ADDITIVITY AND MAXIMUM LIKELIHOOD ESTIMATION OF NONLINEAR COMPONENT BIOMASS MODELS

David L.R. Affleck[1]

Abstract—Since Parresol's (2001) seminal paper on the subject, it has become common practice to develop nonlinear tree biomass equations so as to ensure compatibility among total and component predictions and to fit equations jointly using multi-step least squares (MSLS) methods. In particular, many researchers have specified total tree biomass models by aggregating the expectations of nonlinear component equations. More recently, an alternative approach has been used wherein compatibility is ensured by specifying a total biomass equation plus one or more component disaggregation functions. Yet calibration of such equations typically has not recognized the additivity of the biomass data or the implied stochastic constraints necessary for development of a valid probability model. For model selection based on information criteria, stochastic simulation, Bayesian inference, or estimation with missing data, it is important to base estimation and inference on probabilistic models. Thus, we show how to specify valid stochastic models for nonlinear biomass equation systems and how to estimate parameters using maximum likelihood (ML). We also explain how ML procedures can accommodate unobserved or aggregated component biomass data. We use Parresol's slash pine data set to contrast model forms and demonstrate Gaussian ML from complete and missing data using open-source software.

Forest inventory programs commonly report estimates of total aboveground biomass and carbon in live trees. The estimates are often obtained from individual tree equations that also furnish estimates for foliage, branches, stems, and other tree components. For many purposes, compatibility among component and total tree biomass models is important. As pointed out by Parresol (2001) this compatibility should, at a minimum, ensure that component biomass or carbon estimates do not exceed estimates of whole-tree biomass or carbon, and that component estimates sum to the estimates of the totals. Yet this level of compatibility is insufficient for analytical and estimation procedures such as mixed-effects modeling, stochastic simulation, and Bayesian inference. These techniques require that the additive nature of tree biomass measurements be recognized and that valid probabilistic models be formulated.

[1] Associate Professor (DLRA), College of Forestry & Conservation, The University of Montana, 32 Campus Drive, Missoula, MT 59812. DLRA is corresponding author: to contact, call (406) 243-4186 or e-mail at david.affleck@umontana.edu

The objectives of this research were to synthesize alternative approaches to nonlinear biomass equation specification within a probabilistic modeling framework, and demonstrate how the models and ML algorithms can be extended to accommodate missing component biomoass observations.

ADDITIVITY OF BIOMASS COMPONENTS

Let $Y_1, Y_2, \ldots, Y_M$ denote M biomass components of a tree and Y_t the total of interest. A fundamental identity is $Y_t = \Sigma_m Y_m$. This identity holds when the symbols represent unobserved tree biomass quantities and is often desired of biomass estimates. Yet it also generally holds when the symbols represent biomass measurements. This is because total tree biomass typically goes unmeasured and is obtained instead by summing the component biomass estimates gathered in the field. From the identity stems the result that, given a set of predictors x, the joint probability law for a set of random variables $Y_1, Y_2, \ldots, Y_M$ and Y_t is

That is, the joint probability model for $Y_1, Y_2, \ldots, Y_M$ and Y_t is a simple multiplicative function of the model for $Y_1, Y_2, \ldots, Y_M$ with the multiplier being independent of any model parameters or predictors. As such, the ML estimators of the parameters governing the former model can be obtained by maximizing only the latter model. Given the biomass components and their additivity there is no additional information in the biomass totals.

SYSTEMS OF NONLINEAR BIOMASS EQUATIONS

Parresol (2001) noted that in the nonlinear setting, additivity of component estimates can be guaranteed only by restricting the total biomass equation to be the sum of the component biomass equations. He advocated direct specification of individual component equations $E(Y_m) = g_m(x_m; \beta_m)$, deriving the equation for the total (and/or subtotals) through aggregation $E(Y_t) = \Sigma_m g_m(x_m; \beta_m)$, and completing a statistical model

$$\begin{aligned} Y_m &= g_m(x_m; \beta_m) + \varepsilon_m \qquad m = 1,2,\ldots,M \\ Y_t &= \Sigma_m g_m(x_m; \beta_m) + \varepsilon_t \end{aligned}$$

allowing for non-constant variance as well as cross-correlations on the error terms $\varepsilon_\cdot$. Parresol further recommended joint estimation of the parameters of this system by MSLS using observations of component and total biomass.

An alternative approach, developed initially in the Chinese literature (Tang et al. 2000; see also Dong et al. 2015), is to specify first an equation for the total, then use multiplicative disaggregation functions to yield component equations. For example,

$$\begin{aligned} Y_t &= g_t(x_t; \beta_t) + \varepsilon_t \\ Y_c &= g_t(x_t; \beta_t)\, g_{cs}(x_{cs}; \beta_{cs}) + \varepsilon_c \\ Y_w &= g_t(x_t; \beta_t)\, [1 - g_{cs}(x_{cs}; \beta_{cs})]\, g_{wb}(x_{wb}; \beta_{wb}) + \varepsilon_w \\ Y_b &= g_t(x_t; \beta_t)\, [1 - g_{cs}(x_{cs}; \beta_{cs})]\, [1 - g_{wb}(x_{wb}; \beta_{wb})] + \varepsilon_b \end{aligned}$$

where Y_c, Y_w, and Y_b denote respectively crown, stem-wood, and stem-bark biomass; and $g_{cs}(\cdot)$ and $g_{wb}(\cdot)$ are functions bounded by 0 and 1 that disaggregate, respectively, the total into crown and stem fractions, and the stem fraction into wood and bark fractions. Tang and Wang (2002) describe a MSLS approach for this system that accounts for non-constant error variance and cross-correlations among errors.

While guaranteeing additivity of estimates, neither of the above systems recognizes the additivity of the data and thus neither represents a valid probability model. It follows that the associated MSLS strategies are not ML procedures. Essentially, $Y_t = \Sigma_m Y_m$ together with $g_t(x_t; \beta_t) = \Sigma_m g_m(x_m; \beta_m)$ implies $\varepsilon_t = \Sigma_m \varepsilon_m$, meaning that the variance function and cross-correlations of ε_t are constrained. The MSLS procedures do not recognize these constraints, so using the observed totals in estimation amounts to specifying an internally inconsistent system and estimators with inscrutable properties. The simplest way to align the above equation systems with probabilistic models is to strike the submodels for biomass totals. With component variance and cross-correlation structures otherwise preserved the reduced systems are valid probability models; parameters can be estimated by ML and information criteria such as AIC can be used in model selection. The models can then also be extended to accommodate mixed-effects or Bayesian specifications, or used for stochastic simulation.

MISSING BIOMASS DATA

A further advantage of specifying a valid probabilistic component biomass model is that missing data patterns can be accommodated. In particular, if the missingness mechanism is uninformative (Little & Rubin 1987, ch.1) then the complete data likelihood for impacted trees can be integrated to yield an observed data likelihood for ML estimation.

With component biomass data, there are two important forms of missingness. The first results when individual components (and thus biomass totals) go unobserved. For example, if crown material is lost and only stem biomass components are observed then the likelihood for the tree in question reduces to

A second form of missingness is when all components are represented (and thus the total is computable), but some are known only in aggregate. For example, if crown biomass and overall stem biomass (but not bark and wood biomass) are available for a certain tree, then its contribution to the overall likelihood is

Both integrals are complex in the general case, but easily obtained for Gaussian models.

CASE STUDY: PARRESOL'S (2001) SLASH PINE DATA

Parresol (2001) presented biomass data for 40 slash pine (*Pinus elliottii*) grown in Louisiana, USA. Table 1 reproduces the data and identifies observations masked from missing data analyses.

To the complete and simulated-incomplete data were fit Parresol's (2001) aggregative system of equations with variance functions

$$\mathrm{var}(Y_m) = \theta_{m1}(\text{d.b.h.})$$

and unstructured cross-correlations between crown, bark, and wood biomass observations from the same tree. With the same variance and cross-correlation structures, a disaggregation-based system of equations was also fit with

$$g_t(x_t; \beta_t) = \exp[\, \beta_{t1} + \beta_{t2} \ln(\text{d.b.h.}) + \beta_{t2} \ln(h)\,]$$

$$g_{cs}(x_{cs}; \beta_{cs}) = (1 + \exp[\, \beta_{cs1} + \beta_{cs2} \ln(\text{d.b.h.}) + \beta_{cs2} \ln(h)\,])^{-1}$$

$$g_{wb}(x_{wb}; \beta_{wb}) = (1 + \exp[\, \beta_{wb1} + \beta_{wb2} \ln(\text{d.b.h.}) + \beta_{wb2} \ln(h)\,])^{-1}$$

where h is tree height.

The models were fit by Gaussian ML using the *gnls* function of the *nlme* package (Pinheiro et al. 2014) in R (R Core Team 2014). To do so, a symmetric covariance structure was coded from the basic *gnls* correlation structure; this allowed for reduction of the covariance matrix to account for missing or aggregated components.

Model fit statistics are in Table 2. The systems provide comparable predictive models, and the patterns of missingness shown in Table 1 only slightly degrade model-data agreement. The models are also similar to the two-stage and three-stage least squares models presented in Parresol (2001) but, as they constitute valid probability models, they can be extended to include random effects (e.g., to express dependence among trees within plots) or prior information on parameters (using Bayesian techniques).

Table 1—Slash pine biomass data from Parresol (2001) with tree diameter at breast height (d.b.h.) and total height (h). Shaded cells identify data masked from some analyses; within these cells component values printed in white are assumed missing (i.e. unknown along with the tree total) while those printed in black are assumed known only in aggregate (i.e. stem mass known but not wood or bark separately).

	d.b.h.	height	Green mass (kg)			
Tree	(cm)	(m)	Wood	Bark	Crown	Total
1	5.6	7.9	6.5	2.3	1.0	9.8
2	6.4	8.5	7.4	2.6	2.1	12.1
3	8.1	10.7	17.6	4.5	2.3	24.4
4	8.4	11.3	18.5	4.3	4.2	27.0
5	9.1	11.0	22.6	5.4	5.6	33.6
6	9.9	13.1	30.6	7.4	5.5	43.5
7	10.4	14.3	32.9	6.7	6.4	46.0
8	11.2	14.6	40.6	9.3	6.2	56.1
9	11.7	14.3	46.0	10.7	7.7	64.4
10	12.2	14.9	51.6	13.1	6.1	70.8
11	11.9	16.8	60.4	10.1	5.4	75.9
12	13.2	13.7	62.8	15.2	10.7	88.7
13	12.2	15.8	67.5	12.9	15.3	95.7
14	13.7	18.0	81.2	12.5	8.7	102.4
15	14.2	16.5	94.3	18.2	11.2	123.7
16	15.0	20.1	123.4	16.5	7.7	147.6
17	15.7	16.8	107.3	21.5	19.7	148.5
18	16.5	17.1	123.8	22.1	28.9	174.8
19	16.5	17.1	151.6	24.6	16.8	193.0
20	19.6	13.7	140.4	25.1	46.2	211.7
21	17.5	19.2	170.4	27.4	16.8	214.6
22	17.8	18.3	169.6	31.7	24.0	225.3
23	18.5	17.7	160.3	36.9	47.5	244.7
24	19.6	19.8	199.8	38.7	19.7	258.2
25	18.5	22.9	231.6	29.6	24.6	285.8
26	19.8	18.6	217.9	33.9	45.8	297.6
27	20.6	17.4	216.0	32.6	61.2	309.8
28	21.6	17.7	200.6	40.2	75.4	316.2
29	19.8	18.9	217.5	38.5	62.0	318.0
30	22.9	19.8	314.8	43.1	43.2	401.1
31	23.6	18.3	287.1	63.4	51.7	402.2
32	23.1	18.9	290.9	44.3	76.7	411.9
33	24.1	21.3	320.1	50.6	75.6	446.3
34	26.4	19.2	308.6	65.7	116.0	490.3
35	24.6	25.0	403.0	49.8	69.8	522.6
36	25.1	19.8	390.4	48.8	83.5	522.7
37	29.0	20.4	445.2	60.4	88.0	593.6
38	28.4	26.8	736.4	84.0	79.9	900.3
39	31.8	27.4	770.9	93.8	170.2	1034.9
40	33.0	27.7	921.3	108.0	169.2	1198.5

Table 2—Component root mean squared error (RMSE) and corrected Akaike's information criterion (AICc) for alternative models, parameter estimation routines, and data.

			RMSE[a] (kg)				
Model	**Estimation**	**Data**	**Crown**	**Bark**	**Wood**	**Total**	**AICc[b]**
Aggregative	ML	complete	13.3	5.1	26.9	30.4	784.6
	ML	incomplete	13.2	5.2	26.9	30.7	786.4
	2SLS[c]	complete	13.9	5.0	26.7	31.4	795.6
	3SLS[c]	complete	12.8	5.0	26.7	29.8	808.5
Disaggregation	ML	complete	13.4	4.8	26.2	29.6	789.1
	ML	incomplete	13.4	5.4	26.2	30.3	792.8

[a] Based on the complete data without weight functions or degrees of freedom adjustments.
[b] Based on the complete data and a joint Gaussian model for crown, bark, and wood biomass.
[c] From coefficients published in Parresol (2001).

ACKNOWLEDGMENT

This research was supported by the Inland Northwest Growth & Yield Cooperative and a joint venture agreement with the USDA Forest Service (11-JV-11242305-120).

LITERATURE CITED

Dong, L.; Zhang, L.; Li, F. 2015. A three-step proportional weighting system of nonlinear biomass equation. Forest Science. 61: 35-45.

Little, R.J.A.; Rubin, D.B. 1987. Statistical Analysis with Missing Data. New York: Wiley.

Parresol, B.R. 2001. Additivity of nonlinear biomass equations. Canadian Journal of Forest Research. 31: 865-878.

Pinheiro, J.C.; Bates, D.W.; DebRoy, S.; Sarkar, D.; R Core Team. 2014. nlme: Linear and Nonlinear Mixed Effects Models.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Tang, S.; Wang, Y. 2002. A parameter estimation program for the error-in-variable model. Ecological Modelling. 156: 225-236.

Tang, S.; Zhang, H.; Xu, H. 2000. Study on establish and estimate method of compatible biomass model. Scientia Silvae Sinicae. 36(S1): 19-27 (in Chinese with English abstract).