictions by summarizing results over all corresponding sites. The number of hardwood trees in each validation site is predicted and their coverage probabilities calculated from posterior intervals of hardwood counts. Table 2 shows the corresponding coverage probabilities of 95%, 80%, and 50% Posterior Credible Intervals (CI) for different training data size, averaged over five different simulations. Note that many observed proportions are 0 or 1, and the hardwood count posteriors will not be symmetric. To illustrate this, Figure 9 shows the histograms of hardwood count posteriors for two validation plots where in one all are hardwoods and in the other none. We calculate the narrowest credible intervals for each validation plot, and compute their average coverages and widths as shown in Table 2.

	Empirical Coverage of CI			Average CI Width		
#ofTrain	**95 %**	**80 %**	**50 %**	**95 %**	**80 %**	**50 %**
100 Sites	97 %	87 %	59 %	19.98	11.42	4.41
25 Sites	96 %	86 %	55 %	21.91	12.12	4.49
10 Sites	95 %	78 %	55 %	26.64	13.71	4.35

Table 2: Empirical Coverage of Posterior Credible Intervals and their Average Width. All results are averaged over 5 different simulations.

The empirical coverage probabilities tend to exceed their theoretical values, meaning the intervals provided are on the conservative side. Overall, coverage probabilities are all at a desirable value. Table 2 also includes the average width of the posterior intervals, which shows *on average* wider intervals with fewer training data, as expected. The coverage probabilities from the smaller training datasets are, somewhat surprisingly, closer to their theoretical values than those from the large training datasets. Part of this could be simply due to sampling variation, with a particularly fortunate selection of the 10 training sites in each of the 5 subsamples. It is also possible that model becomes a less accurate approximation to the 'real' underlying natural process when the size of the dataset (and hence information about the natural process) grows.

3.3 Comparison of BGLGM with Bayesian Logistic Regression

In this section we will discuss the difference in performances between BGLGM and a simple non-spatial Logistic Regression where non-statistical audience commonly use (i.e. foresters). To consistently compare with the BGLGM model, we fit a Bayesian Logistic Regression model using the BayesLogit package in R. We will compare their performance through Root Mean Square Error (RMSE) and cross entropy calculated via their posterior mean, as well as the coverage of their predictive distributions.

Table 3 reports the RMSEs and cross entropy measures of hardwood probabilities for the 62 validation sites, computed from runs with 100, 25, and 10 training data, for five different subsamples of validation sites. Both measures are calculated using posterior means of predicted proportions. On average RMSE and cross entropy of BGLGM are smaller compared to Bayesian Logistic Regression, indicating more accurate predictions. RMSE and cross entropy increase with less ground truth data fitted to the model, consistent with the results shown in the previous section.

Predictive distributions of the *total* hardwood counts from all 62 validation sites were also computed for both BayLog and BGLGM, using the relevant 10,000 MCMC posterior sam-

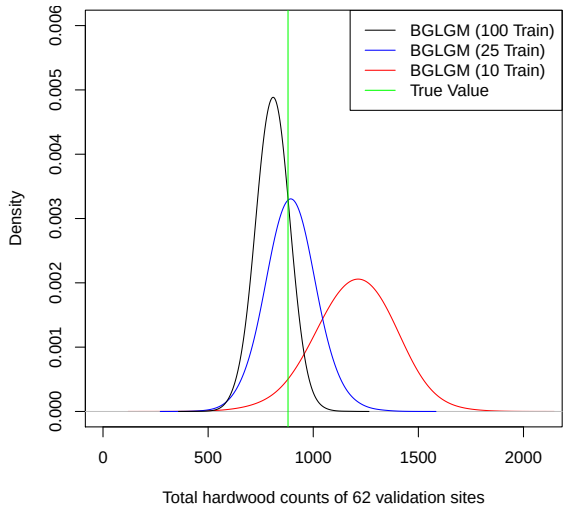
	model	sim 1	sim 2	sim 3	sim 4	sim 5	average
RMSE							
100	BGLGM	0.228	0.207	0.210	0.198	0.157	0.200
	BayLog	0.233	0.206	0.214	0.197	0.158	0.202
25	BGLGM	0.228	0.211	0.234	0.199	0.227	0.220
	BayLog	0.249	0.226	0.246	0.201	0.257	0.236
10	BGLGM	0.291	0.243	0.364	0.208	0.247	0.271
	BayLog	0.361	0.279	0.485	0.237	0.316	0.336
Cross entropy							
100	BGLGM	17.744	13.126	11.351	13.896	12.925	13.809
	BayLog	18.258	13.048	11.536	14.220	13.024	14.017
25	BGLGM	17.588	13.373	12.314	14.056	15.361	14.539
	BayLog	19.294	13.714	12.838	13.880	16.026	15.150
10	BGLGM	20.946	14.747	16.522	14.698	16.447	16.672
	BayLog	32.550	17.565	38.269	14.485	19.029	24.380

Table 3: Root mean square error (RMSE) and cross entropy of predicted hardwood probabilities for the Bayesian Generalized Geostatistical Model (BGLGM) and Bayesian logistic regression model (Bay Log), for training samples of size 100, 25 and 10.

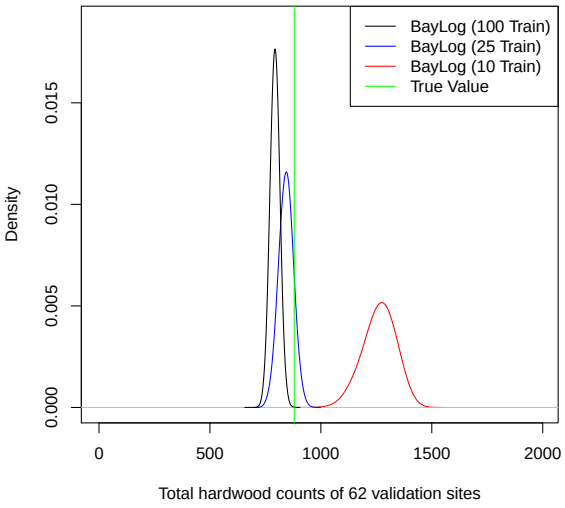
ples. Figures 10a and 10b show these distributions for the first subsample of training plots. Although the distributions from BayLog are narrower compared to those from BGLGM, the BGLGM posterior distributions with all training data sizes capture the true value shown in green within their 95% intervals, while BayLog with 10 and even 100 training data points fails to do so. In addition, we also observe that the posterior distributions become wider with less training data as expected. In conclusion, the BGLGM is a more reliable method compared to the BayLog, in terms of both prediction accuracy and the ability of explaining uncertainties. Note that this process has been repeated for four other simulations with figures shown in the Supplemental Document.

3.3.1 Stratified Sampling of training data

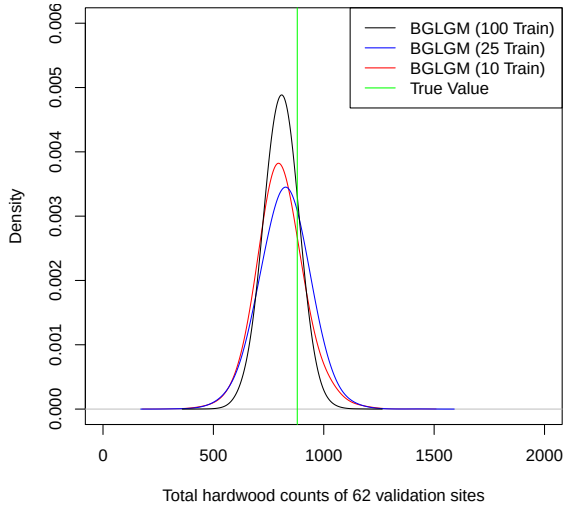
Figures 10a and 10c compares the posterior distributions of the total hardwood trees from both random sampling and stratified sampling on the first of five simulations. Prediction intervals from all five simulations are shown in Figure 10d. The posterior distributions all contain the true value within their 95% posterior interval, however the uncertainty is



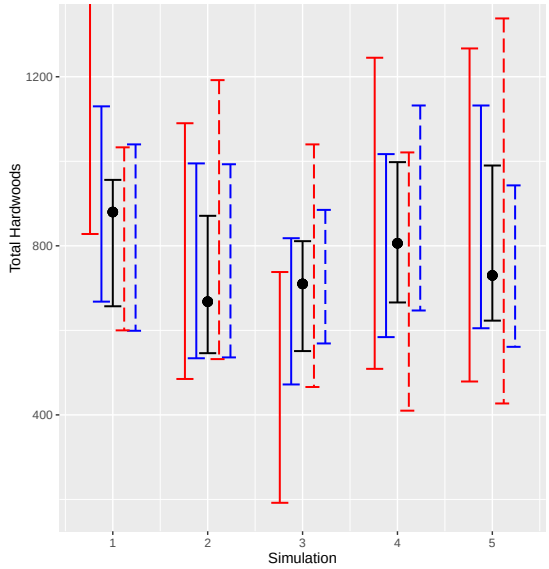
(a) Simulation 1, BGLGM, random



(b) Simulation 1, Logistic Regression



(c) Simulation 1, BGLGM, Stratified



(d) 95% posterior intervals for random (solid) and stratified (dashed) samples.

Figure 10: Posterior distributions of total number of hardwood trees with 10 (red), 25 (blue) and 100 (black) training data points.

generally less under stratified sampling in most cases. In Figure 10d it is notable that the stratified posterior with 10 training data contains the true value near its mode, while with the random posterior it is covered around the tail area. In simulations 2 and 4 results are roughly comparable, while in simulation 3 the stratified posterior with 10 data points captures the true value around its mode. On the other hand, in simulation 5, the stratified posterior with 10 data points becomes more dispersed while 25 is narrower. Overall, the stratified sampling approach shows some potential for improving predictive accuracy when the sample size is small, and a more thorough assessment of possible spatial sampling designs is warranted.

4 Discussion

In this paper, we analyzed the spatial data from the Timiskaming and Abitibi River forests in Ontario, Canada. We have studied the prediction of the proportions of hardwood trees using satellite-derived elevation and vegetation data. A bespoke MCMC algorithm for posterior simulation of a Bayesian Generalized Linear Geostatistical Model (BGLGM) was implemented in order to make spatial predictions for new sites in the forests using the given dataset. We compared BGLGM with a Bayesian Logistic Regression model and although the dataset is imbalanced and contains many zero hardwood counts, the BGLGM provided unbiased estimates with reasonable prediction intervals, while the Bayesian Logistic Regression showed less accurate estimates with underestimated uncertainty associated with the predictions. More importantly, with ground truth data as small as 10 points, BGLGM captured the true value of hardwood tree counts within its 95% posterior intervals, while Bayesian Logistic Regression failed even with 100 training points.

The motivating research question for this work was assessing the feasibility of performing forest inventories with combination of satellite data and a small number of ground truth measurements. The answer to this question has proven to be yes and no. The BGLGM is able to make unbiased predictions with as few as 10 training data points, and posterior

distributions provide a useful quantification of how accurate these predictions are. However, the uncertainty associated with the predictions is considerable, and the prediction intervals associated with 10 training points are arguably too wide to be useful. A stratified sample improved on the simple random sample of ground truth sites for some but not all of the simulations performed. A sample size of 25 ground truth sites was, however, consistently competitive with samples of 100 ground truth sites. Other forest management areas will have different characteristics from the Timiskaming and Abitibi River forests considered here, but these results suggest that a few dozen ground truth sites (more than a handfull, less than a hundred) is the right order of magnitude for a data collection effort.

As future work, one can further extend this model for multiple forests, where forests with similar features are considered to have high correlation indicated within priors and hence facilitate future spatial predictions for similar forests. This will significantly help reduce the redundant collection of data from similar forests.

Data availability statement

The data used in this research are proprietary and owned by First Resource Management Group, see frmginc.com.

Acknowledgements

The authors thank Philip E. J. Green and the First Resource Management Group for introducing us to this problem, providing scientific and technical advice, and making the data available.

Figures 1 and 3 have cartography by The Canada Base Map — Transportation (CBMT) web mapping services of the Earth Sciences Sector (ESS) at Natural Resources Canada (NRCan) licensed as the Open Government Licence — Canada.

Figures 2b and 8 have map tiles by Stamen Design under CC BY 3.0. Data by OpenStreetMap available under the Open Database License.

The second and third authors hold Discovery grants from the Natural Sciences and Engineering Research Council of Canada.

References

Abellan, J. J., Fecht, D., Best, N., Richardson, S., & Briggs, D. J. (2007). Bayesian analysis of the multivariate geographical distribution of the socio-economic environment in England. *Environmetrics*, *18*(7), 745–758.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Brus, D. & De Gruijter, J. (1997). Random sampling or geostatistical modelling? choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, *80*(1-2), 1–44.

Christensen, O. F., Roberts, G. O., & Sköld, M. (2006). Robust Markov Chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *15*(1), 1–17.

Craiu, R. V. & Rosenthal, J. S. (2014). Bayesian computation via Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application*, *1*, 179–201.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Curran, P. J. & Atkinson, P. M. (1998). Geostatistics and remote sensing. *Progress in Physical Geography*, *22*(1), 61–78.

Diggle, P. & Ribeiro, P. (2007). *Model-based Geostatistics*. Springer Series in Statistics. Springer.

Diggle, P. J., Tawn, J., & Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(3), 299–350.

Giorgi, E. & Diggle, P. J. (2017). PrevMap: an R package for prevalence mapping. *Journal of Statis*.

Giorgi, E., Schlüter, D. K., & Diggle, P. J. (2017). Bivariate geostatistical modelling of the relationship between loa loa prevalence and intensity of infection. *Environmetrics*.

Matheron, G. (1962). *Traité de géostatistique appliquée. 1 (1962)*, volume 1. Editions Technip.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, *25*(3), 451–482.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Roberts, G. O. & Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(1), 255–268.

Roberts, G. O. & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, *16*(4), 351–367.

Roberts, G. O. & Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, *2*(4), 341–363.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, *71*(2), 319–392.

Shaby, B. A. & Reich, B. J. (2012). Bayesian spatial extreme value analysis to assess the changing risk of concurrent high temperatures across large portions of European cropland. *Environmetrics*, *23*(8), 638–648.

Stein, M. L. (1999). Interpolation of spatial data: Some theory for Kriging. *Springer-Verlag, New York*.

Wang, J.-F., Stein, A., Gao, B.-B., & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, *2*, 1–14.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, *99*(465), 250–261.

A Appendix

A.1 Reparameterizations Details of section 2.4

As reparamterization and standardization help reduce correlation between variables, they will be play an important role in improving the mixing and convergence of MCMC algorithms. The transformations applied to all the model parameters in (5) follow the recommendations of Christensen et al. (2006) which we will briefly describe here.

Let $\Lambda(t)$ be a diagonal matrix with elements $-\partial^2/\partial t_i^2 \ \log f(y_i|t_i)$ for $i = 1, ..., n$, and denote $\hat{t}_i = \arg\max f(y_i|t_i)$. Assuming a prior $N(\mu, \Omega)$ for β, let $\tilde{\Sigma} = (\Sigma^{-1} + \Lambda(\hat{t}))^{-1}$ and $\tilde{\Omega} = (\Omega^{-1} + X^T(\Sigma^{-1} - \Sigma^{-1}\tilde{\Sigma}\Sigma^{-1})X)^{-1}$. Then by factorizing the posterior distribution in (5) into two parts: $\pi(\beta, \theta, t|y) \propto \pi_1(\theta) f(t, \beta|\theta, y)$, we will be able to simplify the second factor $f(t, \beta|\theta, y)$ as following:

$$\log f(t, \beta|\theta, y) \approx -0.5(t - \hat{t})^T \Lambda(\hat{t})(t - \hat{t}) - 0.5(t - X\beta)^T \Sigma^{-1}(t - X\beta) - 0.5(\beta - \mu)^T \Omega^{-1}(\beta - \mu) \tag{A.1}$$

$$= -0.5(t - \tilde{\Sigma}(\Lambda(\hat{t})\hat{t} + \Sigma^{-1}X\beta))^T \tilde{\Sigma}^{-1}(t - \tilde{\Sigma}(\Lambda(\hat{t})\hat{t} + \Sigma^{-1}X\beta)) \tag{A.2}$$

$$- 0.5(\beta - \tilde{\Omega}(X^T\Sigma^{-1}\tilde{\Sigma}\Lambda(\hat{t})\hat{t} + \Omega^{-1}\mu))^T \tilde{\Omega}^{-1}(\beta - \tilde{\Omega}(X^T\Sigma^{-1}\tilde{\Sigma}\Lambda(\hat{t})\hat{t} + \Omega^{-1}\mu)) \tag{A.3}$$

where the first expression $-0.5(t - \hat{t})^T \Lambda(\hat{t})(t - \hat{t})$ is derived from the Taylor expansion of $\log f(y|t)$ around $\hat{t}$. From equation (A.3), we can simply use the transformations:

$$\tilde{\boldsymbol{t}} = (\tilde{\Sigma}^{1/2})^{-1}(\boldsymbol{t} - \tilde{\Sigma}(\Lambda(\hat{t})\hat{t} + \Sigma^{-1}X\beta)) \tag{A.4}$$

$$\tilde{\boldsymbol{\beta}} = (\tilde{\Omega}^{1/2})^{-1}(\boldsymbol{\beta} - \tilde{\Omega}(X^T\Sigma^{-1}\tilde{\Sigma}\Lambda(\hat{t})\hat{t} + \Omega^{-1}\mu)) \tag{A.5}$$

where $\tilde{t}_1, ..., \tilde{t}_n$ and $\tilde{\beta}_1, ..., \tilde{\beta}_p$ are now approximately uncorrelated with mean zero and variance one. These parameters are also uncorrelated with θ and hence there will be no posterior dependence between $\tilde{t}, \tilde{\beta}$, and θ. However, according to Christensen et al. (2006) and Giorgi

& Diggle (2017), there is posterior dependence within the parameters of $\theta^T = (\theta_1, \theta_2, \theta_3) = (\sigma^2, \phi, \tau)$, and hence a reparameterization is proposed as following:

$$\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3) = (\log \sigma, \log \sigma^2/\phi^{2\kappa}, \log \tau^2)$$

Tables

Parameters	# of training	Mean	2.5% quantile	97.5% quantile
Intercept - β_0	100	-3.47	-4.33	-2.65
	25	-3.54	-5.67	-1.66
	10	-2.37	-6.38	1.31
Elevation - β_1	100	0.53	0.07	0.99
	25	0.12	-1.03	1.13
	10	2.16	-0.96	6.38
SkyF<0.3 - β_2	100	2.89	1.19	4.87
	25	2.09	-0.50	5.26
	10	3.38	-1.39	10.42
SkyF>0.3 - β_3	100	2.61	2.10	3.17
	25	3.02	1.86	4.44
	10	4.20	1.85	7.70
Spatial sd - σ	100	0.04	0.02	0.11
	25	0.04	0.02	0.12
	10	0.06	0.02	0.17
Indep. sd - τ	100	1.98	1.52	2.55
	25	2.38	1.36	4.05
	10	3.09	1.04	7.23
Range(km) - ϕ	100	104.94	21.89	255.14
	25	105.42	22.00	252.93
	10	105.06	21.30	255.27

Table 4: Table 1 Comparison of posterior mean, 2.5 %, and 97.5 % quantiles of model parameters, for different sizes of training data. These results are from only the first of five training samples.

	Empirical Coverage of CI			Average CI Width		
#ofTrain	**95 %**	**80 %**	**50 %**	**95 %**	**80 %**	**50 %**
100 Sites	97 %	87 %	59 %	19.98	11.42	4.41
25 Sites	96 %	86 %	55 %	21.91	12.12	4.49
10 Sites	95 %	78 %	55 %	26.64	13.71	4.35

Table 5: Table 2: Empirical Coverage of Posterior Credible Intervals and their Average Width. All results are averaged over 5 different simulations.

	model	sim 1	sim 2	sim 3	sim 4	sim 5	average
RMSE							
100	BGLGM	0.228	0.207	0.210	0.198	0.157	0.200
	BayLog	0.233	0.206	0.214	0.197	0.158	0.202
25	BGLGM	0.228	0.211	0.234	0.199	0.227	0.220
	BayLog	0.249	0.226	0.246	0.201	0.257	0.236
10	BGLGM	0.291	0.243	0.364	0.208	0.247	0.271
	BayLog	0.361	0.279	0.485	0.237	0.316	0.336
Cross entropy							
100	BGLGM	17.744	13.126	11.351	13.896	12.925	13.809
	BayLog	18.258	13.048	11.536	14.220	13.024	14.017
25	BGLGM	17.588	13.373	12.314	14.056	15.361	14.539
	BayLog	19.294	13.714	12.838	13.880	16.026	15.150
10	BGLGM	20.946	14.747	16.522	14.698	16.447	16.672
	BayLog	32.550	17.565	38.269	14.485	19.029	24.380

Table 6: Table 3: Root mean square error (RMSE) and cross entropy of predicted hardwood probabilities for the Bayesian Generalized Geostatistical Model (BGLGM) and Bayesian logistic regression model (Bay Log), for training samples of size 100, 25 and 10.