Multilevel Urban Tree Allometric equations

erker

October 31, 2018

Introduction

Multilevel models have been used for decades in tree growth equations (Lappi and Bailey 1988). Indeed the test dataset, "orange", in the statistical programming language, R, is used to demonstrate the fitting of nonlinear multilevel (mixed effects) models R Core Team (2016) orange. Multilevel modeling is an attractive approach because it provides a coherent framework to account for the many levels of observation or of groupings in data and to pool information across groups. This paper has two main contributions. First, we demonstrate the use of Stan via the "brms" package in R to fit bayesian nonlinear multilevel models to predict tree diameter growth from age Carpenter et al. (2017); Bürkner (2017). Second, we apply the method to the Urban Tree Database McPherson et al. (2016a,b). This dataset is the result an over a decade long effort to collect age and size data on thousands of trees in 17 cities across the US. Multilevel modeling has the potential to extract more information from the data and improve predictions compared to the existing modeling approach. Improving predictions of tree size from tree age will improve our ability to predict the important ecosystem services these trees provide urban dwellers.

Stan is a probabilistic programming language for bayesian inference Carpenter et al.

(2017). It uses No-U-Turn sampler, an adaptive form of Hamiltonian Monte Carlo sampling, to effectively draw samples from the specified log posterior density. Here, we access Stan via the R package brms Bürkner (2017). brms allows the user to specify the likelihood and priors in syntax similar to the R package lme4 commonly used for frequentist? multilevel (mixed effects) models. This makes harnessing the power of Stan much simpler and concise because it doesn't require the user to know how to write efficient Stan code and can convert a few lines of R code into many lines of Stan. brms is not as flexible as stan, but still can be used to fit many types of models including nonlinear multilevel regression models, such as ours here. Some of the key advantages of fitting a model in Stan via brms include relatively simple syntax and efficient posterior sampling for multilevel non-linear models. The bayesian approach gives better estimates of parameter uncertainty and provides a formal way to include prior information.

The existing approach to modeling the diameter growth of trees in the urban tree database (UTD) was to fit a separate model for each tree species in each city and test several model forms with different weights and then select the model with the lowest Akaike Information Criterion (AIC). This approach has several limitations, many highlighted in the report. First, while the model form selected provided the lowest AIC, many of the estimates are not biologically realistic (for example they begin to increase sharply at old ages, cubic fits, or decrease at old ages, quadratic fits). Therefore, the researchers discouraged applying the models beyond the range of the data, or sometimes even within the range of data if the estimates were unrealistic. These unrealistic estimates and the inability to extrapolate severely limits managers' ability to predict growth over meaningful time scales (a century rather than a few decades). A second limitations is that some models predict negative diameters, an impossibility. Third, models are only provided for the cities and the species sampled. If a manager wants to predict the diameter growth of a tree species in an unsampled city, the researchers recommend using the model from the reference city in the same climate region. However many of the reference cities are on the border of climate regions and there is known large variability in growth within regions (see figure XX in utd report which is from McPherson.... comparing Cheyenne to Ft Collins). Furthermore, if a manager wants to predict growth for an unsampled species or a species that was sampled in a different city it is not obvious which equation/model to use and the additional uncertainty that this introduces is not quantified.

Our approach addresses the above limitations. First, we use a weibull curve, commonly used in foresty growth equations and biologically realistic, which makes extrapolation to ages outside the data range less fraught. Second, using this sigmoidal curve and modeling diameter with a gamma distribution ensures our estimates of diameter are positive. Third, by modeling the weibull curve parameters as functions of species, city, and climate, we are able to borrow information across cities and across species to provide predictions an associated uncertainty of diameter growth even in even in cities or species with very little or no data.

Multilevel modeling in non-urban forestry .... other work. how is this different? Well the urban part, management practices like topping, pollarding, pruning, drastically alter growth and tree dimensions. Others have discussed the difference between forest and urban/open grown trees a fair bit. The Stan and brms part, these software are new. What about the modeling approach? I don't think many past effects have so many levels, or have things vary by climate. The geographic extent of these urban models is very large. There are a number of papers that have done multilevel modeling for tree growth that I should mention.

A paragraph on the impact of these equations for managing forests to predict/forecast ecosystem services.

Methods

Data

The urban tree database (UTD) consists of measurements on 14487 trees of 170 species in 17 cities. However, largely because of the difficulty is measuring tree age, there are only

12687 trees with complete age and diameter data (161 species, 17 cities, 309 species by city combinations 1).

Some species were measured in multiple cities, but not most. The number of trees of each city by species combination sampled ranged from 1 (both Liquidambar styraciflu and Prunus serrulata in Queens, NY) to 79 (Quercus laurifolia in Charleston, SC). The median number of trees in each species-city combination was 37.

Age is defined in this dataset as time since planting, since this is the record kept by cities. Actual age of the trees may be several years more. Diameter (cm) of the trees is measured at breast height (1.37m above ground).

In the UTD, trees are classified taxonomically down to cultivar for some individuals, but here we aggregate cultivars up to the species level. Species are then nested within Genera.

The 17 cities in the UTD cover much of the US geographically, 2, and much of the variation in climate, 3. However, New York City only has a few observations and the data for Indianapolis is missing too.

Rather than using the aggregated sunset zones as done in UTD, we used growing degree days (GDD) and precipitation data from climate NOAA's climate normals to continuously vary equation parameters across climate. Figure 3 shows each census tract centroid in the conterminous US plotted in GDD and precipitation space. We appoximated the GDD and precipitation for each tract by assigning the values of the weather station closest to the centroid. This allows us to vary our model continouously across geographic space in a way that better captures the natural gradients of climate.

Modelling

Model requirements

We sought a model of tree growth that would adequately represent the known biological dynamics of tree growth, namely that diameter growth rate starts slow, reaches a maximum at a young age, and then gradually declines to nearly zero. Diameter for trees much always

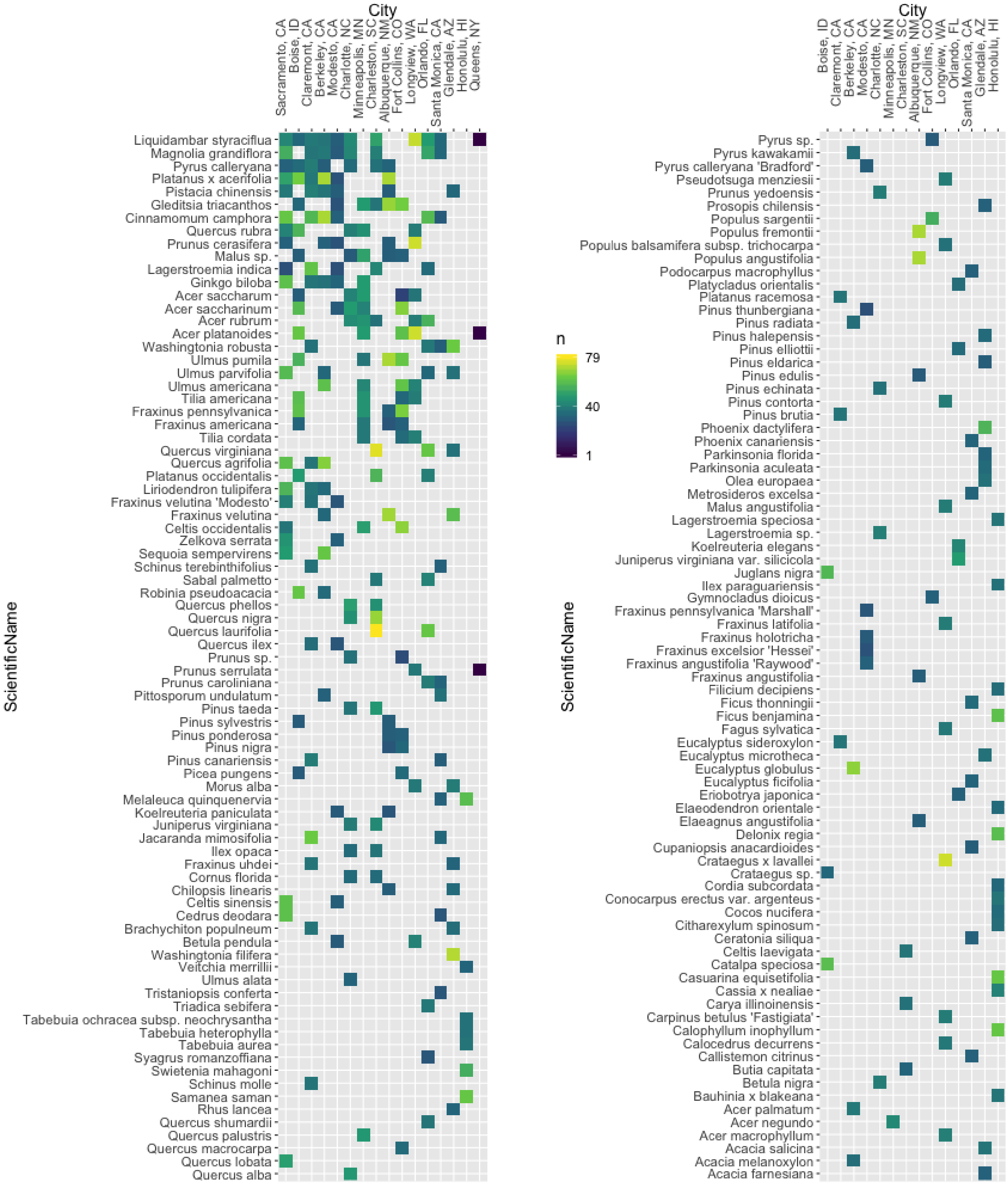


Figure 1: Number of trees sampled of each species and city combination in the urban tree database.

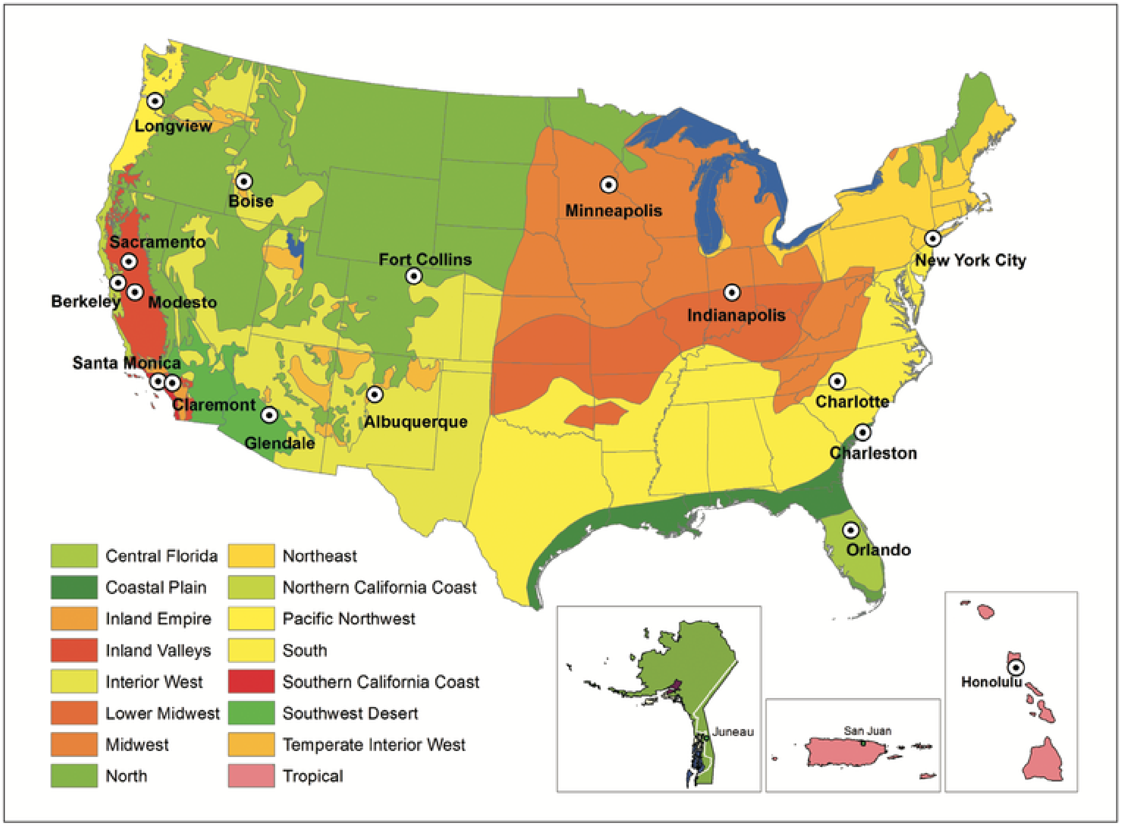


Figure 9—Climate zones were aggregated from 45 Sunset climate zones into 16 zones. Each zone has a reference city where tree growth data were collected. Sacramento, California, was added as a second reference city (with Modesto) to the Inland Valleys zone.

Figure 2: 16 climate regions and 17 representative cities in the UTD (McPherson et al., 2016b).

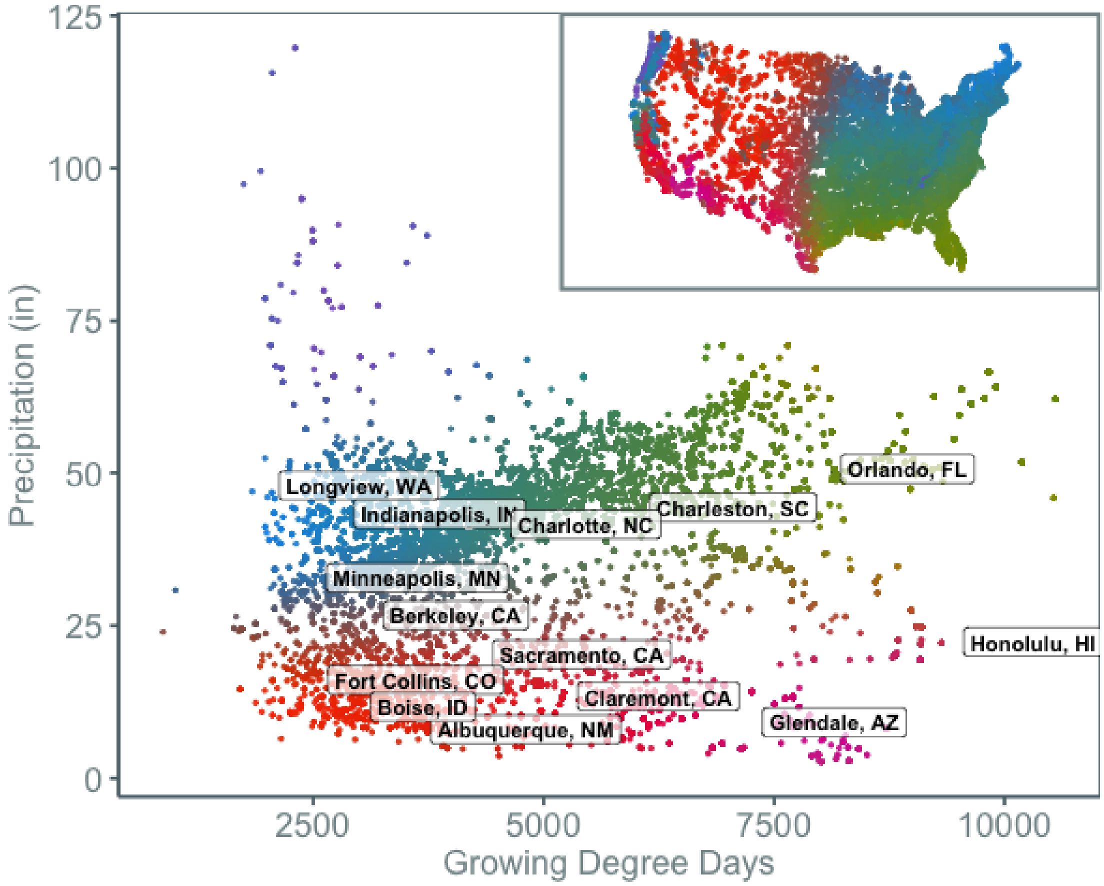


Figure 3: US census tract centroids with UTD reference cities overlaid in growing degree day (GDD) and precipitation climate space and matching color gradient in geographic space. The reference cities cover climate space well, and variation in precipitation and growing degree days is continuous.

increase however slightly because the growth of new wood is essential for proper function. This is different than tree height which often reaches its asympotote. Instead the asympote in our curve could be considered a pragmatic way to constrain diameter growth, or a practical asympote. While trees could theoretically continue to increase in diameter indefinitely, they don't in reality. The asympote represents this practical maximum diameter. An additional feature to the data is that age is time since transplanting. This means trees can have substantial diameter at age 0.

The type of curve that meets these criteria would be an asymmetric sigmoidal curve with an intercept. A modified weibull is such a curve that has worked well in forestry and is the one we use here Weiskittel et al. (2011). However, there are many other curves that meet these criteria and could be used.

Another characteristic of tree growth curves is heterosckedasticity, namely that as the age of a tree increases, so does the variability around the mean. Often past modellers controlled this using log - log transformations Troxel et al. (2013), but we wanted to keep units in their original scale. We tested fitting models where the variance was a linear function or a smoothed spline function of age. However, this still could yield negative predictions at low ages. Instead we adopted the approach of modeling DBH from a gamma distribution, which yielded more realistic posterior predictions.

We fit models of generally increasing complexity starting a single weibull curve for all trees and then varying the curve parameters by city, by genus and species, and by climate. Using approximate leave-one-out cross validation we selected the model with the highest estimated log posterior density Vehtari et al. (2017). The following section details this model, and other models tested are in supplementary materials.

Model

(Notation below. I don't use conventional subscript letters. The nesting of species within genus isn't obvious either until lower levels. I did this because genus could also end up being nested. There is a fair bit going on here, suggestions for how to improve are much appreciated.)

$$y_i \sim Gamma(\mu_i, \alpha_y)$$

$$\mu_i = \beta_{0sc[i]} + \beta_{1sc[i]}(1 - \exp(-\beta_{2sc[i]} x_i^{\beta_{3sc[i]}}))$$

where:

$\mathrm{y_i}$ is the diameter at breast height of tree i and has a gamma distribution with mean, μ_i, and shape, α_y.

β_0 is the intercept, or the diameter of a tree at time of transplanting.

β_1 (plus β_0) is the asymptote of the sigmoidal weibull curve. For most species there are no data near the true asymptote and so this parameter should be considered a highly uncertain estimate of the real maximum dbh of a tree. Pragmatically, it models/causes the slowing of the diameter growth of a tree as it ages.

β_2 and β_3 affect the rate of growth. β_2 provides flexibility to have slow or fast growth at young ages (small x).

All β's must be positive and they are likely correlated with one another, especially β_1 and β_3. Without very old trees that are close to their asympototic dbh, it is harder to separate these two parameters.

for each beta, j = 0,1,2, species, s, and each city, c.

$$\beta_{jsc[i]} = \beta_j + \gamma_{js[i]} + \delta_{jc[i]}$$

for beta$_3$

$$\beta_{3sc[i]} = \beta_3 + \tau_1 * Precip_i + \tau_2 * GDD_i + \tau_3 * (Precip_i * GDD_i) + \gamma_{3s[i]} + \delta_{3c[i]}$$

where β_j is the mean for β coefficient β_j. γ_js is the contribution of genetic (species) effect for species s on β_j. δ_jc is the city effect for city c on β_j.

Species effect:

$$\gamma_{js[i]} \sim N(\gamma_{jg[i]}, \sigma_{js[i]})$$

Genus effect:

$$\gamma_{jg[i]} \sim N(0, \sigma_{jg[i]})$$

for each j. City random effect:

$$\delta_{jc} \sim N(0, \sigma_{jc[i]})$$

Priors:

Priors were selected to be slightly informative and make very biologically unreasonable parameters improbable. The quanitity of data overwhelms the priors, but the relatively narrow priors also helps with sampling.

for gamma $\alpha = \mu^2/\sigma^2$ $\beta = \mu/\sigma^2$

parameter	mu	sd	alpha (of gamma)	beta (of gamma)
β_0	3	1	9	3
β_1	1.75	.3	34.027778	19.444444
β_2	1.25	.15	69.444444	55.555556
β_3 (no climate)	1	.15	44.444444	44.444444
β_3 intercept (climate)	.6	.15	16.	26.666667

$$\beta_0 \sim Gamma(4, 1.33)$$

$$\beta_1 \sim Gamma(34, 19.4)$$

$$\beta_2 \sim Gamma(69.4, 55.5)$$

$$\beta_3 \sim Gamma(16, 26)$$

these priors selected because they create a wide range of possible mean curves, but they are physically possible.

variability by cities and species, something realistic for each parameter

all sigmas: for beta$_0$, the intercept. realistic values for this range from 1 to maybe 10 (that would be a very large average size to plant). Therefore, the sd for δ, $\sigma_{0\ \mathrm{c[i]}}$, is likely less than 1.5. I'll set the prior so that there's a 95% chance it's less than 1.5. Variation between genus/species, the sd for γ, is probably similar.

((10 - 1) / 2) / 3 = 1.5. (take the range of possible values, assume it is ~99.7% range of normal, find the sd by cutting in half and dividing by 3). This is the highest sd I would expect. Set the prior so that I'm saying I think there is a 99.7% chance I think the the sd is less than this. So I make the sd on the prior, 1/3 the highest sd I think is possible.

species is quarter of genus

$$\sigma_{0_g[i]} \sim \text{Normal}(0, .4)$$

$$\sigma_{0_s[i]} \sim \text{Normal}(0, .1)$$

$$\sigma_{0_c[i]} \sim \text{Normal}(0, .3)$$

When fit in stan, β_1 is multiplied by 100, so that it is on a similar order of magnitude as the other parameters (around 1) and the interpretation is the asymptotic diameter in meters (rather than centimeters). Possible values for β_1 might range from .5 to 3 between different genera. Differences between species within genera will likely be less. I expect there to be less difference between cities in the average asymptote.

((3 - .5) / 2) / 3 = .4167 (I'll set genus to slightly less than this because there is also species variation, which I'll set to .1 since I expect most the time species within the same genus to be quite similar, probably within about 1m of one another). I'll set city variation to the same as species.

.4 / 3 = .133 .2 / 3 = .067

for beta$_1$

$$\sigma_{1_{g[i]}} \sim half - Normal(0.1, .4)$$

$$\sigma_{1_{s[i]}} \sim half - Normal(0, .1)$$

for β_2 and β_3 I don't have as good of intuition for β_2 or β_3, but from looking at curves created with different values, I expect their variation to be smaller than for β_0 and β_1. For a start I'll set the variation to about half of what I set it for β_1.

$$\sigma_{2_{g[i]}} \sim half - Normal(0, .1)$$

$$\sigma_{2_{s[i]}} \sim half - Normal(0, .05)$$

$$\sigma_{2_{c[i]}} \sim half - Normal(0, .1)$$

$$\sigma_{3_{g[i]}} \sim half - Normal(0, .1)$$

$$\sigma_{3_{s[i]}} \sim half - Normal(0, .05)$$

$$\sigma_{3_{c[i]}} \sim half - Normal(0, .1)$$

The coefficients for climate variables on β_1 gdd are in thosands. range in data is about 6000 gdd, or 6 thousands gdd. I expect the effects of gdd, precip, and their interaction to be positive. I expect the increase in asympote across this range to be not too big, maybe as little as .1m and as much as 1m. So the coeffcient might be 0.0167 to 0.167. The interaction term will also be positive, but smaller.

precip

$$\tau_1 \sim Normal(.01, .01)$$

gdd

$$\tau_2 \sim Normal(.01, .015)$$

gdd:precip

$$\tau_3 \sim Normal(.005, .005)$$

Results

Model Comparisons

Short descriptions of the models tested and the brms sytax are in table **??**

Note, I removed the scaling of the parameters (multiplying and dividing by 100) in the formula for clarity. In the code they are scaled so that the parameters are on the same order of magnitude and HMC sampling is improved.

Model	Description	brms formula syntax
1	No varying parameters	DBH ∼ b0 + b1 * (1 - exp(-b2 * AGE^{b3}))
		b0 ∼ 1
		b1 ∼ 1
		b2 ∼ 1
		b3 ∼ 1
2	Parameters vary by city	DBH ∼ b0 + b1 * (1 - exp(-b2 * AGE^{b3}))
		b0 ∼ (1 | City)
		b1 ∼ (1 | City)
		b2 ∼ (1 | City)
		b3 ∼ (1 | City)
3	Parameters vary by genus and species	DBH ∼ b0 + b1 * (1 - exp(-b2 * AGE^{b3}))
	Species is nested in genus	b0 ∼ (1 | Genus / Species)
		b1 ∼ (1 | Genus / Species)
		b2 ∼ (1 | Genus / Species)
		b3 ∼ (1 | Genus / Species)
4	Asympotote (β_1) varies by climate	DBH ∼ b0 + b1 * (1 - exp(-b2 * AGE^{b3}))
		b0 ∼ 1
		b1 ∼ gdd * precip
		b2 ∼ 1
		b3 ∼ 1
5	Growth rate (β_3) varies by climate	DBH ∼ b0 + b1 * (1 - exp(-b2 * AGE^{b3}))
		b0 ∼ 1
		b1 ∼ 1
		b2 ∼ 1
		b3 ∼ gdd * precip
6	Parameters vary by city, genus, and species.	DBH ∼ b0 + b1 * (1 - exp(-b2 * AGE^{b3}))
	Growth rate varies by climate.	b0 ∼ (1 | City) + (1 | Genus/Species)
		b1 ∼ (1 | City) + (1 | Genus/Species)
		b2 ∼ (1 | City) + (1 | Genus/Species)
		b3 ∼ precip * gdd + (1 | City) + (1 | Genus/Species)
7	Parameters vary by city, genus, and species	DBH ∼ b0 + b1 * (1 - exp(-b2 * AGE^{b3}))
	(but asympote does not vary by city).	b0 ∼ (1 | City) + (1 | Genus/Species)
	Growth rate varies by climate.	b1 ∼ (1 | Genus/Species)
		b2 ∼ (1 | City) + (1 | Genus/Species)
		b3 ∼ precip * gdd + (1 | City) + (1 | Genus/Species)

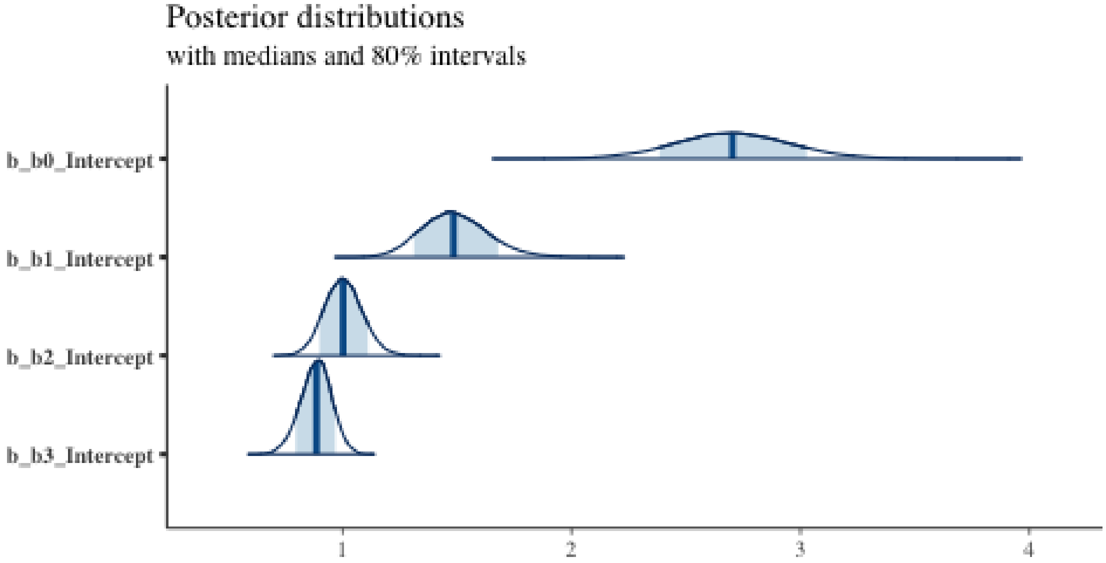
Table 1: $\widehat{elpd}_{\text{loo}}$ is the estimated expected log pointwise predictive density. elpd diff is the difference from the $\widehat{elpd}_{\text{loo}}$ of the top model. se elpd loo is standard error of Vehtari et al. (2017) for descriptions

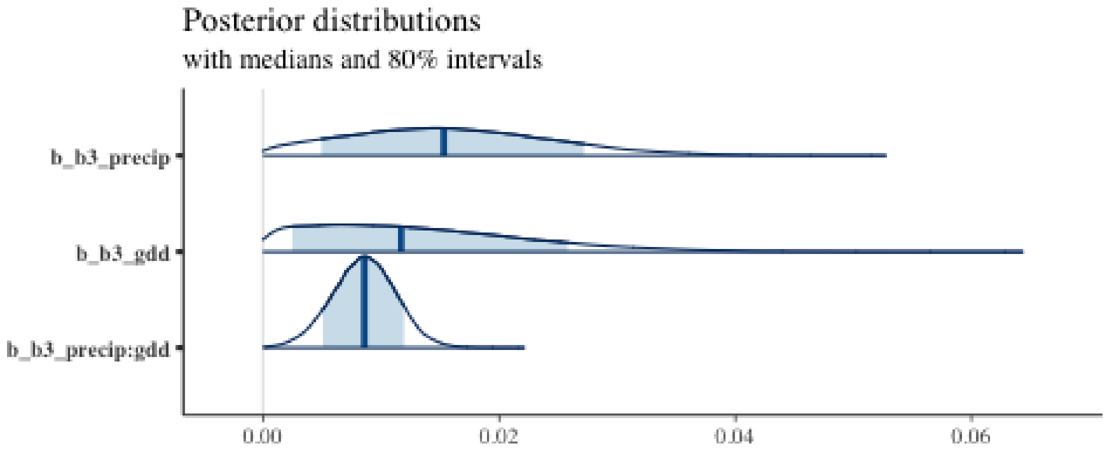
	Model	$\widehat{elpd}_{\text{loo}}$	difference
Best	6	-18845.41	0.00
	7	-18976.38	-130.97
	3	-18989.24	-143.83
	2	-19764.48	-919.06
	5	-20180.41	-1334.99
	4	-20195.21	-1349.80
Worst	1	-20513.12	-1667.70

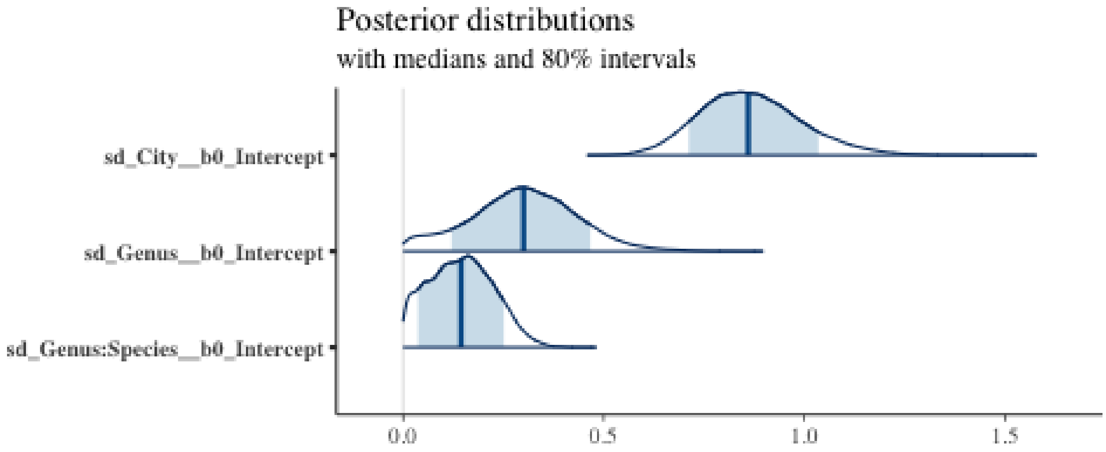
The standard error for the elpd difference of 131 between model 6 and model 7 is 21.4. Therefore, there is strong evidence that model 6 has higher out of sample predictive accuracy than model 7.

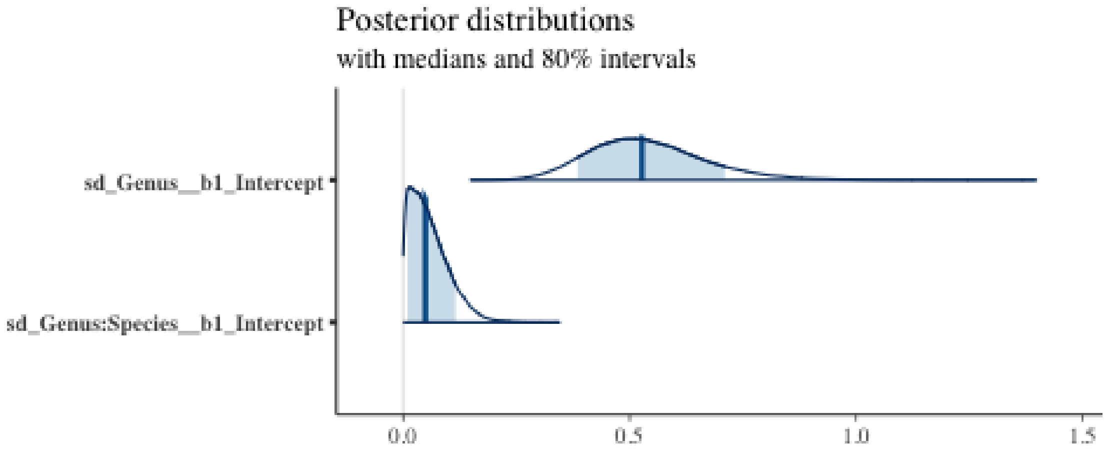
parameter estimates

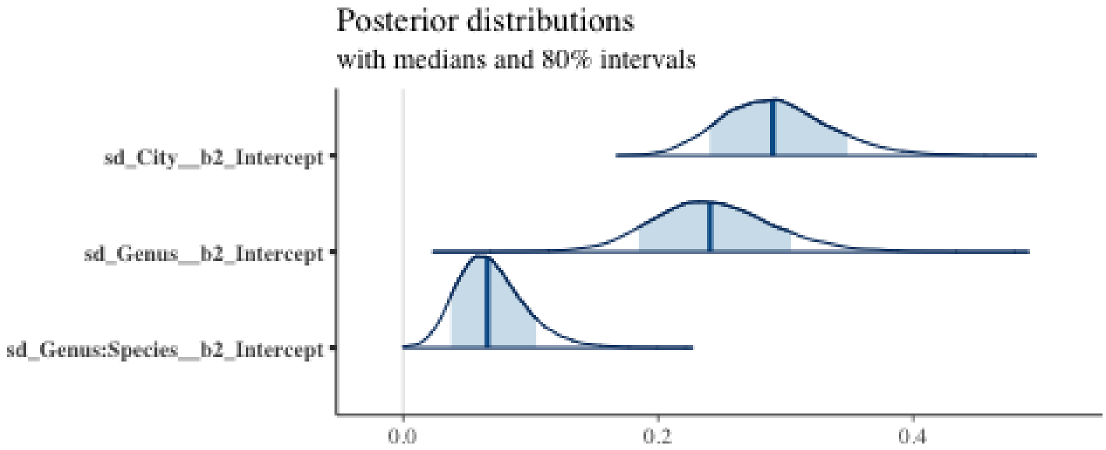
Posterior distributions with 80% interval and median for parameters in

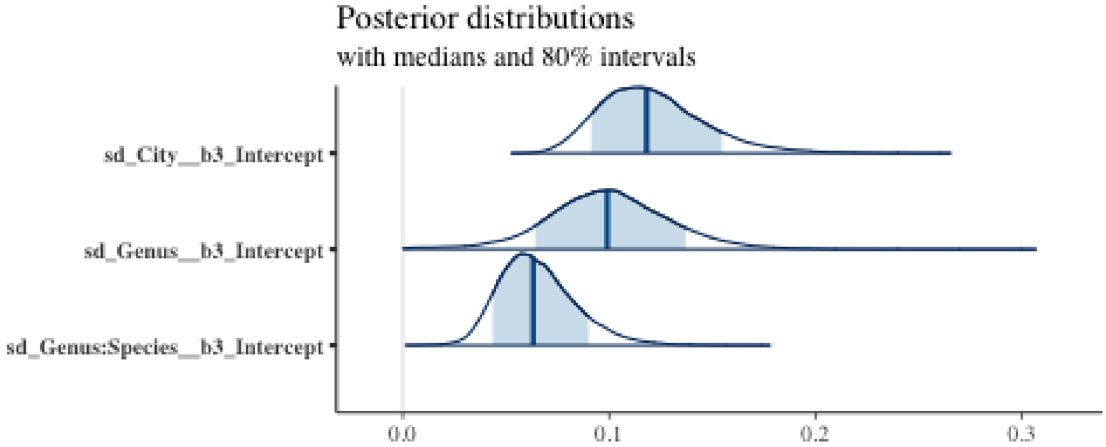


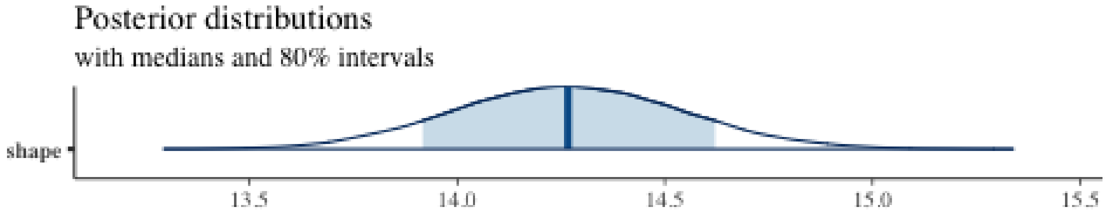

Posterior distributions
with medians and 80% intervals
b_b3_precip
b_b3_gdd
b_b3_precip:gdd
0.00
0.02
0.04
0.06


Posterior distributions
with medians and 80% intervals
sd_City__b0_Intercept
sd_Genus__b0_Intercept
sd_Genus:Species__b0_Intercept
0.0
0.5
1.0
1.5


Posterior distributions
with medians and 80% intervals
sd_Genus__b1_Intercept
sd_Genus:Species__b1_Intercept
0.0
0.5
1.0
1.5

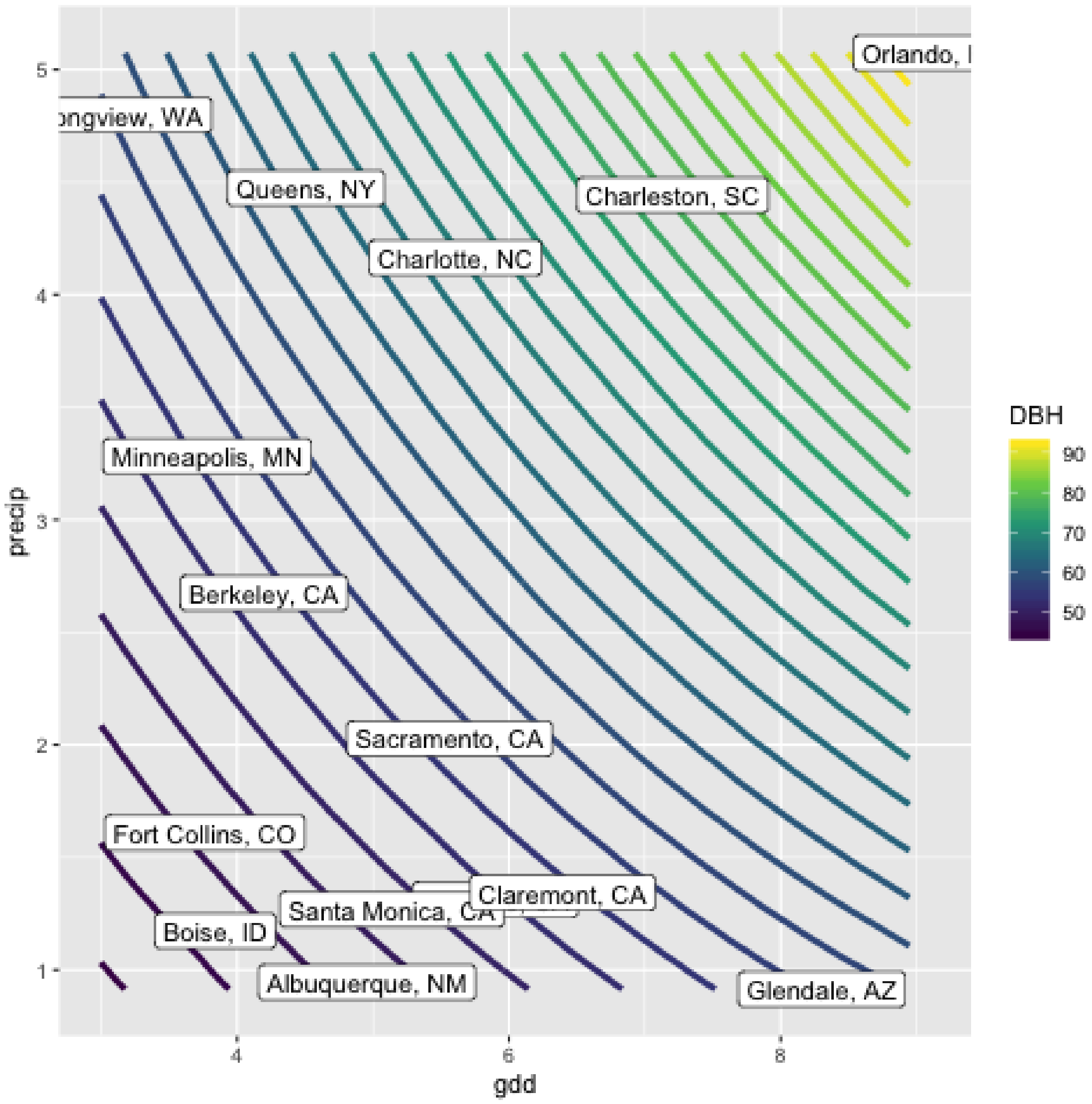






climate effects

There is a positive effect of growing degree days (gdd) and annual precipitation (precip) on tree diameter (dbh), and a postive interaction between the two. Marginal effects of climate on DBH in . There is an estimated 40cm difference in dbh between an average tree in Orlando, FL and one in Boise, ID.


Orlando, I
ongview, WA
Queens, NY
Charleston, SC
Charlotte, NC
Minneapolis, MN
Berkeley, CA
Sacramento, CA
Fort Collins, CO
Claremont, CA
Santa Monica, CA
Boise, ID
Albuquerque, NM
Glendale, AZ
5
4
3
2
1
precip
4
6
8
gdd
DBH
90
80
70
60
50

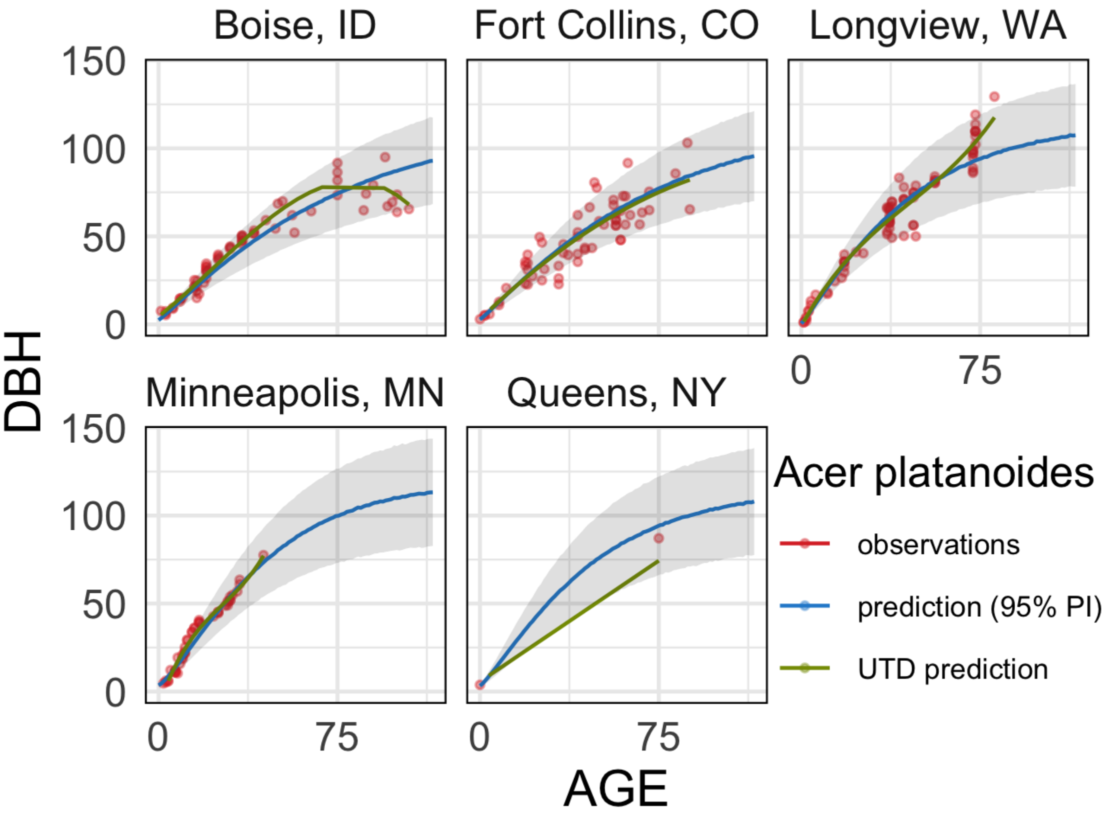
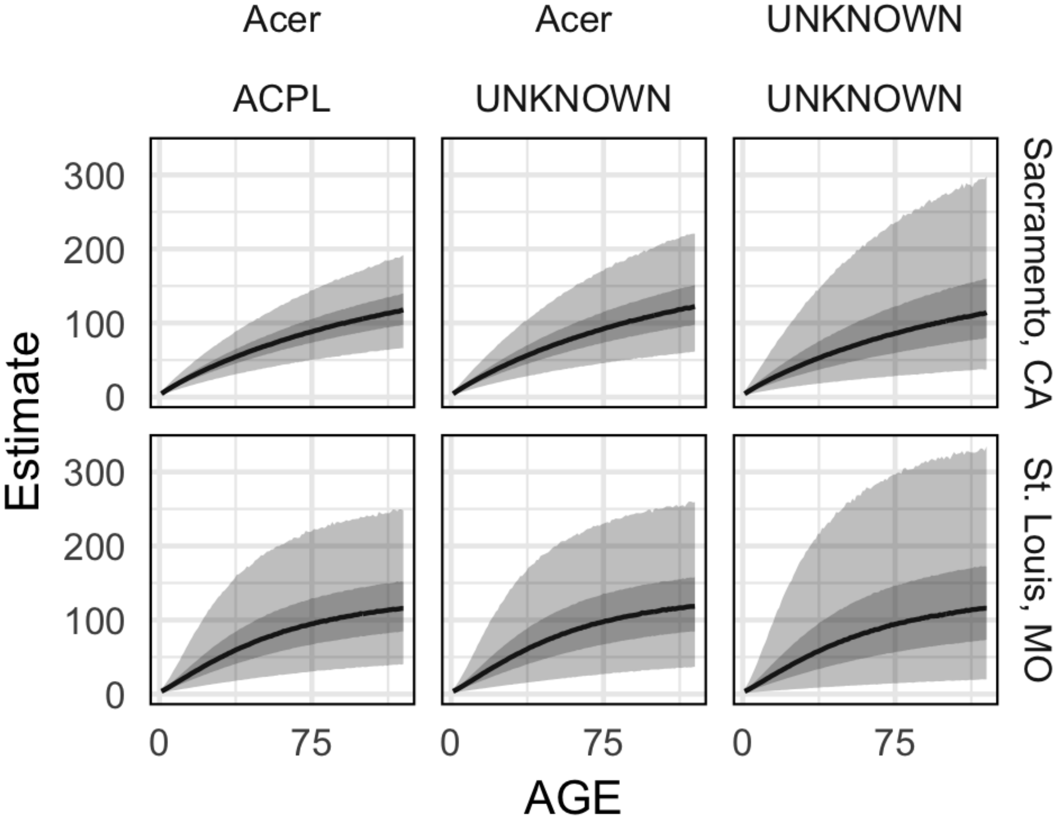


Figure 4: red points are observations, blue lines are predictions of the model and shading indicates 95% prediction interval, green lines show UTD equations

Comparing to existing equations

illustrative comparisons

Uncertainty increases when predicting out of sample cities, genera, and species



Discussoins

An early version of the UTD equations didn't have as much data, but their approach modified parameters based on the number of frost free days Frelich (1992)

Peper et al. (2001) - tested modified weibull following Frelich (1992), but went with logarithm regression model because it had the best in-sample fit. We think weibull would have the best out of sample fit.

Peper et al. (2001) noted that differences in the dimensions of sweetgum and camphor in Modesto and in Santa Monica were due to different pruning regimes, cultural practices. This shows the challenges in modelling. There are some difficult to capture human cultural elements.

several trees

- what is the distribution of maximum age by species? many have very young
- or the distribution of apps max?
- We need to be able to predict to older ages if we want to make realistic predictions.

make better

- more data, duh, perhaps used results to identify where to sample
- more cities, this is important for interpolation across climate space. Could allow for nonlinear relationships and for more variables.
- better climate predictors
- interactions between climate and species.
- use phylogenetic distance, gaussian process, instead of multiple levels of taxonomy.
- extend species with species level predictors (leaf morphology, wood characteristics, shade tolerance, etc).
- smarter priors (e.g. max dbh based on champion trees?) is this possible? I think it would be a very neat extension, but need to think about how these champions are not urban trees most the time. They provide the upper limit on the asymptote, but for urban trees the asymptote could be quite lower.
- incorporate uncertainty in AGE

- There were only 4 trees in Queens NY sampled.
- repeat measures on the same individuals would help much.
- Get more trees in the database, UFIA effort?
- add varying intercept for location of tree in sidewalk/underpowerlines etc.
- There are some funny things with minneapolis data. is it from the same individual for each species?
- challenges of separating β_1 and β_3 without old trees.

References

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):nil.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):nil.

Frelich, L. E. (1992). Predicting dimensional relationships for twin cities shade trees.

McPherson, E. G., van Doorn, N. S., and Peper, P. J. (2016a). Urban tree database.

McPherson, E. G., van Doorn, N. S., and Peper, P. J. (2016b). Urban tree database and allometric equations.

Peper, P. J., McPherson, E. G., and Mori, S. M. (2001). Equations for predicting diameter, height, crown width, and leaf area of san joaquin valley street trees. *Journal of Arboriculture*, 27(6):306–317.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Troxel, B., Piana, M., Ashton, M. S., and Murphy-Dunning, C. (2013). Relationships between bole and crown size for young urban trees in the northeastern usa. *Urban forestry & urban greening*, 12(2):144–153.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Weiskittel, A. R., Hann, D. W., Kershaw, J. A., and Vanclay, J. K. (2011). *Forest Growth and Yield Modeling*. []. John Wiley & Sons, Ltd.