Contents

1 Overview **1**

1.1 Simulations 2

2 PalEON biomass modeling **3**

2.1 Biomass data 3

2.2 Model 3

2.2.1 Fitting the GAM 4

2.2.2 Dividing biomass between taxa 4

2.2.3 Drawing from the "posterior" of biomass 4

2.2.4 Notes on the distribution of biomass among the taxa 4

1 Overview

In ecological applications and probably elsewhere it is common to observe data that are nonnegative and continuous, but that may be exactly equal to zero. An example of such data is the biomass of trees on a plot.

One method for modeling continuous data with exact zeros is to separately make a presence/absence model and a model for the response, given its presence. (Citations)

The Tweedie family of distributions can directly model a process that produces a continuous positive response with exact zeros. The form of the Tweedie distribution is

$$f(y; p, \phi, \mu) = a(y, p, phi) \exp\left[\phi^{-1}\left\{y\theta - \kappa(\theta)\right\}\right]$$

Where $\theta = (1-p)^{-1}\mu^{1-p}$, $\kappa(\theta) = (2-p)^{-1}\mu^{2-p}$, and $a(y, p, \phi)$ is complicated and not of interest right now (we'll get to it later). For $p \in (1, 2)$, the Tweedie distribution is equal to a sum of N independent draws from a Gamma(α, γ) distribution, where $\alpha = -(2-p)/(1-p)$, $\gamma = \phi(p-1)\mu^(p-1)$, and N is drawn from a Poisson(λ) distribution and $\lambda = \phi^{-1}(2-p)^{-1}\mu^{2-p}$.

A Tweedie distribution is a member of the exponential family if the power parameter p is known and fixed. In this case, the mean-variance relationship is that $\mathrm{E}(y) = \mu$ and $\mathrm{var}(y) = \phi\mu^p$. However, when p is unknown it must be estimated and there is currently no good way to

1.1 Simulations

Some simple simulations indicate that maximum likelihood estimates p well when p is near 1.5 and ϕ is near 1. When p is near 1 or near 2 and ϕ is 1, $\hat{p}$ is biased toward 1.5. When μ grows (by adding an intercept) and ϕ is set to 3, then $\hat{p}$ is biased toward 2 (or was for this simulation, where the true p was 1.3.) (These are MLE $\hat{p}$'s)

When ϕ and μ were changed, the location-scale-slope-based $\hat{p}$ was stable, if stil a bit biased.

The simulations noted above were for a simple linear model case:

```
n = 200
x = rnorm(n, 0, 1)
b = 5
y = rTweedie(mu=exp(10+x*b), p=1.3, phi=3)

ppp = seq(1.02,1.98,length.out=49)

models=list()
for (j in 1:length(ppp)) {
    print(ppp[j])
    models[[j]] = gam(y~x, family=Tweedie(link='log', p=ppp[j]))
}

p.sel.1 = function(p) {
    deviance(gam(y~x, family=Tweedie(link='log', p=p)))
}

p.sel.2 = function(p) {
    mod = gam(y~x, family=Tweedie(link='log', p=p))
    sqrt(abs(resid(mod, type='deviance'))) -> scale
    predict(mod, type='link') -> loc
    m = lm(scale~loc)

    return(abs(coef(m)[2]))
}

optimize(p.sel.1, interval=c(1,2))$minimum
optimize(p.sel.2, interval=c(1,2))$minimum
```

To do: further explore accuracy of these estimates of p, and make sure we're using a real MLE (as it is, I minimized the deviance estimated by mgcv - that should be accurate).

2 PalEON biomass modeling

2.1 Biomass data

Per-grid-cell biomass data for the upper midwest, calculated from stem density and basal area via allometry. Simon did those calculations. For prototyping, we use only Wisconsin data (the biomass models take a while to run).

2.2 Model

Biomass is a positive, continuous quantity. It can be exactly zero where there are no trees growing, so we presume that the biomass follows a Tweedie distribution. An alternative would be to model biomass in two stages: presence/absence and then distribution.

We're going to use a generalized additive model (GAM) with the sole predictor being a smooth on spatial location. R-INLA would be a great way to model biomass in a Bayesian context with the linear predictor being a GMRF, but there is no Tweedie likelihood available for R-INLA, and adding one turns out to be difficult (due to Rue's choice of C compiler).

2.2.1 Fitting the GAM

There is not a generally agreed-upon method of estimating the Tweedie distribution's "power" parameter, p. But note that the Tweedie family is a type of "Exponential dispersion model", meaning that the "variance function" relating the variance to the expected value can be written $\mathrm{E}(y) = \mu$, $\mathrm{var}(y) = \phi\mu^{\theta}$ where θ is the power parameter and ϕ is the dispersion parameter.

2.2.2 Dividing biomass between taxa

There is concern that the variance of a sum of biomasses, modeled for individual taxa, would be greater than the variance of a singel model for total biomass. I think this concern is backward, since e.g. $x^2 + y^2 + z^2 < (x + y + z)^2$. In any case, the plan is to make a total biomass model and divvy the total fitted biomass between the taxa based on draws from the biomass models for individual taxa.

2.2.3 Drawing from the "posterior" of biomass

Our goal is to calculate a distribution of total biomass, so we need to be able to make draws from the fitted model.

Consider that the smooth coefficients are drawn from a gaussian distribution with covariance. That distribution is conditional on the GAM's smoothing parameters (as always) and also on the Tweedie power parameter (specific to Tweedie models). We'd like the marginal distribution of biomass (not conditional on those fitted values) so we're going to use the parametric bootstrap to find their distributions. This part is covered in the Simon Wood mgcv book, sections 5.2.7, 5.4.2, and 4.9.3.

This will have to be done for all taxa, I think... it's going to be a slow process.

2.2.4 Notes on the distribution of biomass among the taxa

If the distribution of the biomass was gamma (and all taxa had the same shape parameter) then the distribution of biomass among taxa would be Dirichlet. It's unclear whether there is a friendly analog (generalization?) to the Dirichlet when the biomass follows a gamma distribution.