# Bayesian CART learning for complex stochastic system

## 1 Introduction

Our work is motivated by heavy metal escapement associated with the largest zinc mine in the world, Red Dog Mine, which has operated year-round to produce lead(Pb) and zinc(Zn) concentrates in powder form at the mine site. There is a haul road from mine site to the storage facilities in the coast of northwest Alaska for truck transportation. As trucks travel the haul road, dust with heavy metals and other contaminants is deposited on the plant communities near the haul road. These contaminates are easily absorbed into the tissue of lichen species, especially, so they are good bio-monitors of contaminant concentrations. We consider six contaminants, aluminum(Al), calcium(Ca), iron(Fe), cadmium(Cd), zinc(Zn) and lead(Pd). Lichen tissue was analyzed for contaminant concentration from samples at various distances from both north and south side of the haul road. In addition to contaminant concentration in lichen tissue, for each sample, we measured species richness, lichen species cover and the percent blackening of the midrib of the lichen Hylocomium splendens. Our goal is to build Bayesian hierarchical models to explore the relationship between contaminant concentration and distance from the haul road, and then the relationship between species richness, lichen cover and percent blacking with contaminant concentration as modeled from distance to road.

For each contaminant, we will assume that it follows a multivariate normal distribution conditional on a spatially-varying mean $s_c$. The process of interest is $s_c$, forming the mean of the model at this level of the hierarchy. Then for each latent process vector $s_{c,j}$, $j = 1, \cdots, 6$, it is composed of the spatially structured fixed effect vector $\mu_{c,j}$ and a spatially-autocorrelated random effect $z_{c,j}$, where $\mu_{c,j}$ depends on distance to the haul road $d_{c,j}$ and side of the road(north and south). The details will be described in Section 2.

With $s_c$, we continue to model the ecological responses 1) species richness, 2) lichen cover, and 3) midrib blackening depending on these 6 latent process vectors. For Red Dog Mine is operated for Zn, we believe the variety of $s'_{c,j}s$ differs among contaminants, for example, $s_{c,Zn}$ is supposed to have a larger variety. Generally it is pretty vague to model this difference in a parametric model because little is known about the weight associated with contaminant contributing to the ecological responses. However it can be addressed by a nonparametric Bayesian classification and regression tree (CART), explores by Chipman et al. (1998), Denison et al. (1998), which performs as an efficient stochastic search algorithm and provides a variety of tree structures rather than a simple tree. The non-linear formulation of Bayesian CART is flexible and effective to specify the relationship between response variable (ecological responses) and predictors ($s'_{c,j}s$). The main idea is to build a binary decision tree according to some splitting rules which recursively partition the predictor space into subsets to increase the homogeneity of response variable within each subset. The partition is repeated until no split improves the homogeneity across the tree or the number of observations in the node is below a certain number. Then the node becomes a terminal node. By traversing from tree root to terminal nodes, observations are assigned to terminal nodes. Then within each node, the conditional distribution of response variable given predictors is determined to make inference and forecast, for example, see Gramacy and Lee (2012).

In general, Bayesian CART assumes observations between different terminal nodes are independent. Only the within node observations are correlated. This assumption is very practical for parameter estimation and asymptotic property assessment. However, in our case, the data is a sample of regionalized variables, which presents spatial dependent structure based on the geographic location. Therefore the independence assumption between terminal nodes are unreasonable. The loss of spatial information may lead to biased splitting rules. Here in our paper, instead of independence, we take into account the spatial pattern and propose the spatially correlated Bayesian CART algorithm. The overall spatial pattern only depends on the geographical location and only change up to a scale parameter during the computation which can be achieved after using permutation matrices. Inside of each terminal node, in addition to the overall spatial pattern, we use the multiple linear regression to model the relationship between appropriate transformation of ecological responses and $s_c$.

The rest of the paper is organized as follows. In Section 2, we give the detail of Bayesian hierarchical modeling for the relationship between contaminant concentration and the distance from haul road. The structure of Bayesian CART including model setting, tree operation and sampling schemes for parameters is given in Section 3. A summary of analysis for the Red Dog Mine data is given in Section 4. The paper is closed with a discussion in Section 5.

# 2 Contaminant models associated with distance to the road

## 2.1 Model setting

For each contaminant, let $y_{c,j}(u_i)$ be a random variable for the value of the contaminant at spatial location $u_i$ (a vector containing latitude and longitude), for the jth contaminant, where $j = 1, \cdots, 6$ are indexes for the set {Al, Ca, Fe, Zn, Pb, Cd}. Let $y_{c,j}$ be the vector of all of the random variables $\{y_{c,j}(u_i) : i = 1, \cdots, n_{c,j}\}$. We will use the generic notation $[y \mid s]$ where the bracket [] indicate a distribution, and the bar | denotes conditional dependence. A log (base 10) transformation was used for contaminant values, and conditional on a spatially-varying mean s, we will assume that each contaminant has a multivariate normal distribution,

$$[y_{c,j} \mid s_{c,j}, \lambda_{0,c,j}^{-1}] \sim MVN(s_{c,j}, \lambda_{0,c,j}^{-1} I),$$

where I is the identity matrix. The variance $\lambda_{0,c,j}^{-1}$ is spatially unstructured error representing vary fine scale variation and measurement errors. The process of interest is the mean of the model $s_{c,j}$.

The process $s_{c,j}$ for each contaminant is composed of two parts,

$$s_{c,j}(u_i) = \mu_{c,j}(u_i) + z_{c,j}(u_i),$$

where $s_{c,j}(u_i)$ is the ith component of the vector $s_{c,j}$, $\mu_{c,j}(u_i)$ is a spatially structured fixed effect through two covariates, distance from haul road and side of road (north and south), and $z_{c,j}(u_i)$ is a spatially-autocorrelated random effect.

For $\mu_{c,j}$, we assume 1) the effect of distance from road is modeled differently for the north side of the road and the south side of the road because of prevailing winds, 2) as distance to the road becomes larger, its effect on the contaminant concentration becomes weaker. For the first assumption, we fit separate models to the north and south side of the haul road. For the second assumption, we use a modified Bessel function of second kind with order 0 to model the trend, which is denoted as $k_0(\cdot)$. Therefore the spatially-structured fixed effects is modeled as

$$\mu_{c,j}(u_i) = k_0(d(u_i)/\lambda_{3,c,j})\alpha_{N,j}I_N(u_i) + k_0(d(u_i)/\lambda_{4,c,j})\alpha_{S,j}I_S(u_i),$$

where $I_N(u_i)$ is an indicator function that location $u_i$ is north of the haul road, $I_S(u_i) = 1 - I_N(u_i)$ is an indicator function that $u_i$ is south of the haul road, $d(u_i)$ is a Euclidean distance function from $u_i$ to the nearest point on the haul road, and $\alpha_{N,j}$ and $\alpha_{S,j}$ are parameters. For the spatial random effect vector $z_{c,j}$, we assume a normal distribution with mean 0 and exponential covariance function form which can be expressed as

$$cov(z_{c,j}(u_i), z_{c,j}(u_{i'})) = \lambda_{1,c,j}^{-1} exp(-||u_i - u_{i'}||)/\lambda_{2,c,j}), \tag{1}$$

where $||u_i - u_{i'}||$ is the Euclidean distance between points $u_i$ and $u_{i'}$, and (1) is the $(i,i')th$ element of the covariance matrix $\Sigma_{c,j} = \lambda_{1,c,j}^{-1}R_{c,j}$. $\lambda_{1,c,j}^{-1}$ is the partial sill parameter and $\lambda_{2,c,j}$, $\lambda_{3,c,j}$ and $\lambda_{4,c,j}$ are the range parameters. Let $\theta_{c,j} = (\alpha_{N,j}, \alpha_{S,j}, \lambda_{0,c,j}, \lambda_{1,c,j}, \lambda_{2,c,j}, \lambda_{3,c,j}, \lambda_{4,c,j})$, then process $y_{c,j}$ and $s_{c,j}$ can be defined. With a final specification of the priors on $\theta_{c,j}$, we can obtain the joint distribution $[y_{c,j}, s_{c,j}, \theta_{c,j}]$, and finally to the posterior distributions of interest $[s_{c,j}, \theta_{c,j} \mid y_{c,j}]$. To complete the full Bayesian analysis, we specify non-informative hyper-priors for parameters as follows,

$$\lambda_{0,c,j}, \lambda_{1,c,j} \sim Gamma(1,200), \quad \lambda_{2,c,j} \sim Unif(0.01,100),$$

$$\lambda_{3,c,j} \sim Unif(1000,10000), \quad \alpha_{N,j} \sim N(0, [k_0^T(d/\lambda_{3,c,j})k_0(d/\lambda_{3,c,j})]^{-1}).$$

## 2.2 Parameters sampling scheme

Set $\gamma_0 = 1$ and $\gamma_1 = 200$, the posterior distribution for model in last subsection can be written as

$$\begin{aligned} f(\theta_{c,j}, s_{c,j} \mid y_{c,j}) \propto & f(y_{c,j} \mid s_{c,j}, \lambda_{0,c,j}) f(s_{c,j} \mid \alpha_{N,j}, \lambda_{1,c,j}, \lambda_{2,c,j}, \lambda_{3,c,j}) \\ & f(\alpha_{N,j} \mid \lambda_{3,c,j}) f(\lambda_{0,c,j}) f(\lambda_{1,c,j}) f(\lambda_{2,c,j}) f(\lambda_{3,c,j}) \\ \propto & \lambda_{0,c,j}^{\frac{n}{2}} e^{\frac{\lambda_{0,c,j}}{2}(y_{c,j}-s_{c,j})^T(y_{c,j}-s_{c,j})} \lambda_{1,c,j}^{\frac{n}{2}} |R_{c,j}|^{-\frac{1}{2}} \\ & e^{-\frac{\lambda_{1,c,j}}{2}(s_{c,j}-k_0(d/\lambda_{3,c,j})\alpha_{N,j})^T R_{c,j}^{-1}(s_{c,j}-k_0(d/\lambda_{3,c,j})\alpha_{N,j})} \\ & e^{-\frac{1}{2}\alpha_{N,j}^T k_0^T(d/\lambda_{3,c,j}) k_0(d/\lambda_{3,c,j})\alpha_{N,j}} \lambda_{0,c,j}^{\gamma_0 - 1} e^{\frac{\lambda_{0,c,j}}{\gamma_1}} \lambda_{1,c,j}^{\gamma_0 - 1} e^{-\frac{\lambda_{1,c,j}}{\gamma_1}} \\ & \frac{\lambda_{2,c,j} - 0.01}{100 - 0.01} \frac{\lambda_{3,c,j} - 1000}{10000 - 1000}. \end{aligned}$$

From the posterior distribution, the sampling schemes for all the parameters and the latent process can be obtained as follows,

$$f(\lambda_{1,c,j}|\cdot) \propto \lambda_{1,c,j}^{\frac{n}{2}+\gamma_0-1} e^{-\frac{\lambda_{1,c,j}}{\frac{1}{\frac{1}{2}(s_{c,j}-k_0(d/\lambda_{3,c,j})\alpha_{N,j})^T R_{c,j}^{-1}(s_{c,j}-k_0(d/\lambda_{3,c,j})\alpha_{N,j})+\frac{1}{\gamma_1}}}}$$

$$\lambda_{1,c,j} \sim Gamma(\frac{n}{2}+\gamma_0, \frac{1}{\frac{1}{2}(s_{c,j}-k_0(d/\lambda_{3,c,j})\alpha_{N,j})^T R_{c,j}^{-1}(s_{c,j}-k_0(d/\lambda_{3,c,j})\alpha_{N,j})+\frac{1}{\gamma_1}}).$$

$$f(\lambda_{0,c,j}|\cdot) \propto \lambda_{0,c,j}^{\frac{n}{2}+\gamma_0-1} e^{-\frac{\lambda_{0,c,j}}{\frac{1}{\frac{1}{2}(y_{c,j}-s_{c,j})^T(y_{c,j}-s_{c,j})+\frac{1}{\gamma_1}}}}$$

$$\lambda_{0,c,j} \sim Gamma(\frac{n}{2}+\gamma_0, \frac{1}{\frac{1}{2}(y_{c,j}-s_{c,j})^T(y_{c,j}-s_{c,j})+\frac{1}{\gamma_1}}).$$

$$\begin{aligned} f(\alpha_{N,j} \mid \cdot) \propto & e^{-\frac{\lambda_{1,c,j}}{2}[\alpha_{N,j}^T k_0^T(d/\lambda_{3,c,j})k_0(d/\lambda_{3,c,j})\alpha_{N,j} - 2s_{c,j}^T R_{c,j}^{-1} k_0(d/\lambda_{3,c,j})\alpha_{N,j}] - \frac{1}{2}\alpha_{N,j}^T k_0^T(d/\lambda_{3,c,j})k_0(d/\lambda_{3,c,j})\alpha_{N,j}} \\ \propto & e^{-\frac{1}{2}[\alpha_{N,j}^T(\lambda_1+1)k_0^T(d/\lambda_{3,c,j})k_0(d/\lambda_{3,c,j})\alpha_{N,j} - 2*\lambda_{1,c,j}s_{c,j}^T R_{c,j}^{-1} k_0(d/\lambda_{3,c,j})\alpha_{N,j}]} \end{aligned}$$

Let $Q_{\alpha_{N,j}} = (\lambda_{1,c,j}+1)k_0^T(d/\lambda_{3,c,j})k_0(d/\lambda_{3,c,j})$,

$$\alpha_{N,j} \sim N(Q_{\alpha_{N,j}}^{-1}\lambda_{1,c,j}k_0^T(d/\lambda_{3,c,j})R_{c,j}^{-1}s_{c,j}, Q_{\alpha_{N,j}}^{-1}).$$

$$\begin{aligned} f(s_{c,j} \mid \cdot) \propto & e^{-\frac{\lambda_{0,c,j}}{2}[s_{c,j}^T s_{c,j} - 2y_{c,j}^T s_{c,j}]} e^{-\frac{\lambda_{1,c,j}}{2}[s_{c,j}^T R_{c,j}^{-1} s_{c,j} - 2\alpha_{N,j}^T k_0^T(d/\lambda_{3,c,j})R_{c,j}^{-1}s_{c,j}]} \\ \propto & e^{-\frac{1}{2}[s_{c,j}^T(\lambda_{0,c,j}I+\lambda_{1,c,j}R_{c,j}^{-1})s_{c,j} - 2(\lambda_{0,c,j}y_{c,j}^T+\alpha_{N,j}^T k_0^T(d/\lambda_{3,c,j})R_{c,j}^{-1})s_{c,j}]} \end{aligned}$$

$$s_{c,j} \sim N((\lambda_{0,c,j}I + \lambda_{1,c,j}R_{c,j}^{-1})^{-1}(\lambda_{0,c,j}y_{c,j} + R_{c,j}^{-1}k_0(d/\lambda_{3,c,j})\alpha_{N,j}), (\lambda_{0,c,j}I + \lambda_{1,c,j}R_{c,j}^{-1})^{-1}).$$

For range parameter $\lambda_{2,c,j}$, we use the Metropolis-Hastings algorithm for update. The acceptance rate will be $min(1, p_{2,c,j})$, where

$$p_{2,c,j} = \frac{\frac{1}{|R_{c,j}^{pro}|^{1/2}} e^{-\frac{\lambda_{1,c,j}}{2}(s_{c,j} - k_0(d/\lambda_{3,c,j})\alpha_{N,j})^T (R_{c,j}^{pro})^{-1}(s_{c,j} - k_0(d/\lambda_{3,c,j})\alpha_{N,j})} (\lambda_{2,c,j}^{cur} - 0.01)}{\frac{1}{|R_{c,j}^{cur}|^{1/2}} e^{-\frac{\lambda_{1,c,j}}{2}(s_{c,j} - k_0(d/\lambda_{3,c,j})\alpha_{N,j})^T (R_{c,j}^{cur})^{-1}(s_{c,j} - k_0(d/\lambda_{3,c,j})\alpha_{N,j})} (\lambda_{2,c,j}^{pro} - 0.01)}.$$

For range parameters $\lambda_{3,c,j}$, we also use the Metropolis-Hastings algorithm for update. The acceptance rate will be $min(1, p_{3,c,j})$, where

$$p_{3,c,j} = \frac{p1}{p2},$$

with

$$\begin{aligned} p1 =& e^{-\frac{\lambda_{1,c,j}}{2}(s_{c,j} - k_0(d/\lambda_{3,c,j}^{pro})\alpha_{N,j})^T R_{c,j}^{-1}(s_{c,j} - k_0(d/\lambda_{3,c,j}^{pro})\alpha_{N,j})} \\ & e^{-\frac{1}{2}\alpha_{N,j}^T k_0^T(d/\lambda_{3,c,j}^{pro})k_0(d/\lambda_{3,c,j}^{pro})\alpha_{N,j}} (\lambda_{3,c,j}^{cur} - 1000), \end{aligned}$$

and

$$\begin{aligned} p2 =& e^{-\frac{\lambda_{1,c,j}}{2}(s_{c,j} - k_0(d/\lambda_{3,c,j}^{cur})\alpha_{N,j})^T R_{c,j}^{-1}(s_{c,j} - k_0(d/\lambda_{3,c,j}^{cur})\alpha_{N,j})} \\ & e^{-\frac{1}{2}\alpha_{N,j}^T k_0^T(d/\lambda_{3,c,j}^{cur})k_0(d/\lambda_{3,c,j}^{cur})\alpha_{N,j}} (\lambda_{3,c,j}^{pro} - 1000). \end{aligned}$$

# 3 Bayesian CART structure

## 3.1 Model specification

Classically, geo-statistical data $y_i : i = 1, \cdots, n$ correspond to explanatory variables $x_{ij} : j = 1, \cdots, p$. For a tree with k terminal nodes, a vector Y is composed of all the observations in order by a picked traversal across the tree, X is the corresponding explanatory variable matrix, $\psi$ is the vector of corresponding spatial random effect, then the statistical model here can be expressed as,

$$Y = f(X) + \psi + \varepsilon,$$

where $f(X)$ is a function of X needed to be specified and $\varepsilon$ is Gaussian random error vector with $\varepsilon \sim N(0, \lambda_2^{-1}I)$. The specific example we will consider in detail is a linear regression within each node, $f(x_i) = x_i\beta_i$, with $\beta_i = \{\beta_{ij}\}, j = 1, \cdots, p$ for terminal node i.

For spatial random effect $\psi$, we assume it is a Gaussian random vector only depending on the geographic location and the covariance parameters. Therefore we only define the covariance or precision matrix structure for $\psi$ at the beginning. Then for each tree, a permutation matrix R is determined to guarantee $R\psi$ matching with the traversal. Under this setting, for each tree, the model can be written as,

$$Y = f(X) + R\psi + \varepsilon,$$

where $\psi$ keeps the same structure across the whole computation, while structure of other variables vary from tree to tree. There are two scenarios we particularly want to explore, one is when observations are embedded in a regular lattice and the other one is when observations are from irregular shape area.

When observations are embedded in a $n = r \times c$ regular lattice, for the $(k,l)th$ array cell, we place the following Markov Random Field Gaussian prior for $\psi_{k,l}$,

$$E\{\psi_{k,l} \mid \psi_{k',l'}, (k',l') = (k,l)\} = (\psi_{k-1,l} + \psi_{k+1,l} + \psi_{k,l-1} + \psi_{k,l+1})/4$$

$$var\{\psi_{k,l} \mid \psi_{k',l'}, (k',l') \neq (k,l)\} = 1/(4\lambda_1), \quad \lambda_1 > 0.$$

Define $W_k$ as a $k \times k$ matrix, the only non-zero off-diagonal elements in $W_k$ are $W_{i,i\pm 1} = -1$ with $W_{i,i}$ ensuring zero row and column sums. Then the precision matrix of $\psi$ can be written as the sum of Kronecker products,

$$Q = \lambda_1 (I_c \otimes W_r + W_c \otimes I_r)/2,$$

and $\psi \sim N(0, Q^{-1})$. This structure is the well-known intrinsic autoregressive model. Furthermore let $P_k$ denote the $k \times k$ matrix corresponding to the discrete cosine transformation with entries

$$p_{1,j} = k^{-1/2}, \quad p_{i,j} = (2/k)^{1/2} cos\{\pi(i-1)(j-1/2)/k\}, \quad i = 2, \ldots, k, \quad j = 1, \ldots, k.$$

Suppose that $D_k$ is a diagonal matrix $i$th diagonal entry

$$d_{k,i} = 2[1 - cos\{\pi(i-1)/k\}].$$

It then follows that $P = P_c \otimes P_r$ diagonalizes $Q$,

$$PQP^T = \lambda_1 (I_c \otimes D_r + D_c \otimes I_r)/2 = \Lambda,$$

where $\Lambda$ is $n \times n$ diagonal matrices. The matrix $P$ and $P^T$ corresponding to the two-dimensional discrete cosine transformation and inverse discrete cosine transformation respectively.

When observations are obtained from an irregularly shaped area, for the $ith$ and $jth$ element in $\psi$, we define the covariance function through exponential variogram,

$$cov(\psi_i, \psi_j) = \sigma^2 exp(-d_{ij}/\lambda_0) = exp(-d_{ij}/\lambda_0)/\lambda_1,$$

where $d_{ij}$ is the Euclidean distance between points i and j, $\sigma^2$ is the partial sill parameter and $\lambda_0$ is the range parameter. Let $V = \{v_{ij}\}$ with $v_{ij} = exp(-d_{ij}/\lambda_0)$, then the Gaussian prior for $\psi$ can be written as $\psi \sim N(0, \lambda_1^{-1} V)$.

For a full Bayesian CART analysis, we complete the model specification by assuming the non-informative hyperpriors,

$$\lambda_2, \lambda_1 \sim Gamma(1,200),$$

$$\lambda_0 \sim Uniform(0.01,100).$$

A poisson distribution with parameter $\lambda$ has the probability $P(k) = \lambda^k/[(e^{\lambda}-1)k!]$ under the positive restriction. For tree with k terminal nodes, $v_k$ is the probability to change the variable, $\rho_k$ is the probability to change the value, $b_k$ is the birth probability and $d_k$ is the death probability. The details are as following,

$$v_k = \rho_k, \quad b_k = cmin(1, P(k+1)/P(k)), \quad d_{k+1} = cmin(1, P(k)/P(k+1)),$$

$$b_k + d_k \leq 0.75, \quad b_k P(k) = d_{k+1} P(k+1).$$

The main part of reversible jump MCMC acceptance ratio can be written as $(likelihood \quad ratio) \times (prior \quad ratio) \times (proposal \quad ratio) \times |Jacobian|$, in the Bayesian CART model, $|Jacobian| = 1$.

## 3.2 Change Variable

To get a high acceptance ratio, we need to avoid disturbing the tree structure. Therefore for the changing variable move, after picking up a variable, the value of variable should be from

$$l = (max(min-va-val,\ max-lparent-va-val,\ max-lchild-va-val),$$
$$min(max-va-val,\ min-rparent-va-val,\ min-rchild-va-val)).$$

In this case, $proposal\ ratio = 1$, and

$$the\ prior\ ratio = \frac{P(new-va)|l_{new-val}|}{P(old-va)|l_{old-val}|} = \frac{|l_{new-val}|}{|l_{old-val}|},$$

finally need to compute the likelihood ratio.

## 3.3 Only change value

To avoid disturbing the tree structure, the value should again be selected from the above l. In this case, $proposal\ ratio = prior\ ratio = 1$. We only need to calculate the likelihood ratio.

## 3.4 Birth

For birth step, firstly we need to choose a terminal node and then change it into a split node. In this case, with $k_{die}$ representing the umber of possible location for death in current model,

$$proposal\ ratio = \frac{d_{k+1}\frac{1}{k_{die}+1}}{b_k \frac{1}{k} P(u = b-va) P(u-val = b-va-val)},$$

$$prior\ ratio = \frac{P(k+1)}{P(k)} P(b-va)P(b-va-val),$$

therefore

$$(proposal\ ratio) \times (prior\ ratio) = \frac{k}{k_{die}+1},$$

finally we need to calculate the likelihood ratio.

## 3.5 Death

For death step, firstly we need to choose a valid death node from all of possible locations for death in current model. In this case, with $k_{die}$ being the number of possible locations for death in current model,

$$proposal\ ratio = \frac{b_{k-1}\frac{1}{k-1}P(u=b-va)P(u=b-va-val)}{d_k\frac{1}{k_{die}}},$$

$$prior\ ratio = \frac{P(k-1)}{P(k)P(b-va)P(b-va-val)},$$

therefore

$$(proposal\ ratio) \times (prior\ ratio) = \frac{k_{die}}{k-1},$$

finally we need to calculate the likelihood ratio.

## 3.6 Sampling scheme for parameters

Suppose currently there are $(k-1)$ splittable nodes and k terminal nodes, then the model can be written as

$$Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix}, \quad X = \begin{pmatrix} X_1 & & & \\ & X_2 & & \\ & & \ddots & \\ & & & X_n \end{pmatrix}, \quad X_i = \begin{pmatrix} x_{i11} & x_{i12} & \dots & x_{i1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{in_i1} & x_{in_i2} & \dots & x_{in_ip} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_{11} \\ \vdots \\ \beta_{1p} \\ \vdots \\ \beta_{k1} \\ \vdots \\ \beta_{kp} \end{pmatrix}, \quad \psi = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_{n_1+\ldots+n_k} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix}.$$

For species richness, $y_{ij} = log(obs_{ij})$, for lichen cover, $y_{ij} = log(\frac{obs_{ij}}{1-obs_{ij}})$ and for midrib blackening, $y_{ij} = log(\frac{obs_{ij}}{1-obs_{ij}})$. With above notation and the permutation matrix R, the model for current tree could be written as,

$$Y = X\beta + R\psi + \varepsilon.$$

Assume $\beta \sim N(0,(X^TX)^{-1})$, $\psi \sim N(0,\lambda_1^{-1}V)$, where the $(i,j)$ element of V is $V_{ij} = e^{-d_{ij}/\lambda_0}$ with $d_{ij}$ the distance of (location i, location j) and $\varepsilon \sim N(0,\lambda_2^{-1}I)$. For the hyper-parameters, we assume $\lambda_0 \sim unif(0.01,100)$, $\lambda_1 \sim Gamma(\alpha,\gamma)$ and $\lambda_2 \sim Gamma(\alpha,\gamma)$. Usually with non-informative prior, we set $\alpha = 1$ and $\gamma = 200$. Then the posterior distribution of the parameters can be written as

$$\begin{aligned} f(\beta,\phi,\lambda_0,\lambda_1,\lambda_2 \mid Y) \propto & \lambda_2^{n/2} e^{-\frac{\lambda_2}{2}(Y-X\beta-R\psi)^T(Y-X\beta-R\psi)} \lambda_1^{n/2} \frac{1}{|V|^{1/2}} e^{-\frac{\lambda_1}{2}\psi^T V^{-1}\psi} \\ & |X^TX|^{1/2} e^{-\frac{1}{2}\beta^T(X^TX)\beta} \lambda_2^{\alpha-1} e^{-\lambda_2/\gamma} \lambda_1^{\alpha-1} e^{-\lambda_1/\gamma} \frac{\lambda_0-0.01}{100-0.01}. \end{aligned}$$

From the posterior distribution, the sampling schemes for all the parameters and spatial effects can be obtained as follows,

$$\begin{aligned} \lambda_2 & \propto \lambda_1^{n/2+\alpha-1} e^{-\lambda_2[\frac{1}{2}(Y-X\beta-R\psi)^T(Y-X\beta-R\psi)+1/\gamma]} \\ & \propto Gamma(n/2+\alpha, \frac{1}{\frac{1}{2}(Y-X\beta-R\psi)^T(Y-X\beta-R\psi)+1/\gamma}). \end{aligned}$$

$$\begin{aligned} \lambda_1 & \propto \lambda_1^{n/2+\alpha-1} e^{-\lambda_1[\frac{1}{2}\psi^T V^{-1}\psi+1/\gamma]} \\ & \propto Gamma(n/2+\alpha, \frac{1}{\frac{1}{2}\psi^T V^{-1}\psi+1/\gamma}). \end{aligned}$$

$$\begin{aligned} f(\beta|\cdot) & \propto e^{-\frac{\lambda_2}{2}[\beta^T X^T X\beta - \beta^T X^T(Y-R\psi)-(Y-R\psi)^T X\beta]} e^{-\frac{1}{2}\beta^T(X^TX)\beta} \\ & \propto e^{-\frac{1}{2}[\beta^T(\lambda_2+1)(X^TX)\beta - 2\lambda_2(Y-R\psi)^T X\beta]}, \end{aligned}$$

with $\mu_\beta = \frac{\lambda_2}{\lambda_2+1}(X^TX)^{-1}X^T(Y-R\psi)$ and $\Sigma_\beta = \frac{1}{\lambda_2+1}(X^TX)^{-1}$, $\beta \sim N(\mu_\beta,\Sigma_\beta)$. If we set precision matrix $Q_\beta = (\lambda_2+1)X^TX$, then $\mu_{beta} = Q_\beta^{-1}\lambda_2 X^T(Y-R\psi)$.

$$\begin{aligned} f(\psi \mid \cdot) &\propto e^{-\frac{\lambda_2}{2}[\psi^T R^T R\psi - \psi^T R^T (Y - X\beta) - (Y - X\beta)^T R\psi]} e^{-\frac{\lambda_1}{2}\psi^T V^{-1}\psi} \\ &\propto e^{-\frac{1}{2}[\psi^T(\lambda_2 R^T R + \lambda_1 V^{-1})\psi - 2\lambda_2 (Y - X\beta)^T R\psi]}, \end{aligned}$$

with $\mu_\psi = \lambda_2(\lambda_2 R^T R + \lambda_1 V^{-1})^{-1} R^T (Y - X\beta)$ and $\Sigma_\psi = (\lambda_2 R^T R + \lambda_1 V^{-1})^{-1}$, $\psi \sim N(\mu_\psi, \Sigma_\psi)$. If we set precision matrix $Q_\psi = \lambda_2 R^T R + \lambda_1 V^{-1}$, then $\mu_\psi = Q_\psi^{-1}\lambda_2 R^T (Y - X\beta)$.
For range parameter $\lambda_0$, we will update it through Metropolis-Hastings algorithm. The acceptance ratio will be $min(1, p)$, where

$$p = \frac{\frac{1}{|V_{pro}|^{1/2}} e^{-\frac{\lambda_1}{2}\psi^T V_{pro}^{-1}\psi} (\lambda_0^{cur} - 0.01)}{\frac{1}{|V_{cur}|^{1/2}} e^{-\frac{\lambda_1}{2}\psi^T V_{cur}^{-1}\psi} (\lambda_0^{pro} - 0.01)}.$$

When observations are embedded in a regular lattice with some spots missing, the model we use here can be written as

$$Y = X\beta + RF\psi + \varepsilon,$$

where F is an incidence matrix indicating whether an observation belongs to a certain array cell. Under this setting, we assign precision matrix $\lambda_1 W$ to $\psi$ based on Gaussian random field, where W denote a tridiagonal matrix whose non-zero off-diagonal elements are $W_{i,i\pm1} = -1$ and $W_{i,i} = \sum_{j \neq i} W_{i,j}$. With discrete cosine transformation, $W = P^T \Lambda P$. Then $\psi \sim N(0, (\lambda_1 W)^{-1})$. We also assume $\beta \sim N(0, (X^T X)^{-1})$, $\varepsilon \sim N(0, \lambda_2^{-1} I)$, $\lambda_1 \sim Gamma(\alpha, \gamma)$ and $\lambda_2 \sim Gamma(\alpha, \gamma)$. For non-informative prior, we set $\alpha = 1$ and $\gamma = 200$. Then the posterior distribution of the parameters can be written as

$$\begin{aligned} f(\beta, \phi, \lambda_1, \lambda_2 \mid Y) \propto & \lambda_2^{n/2} e^{-\frac{\lambda_2}{2}(Y - X\beta - RF\psi)^T (Y - X\beta - RF\psi)} \lambda_1^{(n-1)/2} |W|_+^{1/2} e^{-\frac{\lambda_1}{2}\psi^T W \psi} \\ & |X^T X|^{1/2} e^{-\frac{1}{2}\beta^T (X^T X)\beta} \lambda_2^{\alpha - 1} e^{-\lambda_2/\gamma} \lambda_1^{\alpha - 1} e^{-\lambda_1/\gamma} \end{aligned}$$

From the posterior distribution, the sampling schemes for all the parameters and spatial effects can be obtained as follows,

$$\begin{aligned} \lambda_2 &\propto \lambda_1^{n/2 + \alpha - 1} e^{-\lambda_2[\frac{1}{2}(Y - X\beta - RF\psi)^T (Y - X\beta - RF\psi) + 1/\gamma]} \\ &\propto Gamma(n/2 + \alpha, \frac{1}{\frac{1}{2}(Y - X\beta - RF\psi)^T (Y - X\beta - RF\psi) + 1/\gamma}). \end{aligned}$$

$$\begin{aligned} \lambda_1 &\propto \lambda_1^{(n-1)/2 + \alpha - 1} e^{-\lambda_1[\frac{1}{2}\psi^T W \psi + 1/\gamma]} \\ &\propto Gamma((n-1)/2 + \alpha, \frac{1}{\frac{1}{2}\psi^T W \psi + 1/\gamma}). \end{aligned}$$

$$\begin{aligned} f(\beta | \cdot) &\propto e^{-\frac{\lambda_2}{2}[\beta^T X^T X \beta - \beta^T X^T (Y - RF\psi) - (Y - RF\psi)^T X\beta]} e^{-\frac{1}{2}\beta^T (X^T X)\beta} \\ &\propto e^{-\frac{1}{2}[\beta^T (\lambda_2 + 1)(X^T X)\beta - 2\lambda_2 (Y - RF\psi)^T X\beta]}, \end{aligned}$$

with $\mu_\beta = \frac{\lambda_2}{\lambda_2+1}(X^TX)^{-1}X^T(Y-RF\psi)$ and $\Sigma_\beta = \frac{1}{\lambda_2+1}(X^TX)^{-1}$, $\beta \sim N(\mu_\beta, \Sigma_\beta)$. If we set precision matrix $Q_\beta = (\lambda_2+1)X^TX$, then $\mu_{beta} = Q_\beta^{-1}\lambda_2 X^T(Y-RF\psi)$.

$$\begin{aligned} f(\psi \mid \cdot) &\propto e^{-\frac{\lambda_2}{2}[\psi^T F^T R^T RF\psi - \psi^T F^T R^T (Y-X\beta) - (Y-X\beta)^T RF\psi]} e^{-\frac{\lambda_1}{2}\psi^T W\psi} \\ &\propto e^{-\frac{1}{2}[\psi^T(\lambda_2 F^T R^T RF + \lambda_1 W)\psi - 2\lambda_2(Y-X\beta)^T RF\psi]}, \end{aligned}$$

with $\mu_\psi = \lambda_2(\lambda_2 F^T R^T RF + \lambda_1 W)^{-1} F^T R^T (Y-X\beta)$ and $\Sigma_\psi = (\lambda_2 F^T R^T RF + \lambda_1 W)^{-1}$, $\psi \sim N(\mu_\psi, \Sigma_\psi)$. If we set precision matrix $Q_\psi = \lambda_2 F^T R^T RF + \lambda_1 W$, then $\mu_\psi = Q_\psi^{-1}\lambda_2 F^T R^T (Y-X\beta)$.

# 4 Analysis of Red Dog Mine data

# 5 Discussion

# References

H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

D. G. Denison, B. K. Mallick, and A. F. Smith. A bayesian cart algorithm. *Biometrika*, 85(2): 363–377, 1998.

R. B. Gramacy and H. K. Lee. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 2012.