Bayesian Regression of Sap Flux in Saplings on Environmental Conditions

David Clancy*
Duke University
April 20, 2016

Abstract

Currently there is a lack of literature which models physiological processes of saplings using observations taken at a rate of greater than once per day. Drawing on literature about the effects of light, vapor pressure deficit, and soil moisture on the growth of saplings as well as the effects of those quantities on physiological processes on mature trees, a relatively simple bayesian regression model is fit to the sap flux data of 68 saplings in central North Carolina. The results show that saplings tend to vary less between species relative to their mature counterparts with respect to the effect of vapor pressure deficit on sap flux. In addition, there is preliminary evidence that the most common functional forms for the effect of light on sap flux may be lacking due to the inability to allow for a maximum value. While the usefulness of soil moisture is still debated in current literature, in these data the inclusion of soil moisture data did not improve the fit or appropriateness of the model to a noticeable degree. While the methods used here are fairly simple, the findings indicate that further study of saplings at a fine time scale and more complex models for dealing with data from saplings are likely warranted due to differences between the results of this model and existing literature on mature trees.

*David Clancy is a member of the Class of 2016 and graduating with a degree from Department of Statistical Science, Duke University, Durham, NC 27708, USA. (e-mail: david.statistics@gmail.com).
1 Introduction

Many factors go into determining the metabolism of saplings and thus, it may be difficult to isolate differences that arise due to species or specific environmental conditions. Yet understanding how young trees respond to the current conditions may provide valuable insight on recruitment, survival, and forest dynamics. Zhu et al. (2012) show that despite predictions that species would begin to shift or expand their range in response to climate change, most species (58.7%) show signs of range contraction and only very few (4.3%) show signs of expansion and both the northern and southern boundaries of their range.

Modeling the relationships between measured sap flux density of saplings and the surrounding environmental conditions could produce results which could help explain why some species are able to migrate while others are failing to do so. In addition, a deeper understanding of these processes during the early phases of growth for trees may be used in further research to improve forecasts for events such as long-term species migration under different global climate forecasts.

2 Literature Review

The majority of literature on modeling tree physiology focuses on mature individuals or larger scale processes. While there is still a sizable portion of research on saplings, almost all of that research uses a data at an annual or seasonal scale. In general, the response variable of interest is the annual growth rate instead of metrics at a finer scale. Therefore, when reviewing existing research related to the model of interest for this study, one must bear in mind differences between work done at coarse and fine time frames.

Kobe and J.Hogarth (2007) investigated the relationship between measured light levels and growth of Acer saccharum and Fraxinus americana saplings. The authors here chose to use the overall annual growth as the response variable in their analysis which puts the predictions on a much longer time-scale than the analysis which follows in later sections. In fact the metric which produced the best results for this study was simply looking at the percent of canopy openness for each individual. One point of repeated discussion is the effects of higher intensity sunflecks on growth. Direct sensors will obviously pick up these events with no lag, and these sunflecks can record levels much higher than the level of light at which photosynthesis saturates. In the context of the following study, whatever functional form is chosen for light, it must either achieve a maximum or asymptotically approach a limit so that sunflecks do not have a large amount of leverage over the effect of light on sap flux.

A prior study by Pacala et al. (1994) examined effects of not only light but also soil moisture on the growth of saplings. This study included 641 saplings from ten different species across of range of sizes. In this analysis, soil moisture did not contribute enough information to be useful in predicting growth. However, since the time-scale of the response variable was again on an annual scale, this does not necessarily suggest soil moisture should be left out of the following analysis. For the regression analysis, the
authors chose to use Equation 1 for the effect of light on growth.

\[ p(L) = \frac{AL}{A + L} \]  \hspace{1cm} (1)

Equation 1 is a concave, monotonically increasing function in light, \( L \). \( A \) and \( S \) are constants which the paper estimated for each species. The results from the estimates of \( A \) and \( S \) provide insight into how different species react to different light levels. While there was no distinct clustering of species, \( A \) and \( S \) seemed to have an inverse relationship which suggests that certain species prefer low light and will outcompete others in those conditions while others prefer higher light. In contrast to Kobe and J.Hogarth (2007), this regression was only able to explain between 25-50% of the variance in sapling growth.

Despite soil moisture not providing any additional, useful information in Pacala et al. (1994), a more recent study by Kobe (2005) shows that in \( A. \ rubrum \) and \( A. \ saccharum \) (\( n = 73, 70 \) respectively) soil water availability is important predictor. This analysis uses the same form as 1 for modeling the effects of light on the growth of saplings and included additional information about the amount of available nitrogen in the soil, which is not a variable measured in the data below. Once again, this study was not concerned with very frequent measurements of growth or environmental conditions.

Vapor pressure deficit (VPD) is one of the environmental variables about which data is present in the following analysis. VPD is a measure of the difference between the actual water vapor pressure and the saturation water vapor pressure at a particular temperature. This is generally derived from both temperature and relative humidity measurements. In addition, VPD tends to only be used in models using data that is collected at a fairly fine time scale. This poses a possible problem since almost all modeling for saplings is focused on the annual growth rates.

Lendzion and Leuschner (2008) is one of the only studies which discusses the effects of VPD on the growth of any type of sapling. The paper itself also mentions the lack of literature examining this relationship. The authors only focused on European beech (\( Fagus sylvatica \)) saplings which, as the name implies, are not native to North America. While they are used in ornamental landscaping in North America, they are not present in the areas of the individuals used in the following study. The authors set up an experiment and showed that growth rates of the saplings depended on soil moisture and VPD level. Specifically, higher levels of soil moisture tended to increase growth rates and increases in VPD at the lower end of the spectrum increased growth rates while high VPD levels tended to show a decline in growth rates.

Many more studies have been published using data sets from mature trees. In this case, most find positive linear or nearly linear effects of VPD on transpiration and sap flux in a variety of species across North America while the VPD is low (< 1kPA). However, the relationship quickly becomes inversely related and once VPD exceeds a tipping point; increasing VPD results in an exponential decrease in transpiration and sap flux. Hogg and Hurdle (1997) observed that in two aspen stands (\( Populus tremuloides \)) in Canada the relationship between sap flux and VPD was increasing linearly up to a limit (about 1kPA) and then leveled off. Similarly, Oren et al. (1999) found that...
in Bald Cypresses (*Taxodium distichum*), the relationship between daily transpiration and mean daytime VPD was nearly linear. However, in this study, VPD was averaged at a daily level and never exceeded 1kPa, which was the point at which it leveled off in the previously mentioned study. Finally, Oren et al. (1999) show that Equation 2 describes the relationship between stomatal conductance, $g_s$, and VPD, $D$, well for a wide variety of species in mature individuals. In addition, this parameterization results in $m$ and $g_{sref}$ being highly correlated with a slope of approximately 0.6.

$$g_s = g_{sref} - m \ln D$$

Throughout all of these studies, a few key results should be used in the novel modeling of sap flux in the following sections. First of all, whatever functional form is used for the effect of light on sap flux, it should be concave and possibly should be allowed to achieve a maximum. Soil moisture generally is not too useful when looking at an annual scale, though it sometimes contributes enough information for specific species to warrant being left in the model. When working at a finer scale though, moisture could retain usefulness. The effect of VPD should be generally positive at low levels but after reaching a certain point, the effect should decrease and have a convex shape.

## 3 Data

All cases used in this study are come from an experimental warming site near Hillsborough, North Carolina. The site has a mean annual temperature of 14.5°C and annual precipitation of 1208mm, of which almost all falls as rain. The site consists of 18 chambers - three for each combination of temperature (ambient, ambient + 3°C, and ambient + 5°C) and light levels (gap and understory). At each chamber, hourly environmental conditions were monitored and for this study is interested in air vapor pressure deficit (VPD, $D$), soil moisture, and light (Q) (Clark et al., 2014). VPD was not measured directly but rather was derived at time $t$ with the following formula depending on air temperature $T$ and relative humidity $R$,

$$D_t = \frac{1}{10} \left( 6.112 \exp \left( \frac{17.67 \times T_t}{T_t + 243.5} \right) - \frac{R_t \times 6.112 \exp \left( \frac{17.67 \times T_t}{T_t + 243.5} \right)}{100} \right).$$

For the purposes of this modeling, only individuals in the gaps chambers have been included. Several factors were considered for this decision. First of all, sunflecks are an important factor in photosynthetic processes (Robin L. Chazdon, 1991; D. A. Sims, 1993). Since light for the chamber was measured with a single sensor, it may have experienced sunflecks at different times than the individuals themselves. Over longer time periods, one can average over the data or use a method that utilizes photographs to estimate the impacts of sunflecks (Finzi and Canham, 2000; Canham, 1988); however, because of the interest in a time step of less than one day, the disconnect in timing of the sunflecks is much more of a concern. This could also be seen empirically by examining the correlation between trees of a single plot. Those correlation values tended
to be higher in gaps plots than in shaded plots and since all environmental predictors rely on the chamber level measurements being representative, the individuals in shaded chambers were ultimately left out of the modeling process.

Information about sap flux from 68 individuals is included in these data. Table 1 shows the distribution of these individuals across species.

<table>
<thead>
<tr>
<th>Species</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Acer rubrum</em> (Acru)</td>
<td>22</td>
</tr>
<tr>
<td><em>Liriodendron tulipifera</em> (Litu)</td>
<td>20</td>
</tr>
<tr>
<td><em>Quercus alba</em> (Qual)</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 1: Sample sizes for each species.

The sap flux density was measured using Granier-type thermal probes at a rate of once every thirty seconds. These measurement were averaged over 15-minute periods of time and those averages were used here (Granier, 1985; Granier and Gross, 1987). Each individual had a pair of probes which measured a difference in the temperature at time $t$ between the heated probe and the unheated probe. This difference is used to calculate the sap flux density $J_{i,t}$ (Lu et al., 2004). It should be noted that there are known issues with both thermal dissipation probes as well as the empirical relationships used to find the sap flux density (Bush et al., 2010; Steppe et al., 2010). However, these issues are outside of the scope of this study. Again, as before the majority of research on these probes has been done on mature trees and thus using these techniques with saplings may be introducing even more issues. The sap flux data includes a sizable amount of zeros. In general, these zeroes tended to occur in streaks and were more prominent during low light conditions.

4 Model

A bayesian regression model was constructed for the latent sap flux at time $t$ using the data from the environmental sensors in the individual's plot. The model for the latent quantity is given by Equation 3.

$$\tilde{J}_{i,t} = f(Q_{i,t})g(D_{i,t}) + \rho \tilde{J}_{i,t-1} + a_i + \epsilon_{i,t}, \epsilon \sim \mathcal{N}(0, \tau)$$

(3)

Latent sap flux was used because, as previously mentioned, there is a considerable point mass at 0 for the observed values. These zeroes could be due to several things including quantities too small for the probes to detect or some sort of malfunction. Thus, whether or not the observed value is zero should be dependent on at least the previous measurement and the light level. To account for the point mass, a probability $p_{i,t}$ is estimated and with probability $p_{i,t}$ the observed quantity, $J_{i,t}$ is equal to the latent quantity, but with probability $1 - p_{i,t}$ the observed quantity is 0. The estimation of this probability comes from Equation 4.

$$\log \frac{p_{i,t}}{1 - p_{i,t}} = b_0 + b_1 f(Q_{i,t}) + b_2 \tilde{J}_{i,t-1}$$

(4)
For the effect of \( Q_{i,t} \) on \( J_{i,t} \) and \( p_{i,t} \), the following form is used,

\[
f(Q_{i,t}) = \frac{\alpha - \beta Q_{i,t}}{1 + \gamma Q_{i,t}} Q_{i,t} - R
\]

This functional form comes from Ye and Yu (2007). When modeling the effect of light, this set up gives several nice properties. First of all, like most forms used for these applications, it gives a compensation point where there is enough light for photosynthetic processes to force a net gain of CO\(_2\). In addition, \( \alpha \) is the initial slope for the model. There is also the condition that \( 0 \leq \beta, \gamma \leq 1 \) and thus, these two parameters govern the light level at which the model will peak. The result is that \( f \) will grow near-linearly at low light levels but at higher light levels \( f \) will level off and could begin to decrease at very high light levels. Note that while \( f(Q) \) will always have a maximum, that maximum need not occur in the range of light values observed in the dataset. Thus, this form has the flexibility to account for a peak if the data dictate there is one but it does not force a peak. Figure 1 gives an example for the effect of \( Q_{i,t} \) on \( J_{i,t} \). This functional form is consistent with existing research on the effect of light on growth as discussed in the Literature Review section. Given the constraints on \( \beta \) and \( \gamma \), a beta distribution was used for the priors. Using information from Ye and Yu (2007) as well as examination of the range of values involved, the following weakly informative priors were used, \( \beta \sim \text{Beta}(1,9) \) and \( \gamma \sim \text{Beta}(1,9) \).

![Figure 1: Example of \( f \) with mean posterior estimates for \( L. \) tulipifera at the experimental warming site.](image)

The effect of \( D_{i,t} \) on \( J_{i,t} \) is,

\[
g(D_{i,t}) = G_{\text{ref}} - \lambda \ln(D_{i,t}).
\]

This is a form which monotonically decreases with the vapor pressure deficit and is used by Oren et al. (1999) and Bell et al. (2015). However, since these data are
from saplings, when setting up the prior for $G_{\text{ref}}$, there were no differences between the species and the variance was larger than given by their work on mature trees. For the corresponding variable in Bell et al. (2015), the means for the prior on $G_{\text{ref}}$ ranged from 18 to 97 for the species in this data set and so for these data $G_{\text{ref}} \sim \mathcal{N}(50,0.05)$. For $\lambda$, the prior used was equivalent to Bell et al. (2015) and was set as $\lambda \sim \mathcal{N}(0.6 \cdot G_{\text{ref}}, 0.1)$. The reasoning for the relationship between $G_{\text{ref}}$ and $\lambda$ along with the value of 0.6 comes from Oren et al. (1999). While this does not take into account the findings of Lendzion and Leuschner (2008) and others who found an increasing relationship while VPD was small, it is a generally accepted and widely used form when modeling physiological processes.

Implementation of the above model was done in the R language and the rjags package (R Core Team, 2015; Plummer, 2016). After setting up the model described above in rjags, runs of the sampler were made for each of the three species where the first 10,000 results were part of the burn-in period and then the results from the following 12,000 iterations were kept. Appendix A gives the specifications for each of the priors used in the model.

## 5 Results and Discussion

The model outlined above performed reasonably well. Table 2 shows the percentage of variability in sap flow which was explained by the model. The trend of the model performing the worst for the red maples ($Acer \text{ rubrum}$) can be seen throughout the results for the estimation of all of the parameters. In addition, the point mass at 0 seems to have been handled very well by the model. About 35% of observations were equal to zero and of those, 85% were predicted correctly and only 4% of non-zero values were assigned a prediction of 0.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qual</td>
<td>0.6542023</td>
<td>0.6320572</td>
<td>0.6748933</td>
</tr>
<tr>
<td>Litu</td>
<td>0.5997571</td>
<td>0.5816373</td>
<td>0.6165520</td>
</tr>
<tr>
<td>Acru</td>
<td>0.5011325</td>
<td>0.4752171</td>
<td>0.5252726</td>
</tr>
</tbody>
</table>

Table 2: Mean posterior estimates for the proportion of variability explained by the model along with the 95% credible intervals.

To check the appropriateness of the model and functional forms as well as to check for a well behaved sampler, trace plots and other diagnostic plots were examined for each of the parameters being estimated. In addition, residuals were checked against fitted values and environmental predictors to ensure the functional forms were not failing to capture any patterns. Due to space, only a couple trace plots are shown in Figure 2, but they are representative of what each parameter showed for each species. In addition to just plotting the values by iteration, a cumulative mean was also included to insure that the posterior distributions were well behaved.

One difference between the model implemented here and the models in the cited literature on which it was based is that there was not an assumption that values of
Figure 2: Example of two trace plots from the sampler dealing with data from white oak saplings. The top shows the trace plot for $\beta$ while the bottom is the trace plot for $\gamma$. The cumulative mean is shown by the dark green line. Plots are indicative of good mixing in the sampler and reliable parameter estimation.

$G_{\text{ref}}$ for mature trees hold for saplings. Table 3 shows the mean posterior estimates and 95% credible intervals for $G_{\text{ref}}$ across all three species. According to Oren et al. (1999) the mean values for Q. alba, L. tulipifera, and A. rubrum, are 18, 97, and 88 respectively. While this model still shows white oak to have a considerably smaller value of $G_{\text{ref}}$ when compared to the other species, the differences between the three are greatly diminished. Thus, future models for saplings should not necessarily use priors for the functional form of VPD which come from estimates taken from mature trees. It is worth noting though that these results lend credibility to the idea that the ordering of species may stay the same or similar for saplings. If this holds true in future study, priors may take this into account.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qual</td>
<td>47.81665</td>
<td>39.66760</td>
<td>56.65878</td>
</tr>
<tr>
<td>Litu</td>
<td>63.05251</td>
<td>63.02385</td>
<td>63.07558</td>
</tr>
<tr>
<td>Acru</td>
<td>68.61502</td>
<td>68.61421</td>
<td>68.61587</td>
</tr>
</tbody>
</table>

Table 3: Mean posterior estimates for the value of $G_{\text{ref}}$ along with the 95% credible intervals.

A novel aspect of this model is the incorporation of $f(Q)$ as, to the authors knowledge at the time of writing, this functional form has not been incorporated into a model incorporating elements other than simply light. The primary difference between using this form and forms such as 1 or the form used by Bell et al. (2015) is that in this case
$f(Q)$ is not necessarily monotonically increasing within the support of the data. However, the drawback for this flexibility is that $f(Q)$ incorporates four distinct parameters and thus comes at a higher computational cost than other forms which always increase. This cost comes in the form of extra unobserved nodes in the graphical model used by JAGS which results in a longer runtime.

Figure 3 shows the estimates for $f(Q)$ in the three different species as well as a 95% credible intervals throughout the range of $Q$. It’s clear that in all three species, it’s reasonable to think that in the upper limits of the light data present, the effect of light on sap flux may begin to decrease. All of the lower bounds of the intervals show a maximum in the observed range. Using the mean posterior estimates, tulip poplar and white oak saplings both respond to greater levels of light with greater sap flux. However, in red maples, a maximum is very clearly reached within the support of the data. Therefore, based on the results of this model, the use of this flexible functional form for light is justified and future models for saplings should consider using a functional form which allows for a maximum.

In addition to the model discussed up to this point, several other variations were fit. The most interesting finding from this comparison is that the addition of soil moisture does not provide substantially more information than the form of Equation 3. When incorporating soil moisture $M$ into the model, the following equation for latent sap flux was used,

$$
\tilde{J}_{i,t} = f(Q_{i,t})g(D_{i,t})M_{i,t} + \rho \tilde{J}_{i,t-1} + a_i + \epsilon_{i,t}.
$$

Other modifications included removing the dependency on the previous observation of sap flux and/or removing the random effect term. In both of these cases, Equation 3 performed noticeably better. Keeping those terms is also further supported by the fact that the 99% credible intervals for $\rho$ and most $a_i$ do not include 0. An interesting observation is that the random effects for a specific plot tend to fall close to one another. However, given that most plots have no more than 2 individuals of a given species, the effect was kept at an individual level instead of at a plot level as there simply wasn’t enough evidence to move it to the higher level and it is more interpretable at the individual level.
Figure 3: Posterior mean estimates with 95% credible intervals for $f(Q)$ throughout the range of $Q$ for each species. The fact that all lower bounds achieve a maximum along the support of the data and the posterior mean of *A. rubrum* achieves the maximum at a relatively low value of $Q$ gives support to the idea that it is worth using this slightly more complex functional form for the effect of light.
In addition, the modeling approach used was limited by computational limits. R performs much slower than lower level languages such as C, C++, and Fortran. As a result, if a custom sampler were to be implemented in such a language, a more complex analysis could be possible. All of the above computation was performed on a personal computer and thus the time needed for samplers was often limited by the amount of time the computer could stay in one place. One possible area of further exploration of this dataset would be the incorporation of the experimental warming treatments in the analysis.

Acknowledgements

Thank you to all those who have helped me throughout this process, especially Dr. Alan Gelfand, Dr. Jim Clark, Dr. Mine Cetinkaya-Rundel, and Matthew Kwit.

A Appendix: Priors

The table below gives the specifications of the priors for the parameters used in the model outlined in the text. Note that many of the parameters use a truncated normal form given by appending at least one bound through use of the indicator function, 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>$\mathcal{N}(1,1)1(\rho \geq 0)$</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Inv-Gamma(1,20)</td>
</tr>
<tr>
<td>$a_i$</td>
<td>$\mathcal{N}(0,10)$</td>
</tr>
<tr>
<td>$b_{0,1,2}$</td>
<td>$\mathcal{N}(0,0.01)$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\mathcal{N}(0,0.01)1(\alpha \geq 0)$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Beta(1,9)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Beta(1,9)</td>
</tr>
<tr>
<td>$R$</td>
<td>$\mathcal{N}(0,0.001)1(R \geq 0)$</td>
</tr>
<tr>
<td>$G_{\text{ref}}$</td>
<td>$\mathcal{N}(50,0.05)1(G_{\text{ref}} \geq 0)$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>$\mathcal{N}(0.6G_{\text{ref}},0.1)1(0.4G_{\text{ref}} \geq \lambda \geq 0.8G_{\text{ref}})$</td>
</tr>
</tbody>
</table>
References


