

Spatial smoothing of zero-inflated abundance data

Wesley Brooks

1. Introduction

The biomass of several taxa of trees was estimated in each grid cell of the upper midwestern United States (Minnesota, Wisconsin, and part of Michigan) from the Public Land Survey. This work describes the model that was used to spatially smooth the biomass estimates in order to produce draws from an underlying distribution of the per-taxon biomass in each grid cell.

In order to more quickly assess the model, it was fit to just the data from Wisconsin before being applied to the entire upper midwest. Also, because the model failed to converge for about half of the individual taxa, the decision was made to reduce the taxa to two categories: hardwoods and softwoods. The following taxa were summed in each grid cell to generate the observations of the softwood biomass:

- Tamarack
- Pine
- Fir
- Cedar
- Spruce

- Hemlock

The following taxa were summed in each grid cell to generate the observations of the hardwood biomass:

- Ashes
- Birches
- Elms
- Maple
- Poplar
- Basswood
- Oaks
- Willow
- Alder
- Ironwoods
- Walnuts
- Hickory
- Beech
- Celtis
- Cherries
- Juniper
- Rose trees

- Sycamore
- Cornus
- Buckeye
- Undifferentiated hardwood

2. Models

2.1. Tweedie distribution

The model for biomass is a Tweedie model, which can be thought of as a Poisson-Gamma mixture. Let the response be Y and $EY = \mu$. Now Y is a sum of N iid Gamma-distributed random variables Z_1, \dots, Z_N . The number, N , of iid Gamma random variables, has a Poisson distribution, which means it could be exactly zero. That is,

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

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

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

Two variables control the Poisson mean λ ; a power parameter $\theta \in (1, 2)$ and a dispersion parameter $\phi > 0$.

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

The parameters of the Gamma distribution are

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

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

And, as usual for a generalized linear model (GLM), we define a link function, $g(\cdot)$, that converts the fitted values \hat{y} to the linear predictor η . In this case, the log link is used.

$$\eta = g(\mu)$$

2.2. Generalized additive model

The model is fit to the data using a generalized additive model (GAM) that has the spatial coordinates $\mathbf{s} = (\text{lat}, \text{long})$ as its only covariates. The GAM for just the Wisconsin data used a spline smooth with 500 knots.

2.3. Tuning the Tweedie power parameter

The Tweedie distribution is in the exponential family only when the power parameter θ is prespecified. The usual parameter estimation machinery for a GAM or GLM requires that the observations be from a distribution in the exponential family, so using the Tweedie distribution requires the power parameter to be estimated separately from the parameters of the GAM. In the model for biomass, R's `optimize` function was used to find an estimate $\hat{\theta}$ that maximized the likelihood of the observed data.

2.4. Drawing from the model output

As in section 5.4.2 of (1)

3. Results

4. Bibliography

- [1] S. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall, 2006.

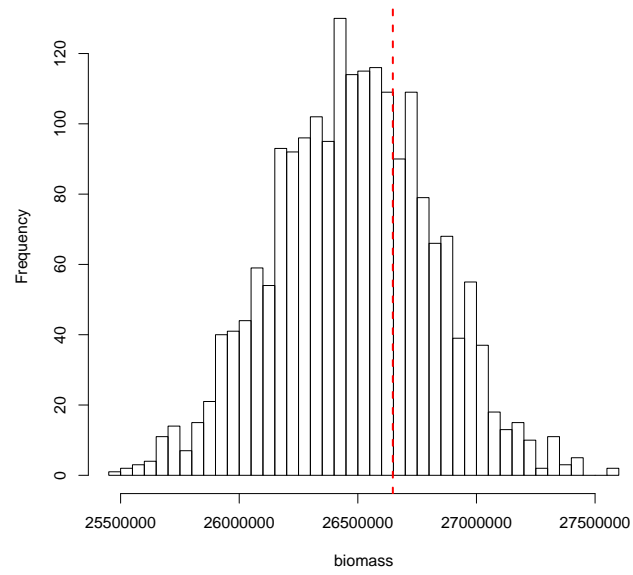


Figure 1: Distribution of the total hardwood biomass for Wisconsin. The vertical line represents the sum of the Wisconsin hardwood biomass observations.

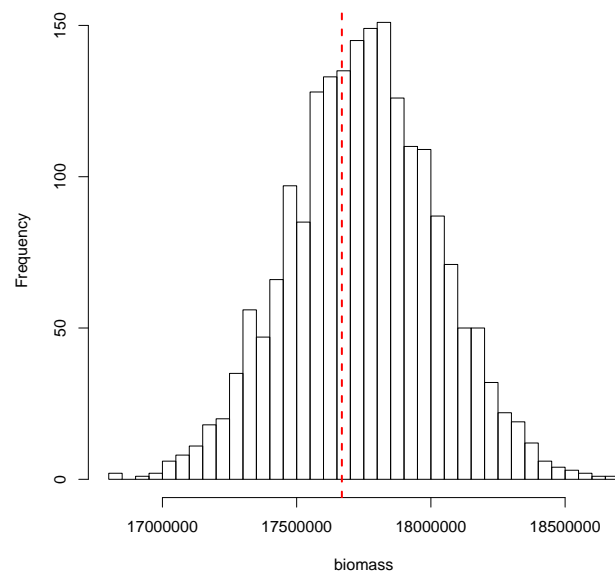


Figure 2: Distribution of the total softwood biomass for Wisconsin. The vertical line represents the sum of the Wisconsin softwood biomass observations.