

Modeling PaleON biomass

Wesley Brooks

1. Introduction

Our objective is a model of the biomass of each species in each grid cell, where the only parameters used to calibrate the models are location and species composition. It is important to come up with an estimated probability that a given cell will have no biomass of a given species, and also to model the variance of the biomass estimate.

2. Philosophy

Consider two different kinds of errors:

- Natural variability in biomass (noise)
- Measurement error

Jun and I discussed the interpretation of a model for biomass. At the time I saw the effort being directed toward getting a model that tells us, based on our survey of the forests at time of settlement, how much biomass there was on the landscape at time of settlement. On the other hand, Jun says that the goal is to model the process that gives rise to biomass on the landscape, observed through the survey at time of settlement.

As a practical matter, one implication is that grid cells where there were small but nonzero biomass observations are treated differently. In my view, the fact that, say, a spruce tree was seen in grid cell 517 is enough to say that there is zero probability of there being no spruce biomass in cell 517. In Jun's view, we could easily imagine the same process populating the landscape in such a way that, randomly, cell 517 has no spruce trees. So there's a positive probability of zero spruce in cell 517.

If measurement error was the only possible error, then it we could say that observing even a single spruce tree in a cell means that spruce biomass is nonzero in that cell. But since there is also random noise, observing a spruce tree doesn't guarantee that spruce biomass will be nonzero the next day or year or whenever...

3. Chris' proposal

Chris proposed the following: Let $Y_{k,s}$ denote the biomass of taxon k in grid cell s . Let n_s be the number of PLS survey points within cell s , and note that two trees are sampled at each survey point. Then, letting $\hat{\lambda}_{s,i}$ be the estimate for total stem density (i.e. taxa are aggregated) at survey point s_i , $d_{s,i,j}$ the measured DBH for tree j at survey point s_i , and (a_k, b_k) allometric parameters for taxon k , the biomass measurement for taxon k in grid cell s is:

$$\begin{aligned} Y_{k,s} &= [n_s]^{-1} \sum_{i=1}^{n_s} \hat{\lambda}_{s,i} \sum_{j=1}^2 2^{-1} a_k d_{s,i,j}^{b_k} I(\text{tree}_{s,i,j} \text{ is in taxon } k) \\ &= [2 \cdot n_s]^{-1} \sum_{i=1}^{n_s} \hat{\lambda}_{s,i} \sum_{j=1}^2 h_k(d_{s,i,j}) A_{s,i,j}(k) \end{aligned}$$

Where $h_k(\cdot)$ is the allometric equation for taxon k and $A_{s,i,j}(k)$ is the indicator of the event that $\text{tree}_{s,i,j}$ is in taxon k .

Now we want to use these measurement of biomass as the observations on which a model of biomass will be based. For this, Chris proposes a log-normal model for $Y_{k,s}$ based on taxon k 's composition fraction, $p_k(s)$:

$$\begin{aligned}\log Y_{k,s} &\sim \mathcal{N}\left(g_k(s), \hat{V}_{k,s}\right) \\ g_k(s) &\sim \text{GP}\left(f(p_k(s)), \Sigma\right)\end{aligned}$$

That is, the Gaussian process $g_k(\cdot)$ sets the mean for a log-normal model of $Y_{k,s}$. The variance components are Σ , which is a spatial covariance, and $V_{k,s}$, which is random noise.

4. First look

Consider the ‘Oaks’ taxon. In Figure ??, the log of oak biomass is plotted on a heatmap. In order to

5. Models

Here are some ideas for the biomass model:

- Treat $g_k(\cdot)$ as a Gaussian Markov random field, coding the cells’ neighborhood structure, and then fit the model with INLA
-

There are two possible directions here: a two-stage hierarchical model, where in stage one we randomly decide whether a given cell has nonzero biomass, and then if so stage two randomly sets the biomass; alternatively, a single-stage model with zero-inflation, maybe one that could be tuned like a `glm` (e.g. a Tweedie model).

5.1. Two-stage model

The two-stage model has one stage for determining the probability that a given species has zero biomass in a cell and another stage for estimating the distribution of biomass, given that it is nonzero. Since in the observed data there is an almost perfect correspondence between cells where biomass is zero and cells where composition is zero, the first stage of the model needs to look at the composition of neighboring cells:

$$B_{k,s} \sim \text{Bernoulli}(\mu_{s,k})$$

$$\mu_{s,k} = \alpha_{s,k} + \sum_{i \in \mathcal{N}_s} \beta p_k(s)$$

5.2. Single-stage (Tweedie model)

Tweedie family of distributions for $\theta \in (0, 1)$:

$$N \sim \text{Poisson}(\lambda)$$

$$Z_i \sim \text{Gamma}(\alpha, \tau)$$

$$Y = \sum_{i=1}^N Z_i$$

$$\lambda = \phi^{-1} \frac{\mu^{2-\theta}}{2-\theta}$$

$$\alpha = \frac{2-\theta}{\theta-1}$$

$$\tau = \phi(\theta-1)\mu^{\theta-1}$$

$$\eta(s) = f(\mathbf{x}(s)) = g(\mu(s))$$

$$\mu(s) = g^{-1}(\eta(s))$$

Where $g(\cdot)$ is the link function, $\mathbf{x}(s)$ is the vector of covariates (x, y coordinates and composition fraction) at location s .

6. Figures

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

7. References

List of Figures

Oaks observed biomass (log scale)

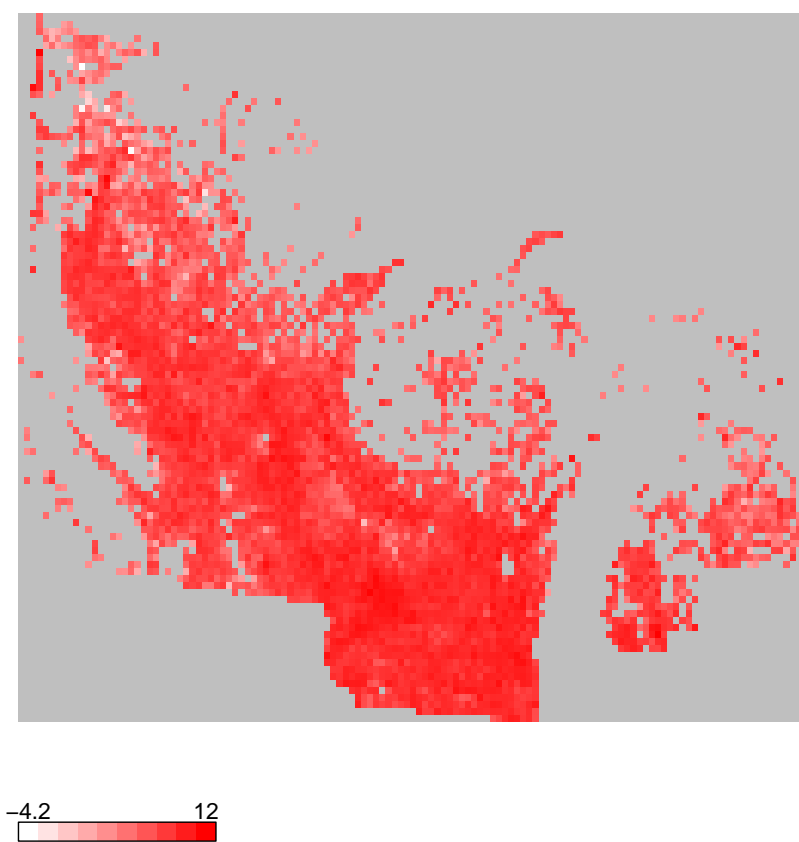


Figure 1

Oaks fitted biomass (log scale)

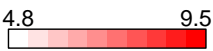
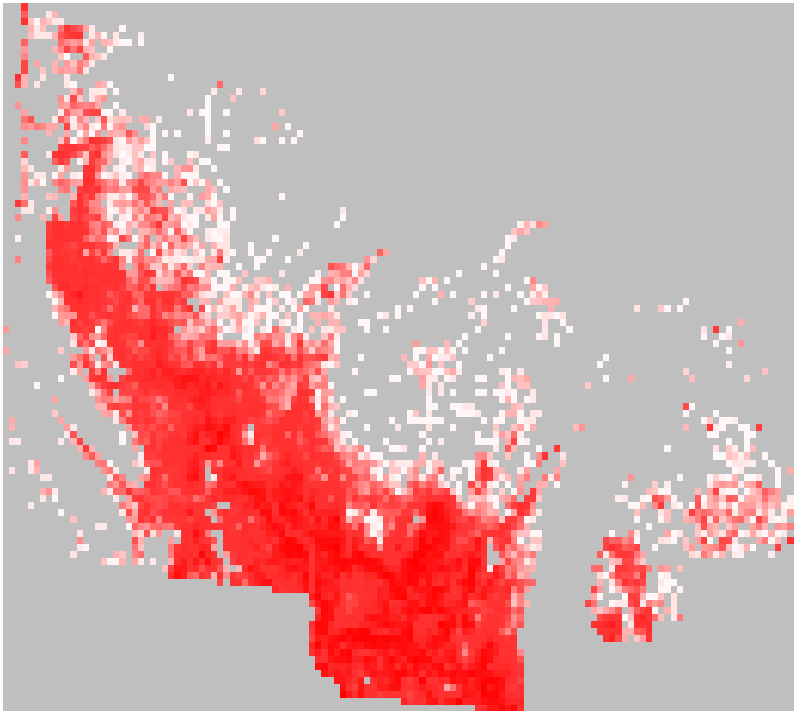


Figure 2

Oaks biomass residual (log scale)

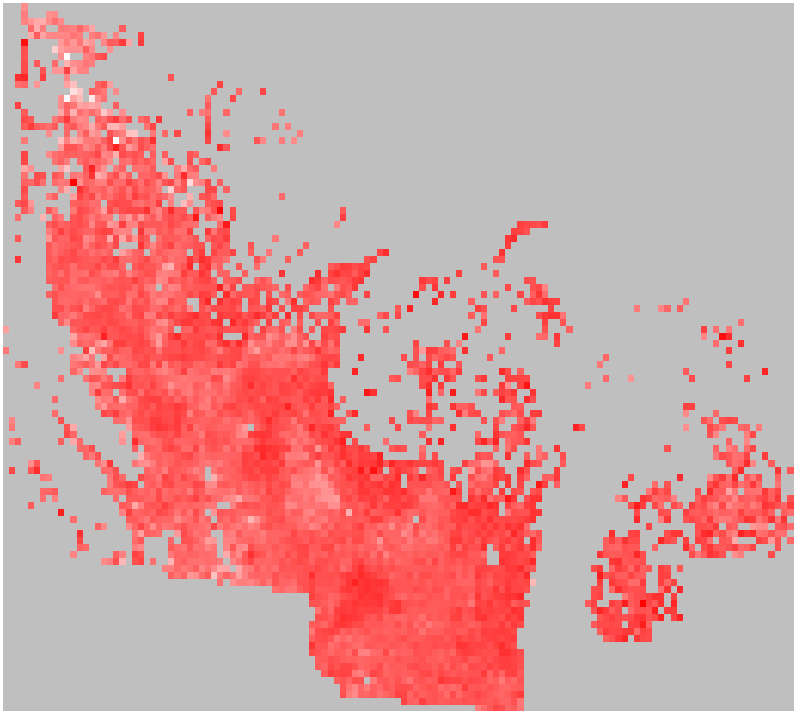


Figure 3

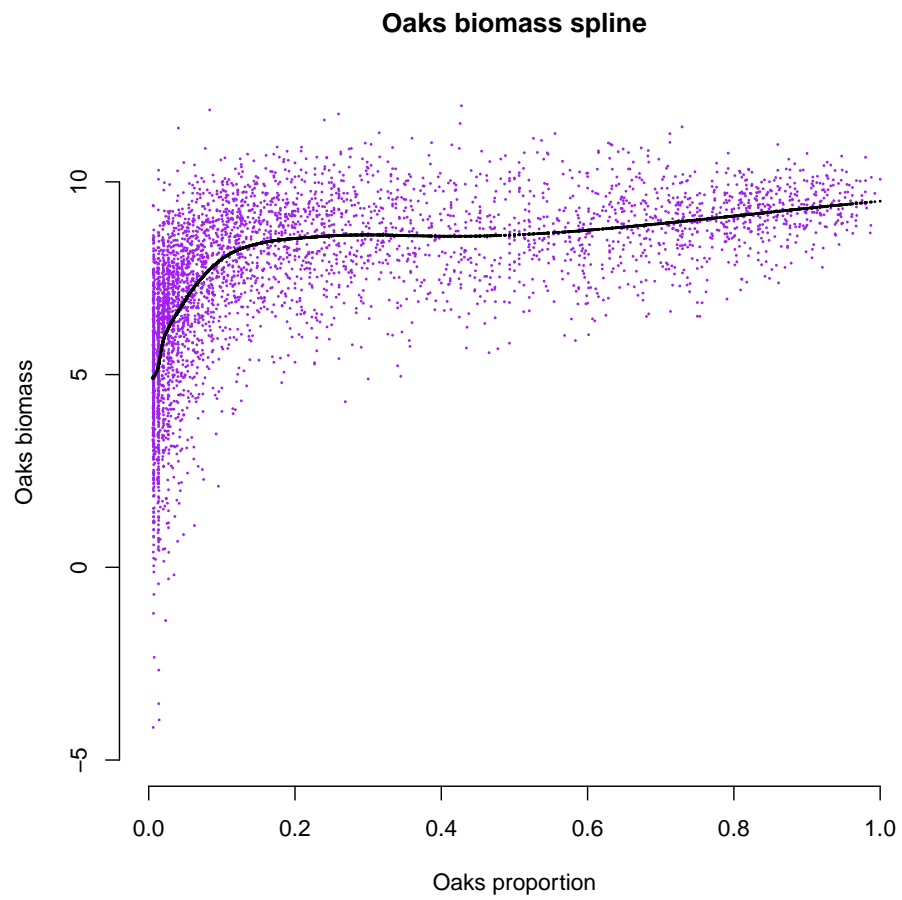


Figure 4