# B53J-2527 - Quantifying Uncertainty in Allometric Models: Fake Forests for the Real World

Craig Wayson[1], Ruth D. Yanai[2], Oswaldo Carrillo[3], Marcela Olguin-Alvarez[4]

*(1) USDA Forest Service Washington DC, Washington, DC, United States, (2) College of Environmental Science and Forestry, Syracuse, NY, United States, (3) National Forestry Commission of Mexico, National Forest Monitoring System, Guadalajara, JA, Mexico, (4) Consultant, Mexico City, Mexico*

## Abstract

Allometric equations are essential for estimating forest biomass, but they are expensive to construct, so estimates are usually based on published equations. The uncertainty associated with choosing an equation is not often reported. To examine the impacts of allometric equation choice on levels of uncertainty, we performed a thought experiment by creating a fake forest with characteristics similar to Mexican forests in terms of diameter distribution and a 'known' biomass with 100,000 individuals. An optimal sample size (n) and distribution of diameters for an allometric equation were calculated. From here, we analyzed the effects of n, distribution of n, range of n, and equation form on the uncertainty estimates of biomass for the forest as a whole. This work tries to illuminate the effects of allometric equation choice on uncertainty for forest inventories, such as that of Mexico, where a decision tree algorithm has been created to select the most appropriate equation from their equation database. This is an effort to reduce overall uncertainty in carbon estimates for structurally heterogeneous and biodiverse forests.

## Introduction

In an effort to improve a decision tree designed to select the most appropriate allometric equation where multiple ones exist for a given species we designed a thought experiment to test the assumptions used. By creating a fake forest of known characteristics we could examine the properties allometric equations used to generate estimates of biomass, specifically their effect on uncertainty in the estimates.

The analysis provides concrete data on allometric equation generation best practices that have been published elsewhere. Often these guidance documents have logical conclusions, but the relative impact of sample design and equation generation criteria are ignored. For example, is sample size more important than sample distribution and by how much? This thought experiment allows us to numerically compare the impacts of each design criteria on overall uncertainty. Analyzing the behavior and statistical characteristics of tree allometric models improves our understanding of them in an effort to strengthen selection criteria.

## Methods

- ***Real-world and simulated forests***

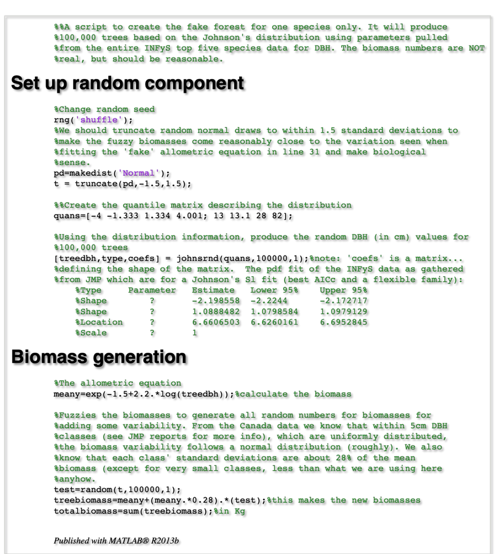


Figure 1. Sample of Matlab code used to create a Fake Forest for analysis. Biomass values were created using three diameter distributions (Figure 3) with random variation added to simulate real-world variability.

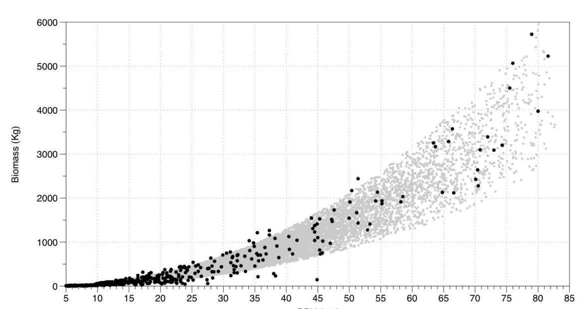

Figure 2. Black dots represent real-world dbh and biomass values for Latin America from http://www.globallometree.org. The gray dots represent values from our Fake Forest.

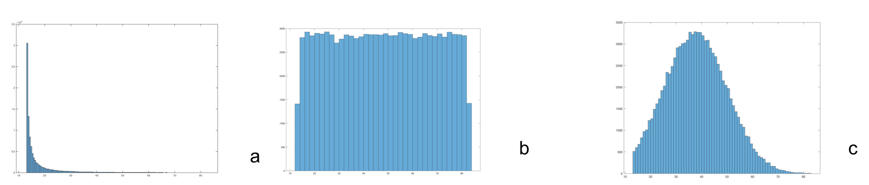


Figure 3. We created three simulated forests, each with 100,000 trees, with contrasting diameter distributions: (**a**) a reverse-J distribution typical of uneven-aged stands, dominated by small stems, described with Johnson's SU distribution fitted to the data from the national forest inventory of Mexico (CONAFOR); (**b**) a uniform distribution of diameters ranging from 13 to 83 cm, the DBH range found in the Mexican forest inventory, with random variation added; and (**c**) a truncated normal distribution with a mean of 37.5 and standard deviation of 12.5 cm, which might be expected of a cohort of trees of a single age.

- ***Simulated samples***

We used 10 sampling intensities from 10 to 1,000 'harvested trees' for use in creating allometric models. We used a total of 64 unique sampling strategies defined used 8 diameter classes (Figure 4). To examine model selection effects we choose 3 common model forms:

$$biomass = e^{a+b(\log(DBH))}$$
$$biomass = e^{a+b(\log(DBH^2))}$$
$$biomass = aDBH^b$$

- ***Costs***

We obtained real-world harvest and processing costs for trees in Mexico as a function of diameter, to look for the most optimal sample size and strategy as part of a cost-benefit analysis. What is most scientifically rigorous may not be practical in the real world.

- ***Simulation runs***

For each sampling intensity (10), sampling scheme (64), and model type (3), we ran 10,000 simulations to ensure random selection effects were removed. This resulted in pulling 57.6 million sets of trees to harvest to create an allometric equation. For each trial, the resulting equation was used to predict the biomass for all trees in the Fake Forest and compared against the known biomass of the forest. Uncertainty values for each combination were recorded and then multiplied by the costs associated with the creation of that particular model.

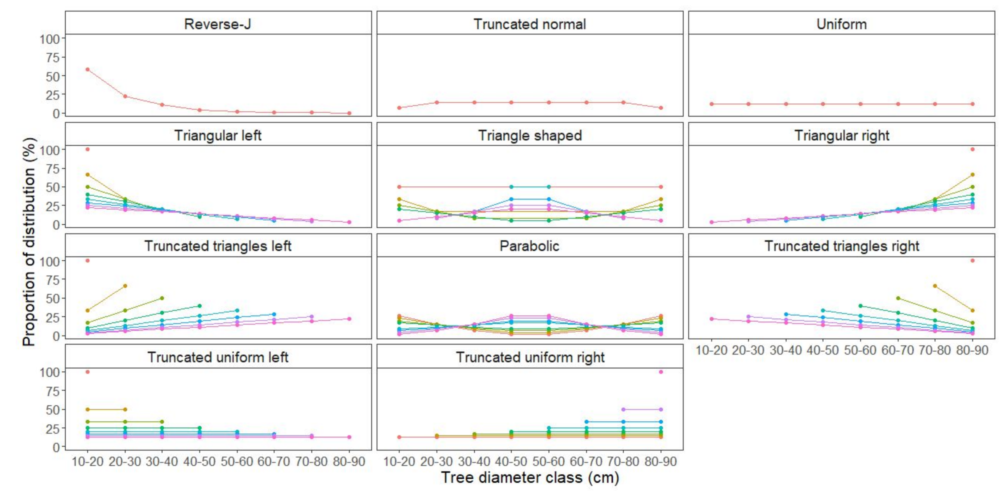


Figure 4 shows the 64 sampling strategies examined. For each sample size, trees were 'harvested' for allometric model creation in the proportions from each DBH class shown above. This resulted in 64 unique sample strategies including 3 (top row) proportional to the distributions of the three underlying fake forests (Figure 3).

## Results

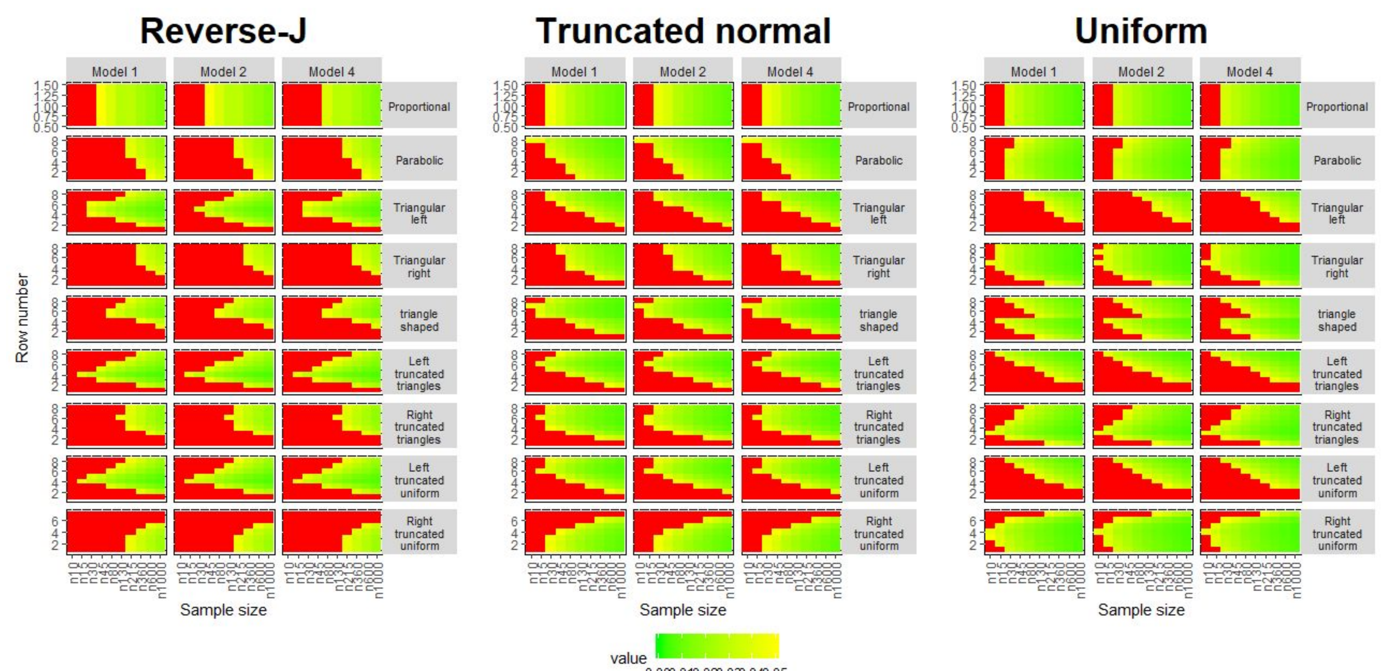


Figure 5. Predicted forest biomass compared to the underlying Fake Forest, with combinations in red having > 5% median absolute error. Combinations of small sample sizes can yield good results if an adequate sampling scheme is chosen. Proportional sampling did not perform as well as some other sampling schemes. The Reverse-J DBH distribution (Figure 3) is the hardest to characterize (Table 1). Model choice is less important than sampling choices.

| Distribution | Model | Proportion < 5% uncertainty | Lowest cost |
|---|---|---|---|
| Normal | 1 | 58% | $5,865 |
| | 3 | 58% | $5,865 |
| | 4 | 57% | $6,200 |
| Uniform | 1 | 63% | $5,780 |
| | 3 | 63% | $6,170 |
| | 4 | 62% | $6,170 |
| Reverse-J | 1 | 37% | $6,200 |
| | 3 | 36% | $6,200 |
| | 4 | 37% | $6,200 |

Table 1. Summary table of model performance and costs. The proportion with <5% uncertainty is the percentage of the 640 combinations that resulted in an equation that predicted the Fake Forest biomass within 5% of the true value. Lowest costs represents the least cost to create a model with <5% uncertainty, which depends on the number of trees in the sample and their size. Larger trees are more expensive to process, and if it is feasible to cut fewer large trees to obtain a good allometric model, cost savings can result.

With 10,000 simulations for each sample size and sampling scheme, the median uncertainty was often < 5% (Figure 5). However, researchers collect only one sample from a population, and that sample could be worse than the median uncertainty. Figure 6 shows 5 examples of equations fit to 15, 30, or 500 trees sampled using from the Reverse-J Fake Forest (gray points) following the best performing sample scheme (Figure 5).

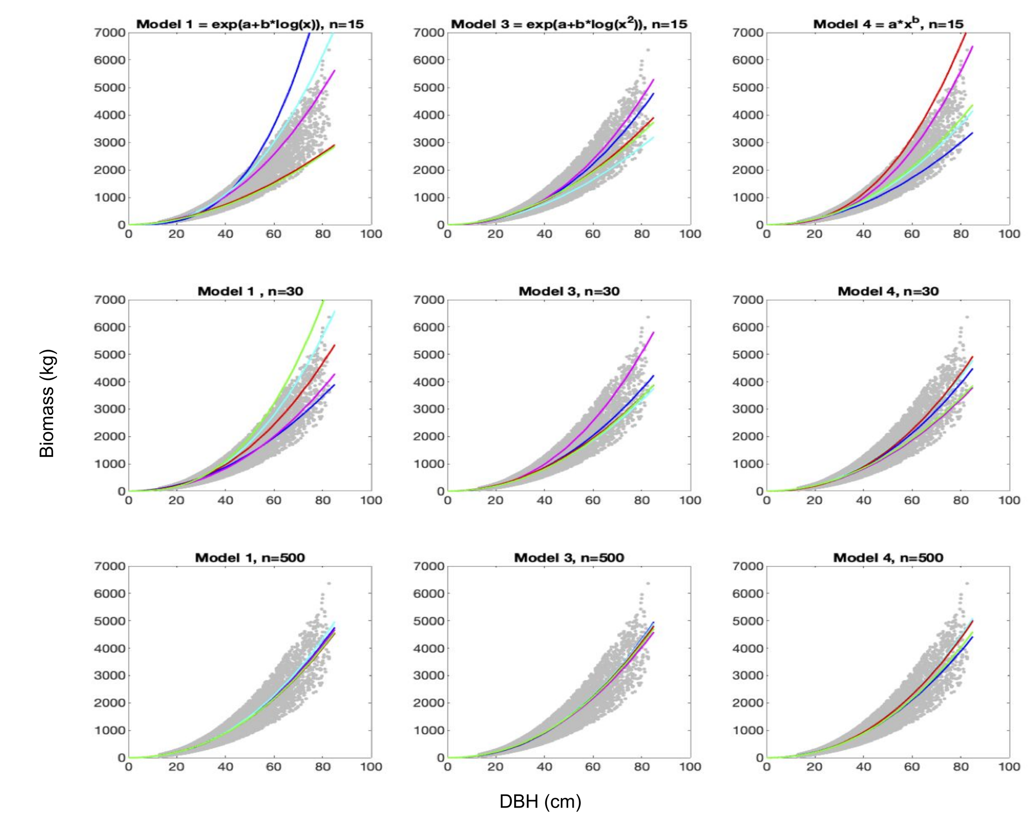


Figure 6. As expected, larger samples better characterize the underlying population (bottom row, *n*=500). Model choice is also important; even with *n*=500, Model 4 shows more variation in model predictions.

## Conclusions

- The sample size and diameter distribution ranges change the type of model that fits the data best.
- Efficient sample schemes vary depending on the underlying forest DBH diameter distributions. With extreme distributions, care must be taken to select the best sample scheme.
- The total uncertainty of predictions decreases when models were developed with larger and better distributed samples in different diameter classes. However, smaller sample sizes and cheaper sample schemes can lead to acceptable levels of uncertainty for the 'real world.'
- Proportional sampling generally performs well, but other, cheaper, sampling schemes can do just as well.
- The preliminary results have already been used to adjust the decision tree used for biomass calculations in the national forest inventory of Mexico.

## Next steps

- Develop a quantitative method to balance costs, model selection, and the probability that one sample will give acceptable results to better guide future allometric model creation.
- Estimate uncertainties at the level of individual predictions for nonlinear models using a Monte Carlo method and a Delta method.
- Graphical analysis of RMSE, $R^2$, uncertainties of predictions and actual uncertainty

## Acknowledgments

Alex Young, SUNY-ESF, helped with graphics. Benedicto Vargas, Instituto Tecnológico de El Salto, provided cost data. QUEST (Quantifying Uncertainty in Ecosystem Studies) is a Research Coordination Network funded by the National Science Foundation. www.quantifyinguncertainty.org