Canadian Science Publishing

ARTICLE

Additional biomass estimation alternatives: nonlinear two- and three-stage least squares and full information maximum likelihood for slash pine

Dehai Zhao, Thomas B. Lynch, James A. Westfall, and John W. Coulston

Abstract: A system of nonlinear biomass component equations was developed for slash pine (*Pinus elliottii* Engelm. var. *elliottii*) trees using an econometric approach in which endogenous right-hand-side variables were included in some equations. The system was fitted to component biomass data from 306 slash pine trees sampled in the southeastern United States with weighted two-stage (2SLS) and three-stage (3SLS) least squares and full information maximum-likelihood (FIML) estimation methods. The predictive performances of the system fitted with these three estimation methods were ranked based on an array of statistics, and the ranking follows the order of FIML > 3SLS > 2SLS. The new system performed as well or better than previously published biomass equation systems developed using the aggregation and disaggregation approaches and fitted to the same data. The results demonstrated that the econometric approaches such as FIML and 3SLS have the potential to be useful for tree biomass modeling.

Key words: southern pine, systems of equations, econometrics, carbon sequestration.

Résumé : Un système d'équations de composantes de la biomasse non linéaire a été développé pour les pins d'Elliot (*Pinus elliottii* Engelm. var. *elliottii*) en utilisant une approche économétrique dans laquelle les variables endogènes du côté droit ont été incluses dans certaines équations. Le système a été ajusté aux données des composantes de la biomasse de 306 pins d'Elliot échantillonnés dans le sud-est des États-Unis avec des moindres carrés pondérés en deux étapes (2SLS) et en trois étapes (3SLS) et des méthodes d'estimation du maximum de vraisemblance à informations complètes (FIML). Les performances prédictives du système ajusté avec ces trois méthodes d'estimation ont été classées en fonction d'un éventail de statistiques et le rang suit l'ordre du FIML > 3SLS > 2SLS. Le nouveau système a performé aussi bien ou mieux que les systèmes d'équations de la biomasse publiés antérieurement qui avaient été développés en utilisant les approches d'agrégation et de désagrégation et ajustés aux mêmes données. Les résultats ont démontré que les approches économétriques telles que le FIML et le 3SLS ont le potentiel d'être utiles pour la modélisation de la biomasse des arbres. [Traduit par la Rédaction]

Mots-clés : pin du sud, systèmes d'équations, économétrie, séquestration du carbone.

Introduction

Models for estimating the components of individual tree dry biomass generally include predictors for bole wood, bole bark, branches, and foliage. Several approaches have been used to develop prediction models for total tree biomass and its components. Among them aggregative and disaggregative modeling strategies are commonly used to develop additive biomass equations, in which the predictions for the components sum to the prediction from a total tree biomass equation. The aggregative strategy is to develop biomass prediction models for each biomass component and obtain the total biomass by adding the components. Parresol (2001) proposed simultaneous estimation of a system of biomass component equations and the aggregated total biomass equation using a weighted nonlinear seemingly unrelated regressions method (SUR, referred to as SUR1). The SUR1 approach has been used by several others including Sabatia et al. (2008) for shortleaf pine (*Pinus echinata* Mill.) biomass components, Zhao et al. (2015) for loblolly pine (*Pinus taeda* L.), and Zhao et al. (2019) for slash pine (*Pinus elliottii* Engelm. var. *elliottii*) biomass components. Affleck and Diéguez-Aranda (2016) proposed to jointly fit only biomass component equations using maximum-likelihood (ML) estimation, arguing from the standpoint of how biomass data are collected that it was more appropriate to simply add predictions of biomass components to obtain total biomass. Zhao et al. (2019) used SUR to fit biomass component equations only (referred to as SUR2), and they demonstrated both analytically and empirically that the SUR2 should be more reasonable for estimating the aggregative models than the SUR1. The SUR method accounts for the fact that the residual errors for models in the system may be correlated (Judge et al. 1985, p. 468; Zellner 1962).

The disaggregative strategy is the "component proportion" approach in which total tree biomass is disaggregated into tree components based on their estimated proportions (Tang et al. 2000; Jenkins et al. 2003; Zhao et al. 2019). Tang et al. (2000) developed a disaggregation approach in which a total biomass model is first developed,

Received 30 November 2021. Accepted 8 February 2022.

D. Zhao. Warnell School of Forestry and Natural Resources, The University of Georgia, Athens, GA 30602, USA.
T.B. Lynch. Department of Natural Resource Ecology and Management, Oklahoma State University, Stillwater, OK 74078, USA.
J.A. Westfall. US Forest Service, Northern Research Station, 3460 Industrial Drive, York, PA 17402, USA.
J.W. Coulston. US Forest Service, Southern Research Station, 1710 Research Center Drive, Blacksburg, VA 24060, USA.
Corresponding authors: Dehai Zhao (email: zhaod@uga.edu) and Thomas B. Lynch (email: tom.lynch@okstate.edu).

and the biomass component models are defined and then the component proportions are derived from these biomass component models and total biomass model (also see Dong et al. 2015). The estimated value of a dependent (endogenous) variable — total biomass — is used as explanatory variables to solve the parameters of biomass component equations with two-stage nonlinear error-in-variable models (TSEM) (Tang et al. 2001). Affleck and Diéguez-Aranda (2016) used a different option to disaggregate the total biomass, i.e., specifying models for the total and discrimination functions between components (component proportions) but then fitting the system using component biomass observations on the mass scale via ML or SUR. Unlike the approaches of Tang et al. (2000) and Affleck and Diéguez-Aranda (2016), Zhao et al. (2019) proposed another disaggregation approach in which biomass component proportions are directly modeled using the Dirichlet regression model (DRM) (Zhao et al. 2016) and the total biomass model is separately developed. A potential problem with this approach is that biases can occur when predictions from a model fitted by ordinary least squares (OLS) are used to obtain predictions in a second model fitted independently by OLS. A formula that could be used to quantify this bias is derived in Appendix A.

Except for systems of biomass equations developed by the approach of Tang et al. (2000), the right-hand sides of additive biomass equations developed recently include the independent (exogenous) variables only. Some biomass components may be highly related to each other. Taking advantage of such relationships provides an alternative, that is, using the dependent (endogenous) variable from one component model equation as an explanatory variable in another component model equation. For example, we may expect foliage to be highly related to branch biomass and bark biomass to be related to stem wood biomass. So, we may use branch biomass, typically the dependent (endogenous) variable in branch component equation, as an explanatory variable to predict foliage biomass in the foliage component equation. Similarly, we may use stem wood biomass as an explanatory variable to predict bark biomass even though stem wood biomass is the dependent (endogenous) variable in another component prediction equation. This approach leads to simultaneous equations in which at least one of the equations contains endogenous right-hand-side variable(s), thus, yielding biased and inconsistent parameter estimates upon parametrization via SUR. Alternatively, however, two-stage (2SLS) and three-stage (3SLS) least squares (e.g., Judge et al. 1985, pp. 597–600) and full information maximum likelihood (FIML) (Judge et al. 1985, p. 601; Theil 1971, pp. 524–526) are econometric methods that can be used to estimate multiple equations with endogenous right-hand-side variables.

2SLS is applied to a system of model equations by regressing each endogenous variable on all exogenous variables in the first-stage regression model, and then replacing the original values of the endogenous right-hand-side variables in the second-stage regression model with the predicted values from the first stage and fitting it with OLS. This single equation method leads to consistent estimators but generally has the disadvantage of being asymptotically inefficient. 3SLS results when the estimated variances and covariances of the residuals from the 2SLS are used to re-estimate model parameters using generalized least squares in the third stage. 3SLS can be more efficient than 2SLS when the cross-equation covariation (i.e., cross-equation error correlation) is large (Belsley 1988). 2SLS and 3SLS procedures have been used by several authors in the forest biometrics literature (e.g., Furnival and Wilson 1971; Murphy and Sternitzke 1979; Borders 1986; Lynch and Clutter 1998). FIML, another full system method, uses ML estimation with the multivariate distribution of errors for the system of model equations. The multivariate error distribution used in FIML has usually been the multivariate normal distribution. Rothenberg and Leenders (1964) have demonstrated that FIML and 3SLS have the same asymptotic limiting covariance matrix (Theil 1971, p. 526).

2SLS, 3SLS, and FIML could conceivably be applied to the problem of component proportion estimation, especially for biomass equations developed by the approach of Tang et al. (2000). However, parameterization software for TSEM is not currently available. Conversely, 2SLS, 3SLS, and FIML could be easily performed using the SAS/ETS® MODEL Procedure (SAS Institute Inc. 2011).

The objective of this study was to develop and compare biomass prediction models for estimating components directly using 2SLS, 3SLS, and FIML for slash pine trees. Their predictive performances were evaluated and compared with other model systems developed by Zhao et al. (2019) using SUR and DRM approaches on the same slash pine dataset.

Materials and methods

Tree biomass data

Because of our desire to compare the new biomass estimation methods proposed here with previous methods, we intended to use the same slash pine plantation data as Zhao et al. (2019). The data consisted of two sets: one from destructive biomass sampling of 96 slash pine trees conducted in 2016 and a second set of 210 slash pine trees from a legacy biomass database available for download at legacytreedata.org assembled by Radtke et al. (2015). Combining the two data sources provided 306 individual slash pine tree observations from plantations on the coastal plain of Georgia and north Florida. Each tree had observations for wood, bark, branch, and foliage oven-dry biomass values, which when summed provided total biomass. Each tree also had diameter at breast height (DBH) and tree total height (HT) measurements. Summary statistics for these data are given in table 1 of Zhao et al. (2019), which indicates that DBH ranged from 3 to 53.3 cm with a mean of 18.4 cm, HT ranged from 2.9 to 30.2 m with a mean of 18.4 m, and total biomass ranged from 0.98 to 1861.9 kg with a mean of 201.83 kg.

Stem wood, stem bark, branch, foliage and tree total aboveground biomass of all trees and their relationships with tree DBH and HT are shown in figure 1 of Zhao et al. (2019). Figure 1 presented here illustrates relationships between the natural logarithms of DBH and HT and the natural logarithms of the wood, bark, branch, and foliage biomass components. Nearly linear relationships between the logarithms of DBH and HT with biomass components indicates that power functions of DBH and HT may be effective starting points for modeling the component relationships. A close linear relationship between the logarithms of wood biomass and bark biomass (Fig. 1) indicates that power function relationships between the two variables should be appropriate. Similarly, a linear trend between the logarithms of branch and foliage biomass supports the use of a power function relationship between these variables.

Model development and estimation

Based on the relationships mentioned above, the following biomass component equations with endogenous right-hand-side variables were developed for slash pines:

$$(1) \qquad y_1 = b_{11}\mathrm{DBH}^{b_{12}}\mathrm{HT}^{b_{13}} + \varepsilon_1$$

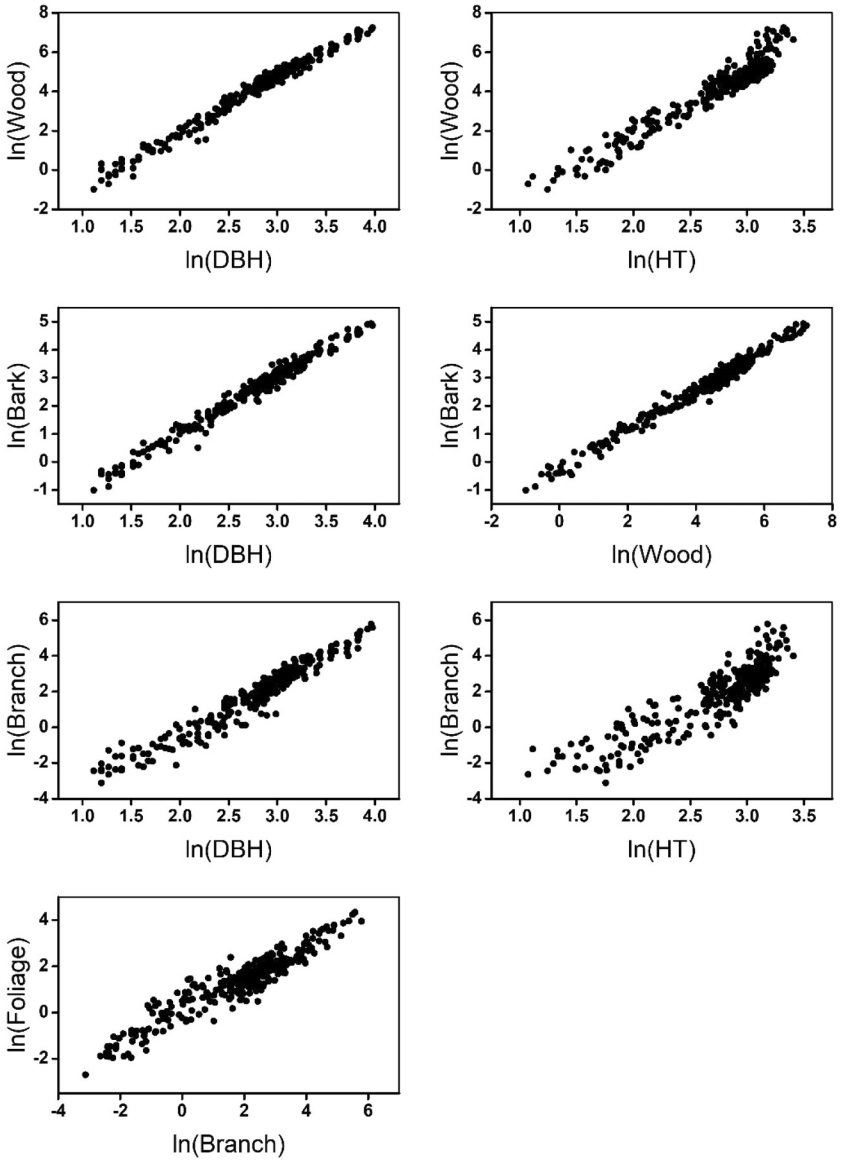
$$(2) \qquad y_2 = b_{21}(y_1)^{b_{22}}\mathrm{DBH}^{b_{23}} + \varepsilon_2$$

$$(3) \qquad y_3 = b_{31}\mathrm{DBH}^{b_{32}}\mathrm{HT}^{b_{33}} + \varepsilon_3$$

$$(4) \qquad y_4 = b_{41}(y_3)^{b_{42}} + \varepsilon_4$$

where y_1 is wood biomass (kg), y_2 is bark biomass (kg), y_3 is branch biomass (kg), y_4 is foliage biomass (kg), b_{ij} (i = 1, 2, 3, 4; j = 1, 2, 3) are unknown parameters, ε_i (i = 1, 2, 3, 4) are random equation errors, DBH is diameter at breast height (cm), and HT is tree total height (m).

Fig. 1. Relationships among the natural logarithms of DBH (ln(DBH)), total height (ln(HT)), stem wood biomass (ln(Wood)), bark biomass (ln(Bark)), branch biomass (ln(Branch)), and foliage biomass (ln(Foliage)) for slash pine.



y_1, y_2, y_3, and y_4 would be termed endogenous variables in econometrics because they are determined within the model system while DBH and HT would be considered exogenous variables because they are not determined within the system of model equations (Theil 1971, pp. 430–431). The term "explanatory variables" is often used for variables on the right-hand side of econometric prediction models because models used to predict endogenous variables may be composed of exogenous variables as well endogenous variables from other system model equations. Stem wood and branch component equations (Models 1 and 3) were modeled using power functions of exogenous variables DBH and HT. This is a nonlinear form of volume equation originally proposed by Schumacher and Hall (1933) and has been used successfully in many tree content modeling efforts including biomass component models (e.g., Clutter et al. 1983, p. 8; Zhao and Kane 2017; Zhao et al. 2019). In Model 2 the endogenous variable bark biomass y_2 is a function of the endogenous variable — wood biomass y_1 — and the exogenous variable DBH. Foliage component equation (Model 4) has only the endogenous variable — branch biomass y_3 — as explanatory variable. The variable DBH was added to the bark prediction model equation because the amount of bark needed to cover a stem having a given wood content can vary systematically due to factors such as bark thickness and stem form that are related to individual tree DBH. Econometric methods such as 2SLS, 3SLS, and FIML allow both endogenous and exogenous variables to be used as explanatory variables as in Models 2 and 4 above.

This model system follows the perspective of Affleck and Diéguez-Aranda (2016) and the findings of Zhao et al. (2019) in that total biomass is obtained by adding predictions from all the components but there is no model in the system to fit directly to total biomass as was the case with Parresol (2001). The system of biomass component equations (Models 1–4) was fitted to the biomass data of 306 slash pine trees using the weighted 2SLS, 3SLS, and FIML, respectively, by using the SAS/ETS® MODEL Procedure (SAS Institute Inc. 2011). This includes three-fitting steps: (1) fitting the system using PROC MODEL with 2SLS, 3SLS, or FIML, without taking into account the heteroscedasticity problem in model residuals; (2) squaring the estimated residuals of the unweighted model from the first step as the dependent variable and then fitting it as a function of DBH and HT on the natural log scale to determine the

Table 1. Parameter estimates and their standard errors (SE) and p values for the system of biomass component model equations fitted with three-stage least squares (3SLS), two-stage least squares (2SLS), and full information maximum likelihood (FIML).

	3SLS			FIML			2SLS		
Parameter	Estimate	SE	p value	Estimate	SE	p value	Estimate	SE	p value
b_{11}	0.0125	0.0006	<0.0001	0.0115	0.0008	<0.0001	0.0124	0.0007	<0.0001
b_{12}	2.0956	0.0344	<0.0001	2.0662	0.0318	<0.0001	2.0930	0.0338	<0.0001
b_{13}	0.9881	0.0425	<0.0001	1.0469	0.0477	<0.0001	0.9935	0.0426	<0.0001
b_{21}	0.1488	0.0303	<0.0001	0.1507	0.0289	<0.0001	0.1395	0.0281	<0.0001
b_{22}	0.2861	0.0522	<0.0001	0.2923	0.0529	<0.0001	0.2694	0.0518	<0.0001
b_{23}	1.2013	0.1503	<0.0001	1.1873	0.1467	<0.0001	1.2498	0.1486	<0.0001
b_{31}	1.903E-3	3.35E-4	<0.0001	2.334E-3	3.54E-4	<0.0001	1.649E-3	2.24E-4	<0.0001
b_{32}	3.1554	0.0867	<0.0001	3.1941	0.0964	<0.0001	2.9986	0.0446	<0.0001
b_{33}	–0.2093	0.1188	0.0790	–0.3206	0.1202	0.0081	—	—	—
b_{41}	1.1609	0.0527	<0.0001	1.0557	0.0508	<0.0001	1.1610	0.0527	<0.0001
b_{42}	0.6345	0.0175	<0.0001	0.6818	0.0189	<0.0001	0.6345	0.0175	<0.0001

weighting functions for each component equation (Parresol 2001; Zhao et al. 2015); and (3) finally, utilizing the resultant component-specific weighting functions in refitting the equation system using PROC MODEL with 2SLS, 3SLS, or FIML.

Evaluation

The following criteria were used to evaluate the performance of 2SLS, 3SLS, and FIML for the slash pine data: mean error (E), percent mean error (E%), mean absolute error (MABE), percent mean absolute error (MABE%), root-mean-squared error (RMSE), percent root-mean-squared error (RMSE%), and pseudo R^2, defined as follows:

$$E = \frac{\sum_{j=1}^{N}(y_{ij} - \hat{y}_{ij})}{N} \quad (5a)$$

$$E\% = \frac{100}{N}\sum_{j=1}^{N}\frac{y_{ij} - \hat{y}_{ij}}{y_{ij}} \quad (5b)$$

$$\text{MABE} = \frac{\sum_{j=1}^{N}|y_{ij} - \hat{y}_{ij}|}{N} \quad (6a)$$

$$\text{MABE\%} = \frac{100}{N}\sum_{j=1}^{N}\frac{|y_{ij} - \hat{y}_{ij}|}{y_{ij}} \quad (6b)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{N}(y_{ij} - \hat{y}_{ij})^2}{N}} \quad (7a)$$

$$\text{RMSE\%} = 100\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\frac{y_{ij} - \hat{y}_{ij}}{y_{ij}}\right)^2} \quad (7b)$$

$$R_i^2 = 1 - \frac{\sum_{j=1}^{N}(y_{ij} - \hat{y}_{ij})^2}{\sum_{j=1}^{N}(y_{ij} - \bar{y}_i)^2} \quad (8)$$

where y_{ij} and $\hat{y}_{ij}$ are the jth observed and predicted biomass for the ith component or total, and $\bar{y}_i$ is the mean of N biomass observations for the same component or total.

In this study, the biomass equation system was fitted to the entire data set (N = 306 trees). Model validation was accomplished by the leave-one-out (LOO) cross-validation technique, in which the model system was fitted using all but one tree (leaving one tree out), and then the fitted model system was used to predict the values of all components and total biomass for that left-out tree. The summary statistics were calculated using the same formulas (eqs. 5*a*–8).

The predictive performance of the new system of component equations was also compared with the previously reported biomass equation systems by Zhao et al. (2019) with aggregation and disaggregation approaches and fitted to the same slash pine data used here. The same evaluation criteria and the same data facilitate these comparisons.

Results

Model fitting by the weighted 2SLS, 3SLS, and FIML methods

Parameter estimates for the system of biomass equations (Models 1–4) fitted with the weighted 2SLS, 3SLS, and FIML are given in Table 1. The b_{33} parameter estimate was not significantly different from zero for the 2SLS option (p = 0.712) so that parameter was not considered in the final 2SLS estimation process. Conversely, the parameter estimate b_{33} was appreciably different from zero when fitted by 3SLS (p = 0.079) and FIML (p = 0.08). Table 1 indicates that the other parameter estimates were highly significant, with $p < 0.0001$ for the weighted 2SLS, 3SLS, or FIML estimation.

The equation system (Models 1–4) was first fitted using 2SLS method. Scatterplots of the residuals against the predicted values for each biomass component in the system equations fitted with 2SLS revealed significant heteroscedasticity (Appendix Fig. B1). Residual variances from the unweighted model system were modeled as a power function of DBH for stem bark and branch equations, and a power function of DBH and HT for stem wood and foliage biomass equations, leading to the weighting functions $\text{DBH}^{4.205}\,\text{HT}^{1.137}$, $\text{DBH}^{3.766}$, $\text{DBH}^{5.139}$, and $\text{DBH}^{6.061}\,\text{HT}^{-3.377}$, which were used for stem wood, stem bark, branch, and foliage biomass equations, respectively. In the PROC MODEL, the weights for different component equations are specified as an inverse to a square root of the corresponding weighting functions (Zhao et al. 2015). Finally, the equation system was refitted using 2SLS and the weighting functions. After fitting with weighting function for each equation, scatterplots of Pearson residuals against the predicted values for each biomass component showed that there were no marked departures that would nullify the homogeneous error variance assumption (Appendix Fig. B1). The 2SLS with or without weight functions did not consider the contemporaneous correlations among different biomass component equations but could estimate them. The estimated cross-correlation matrix with the weighted 2SLS was

Table 2. Fit statistics for the component biomass of slash pine trees from the model equations developed in this study fitted with weighted two-stage least squares (2SLS), three-stage least squares (3SLS), and full information maximum likelihood (FIML), previously developed equations using the aggregative approach fitted biomass component equations with weighted nonlinear seemingly unrelated regression (SUR2) (Zhao et al. 2019).

Method	Biomass	*E*	*E*%	MABE	MABE%	RMSE	RMSE%	R^2
2SLS	Stem wood	1.699	−1.698	14.217	10.338	31.354	14.331	0.980
	Stem bark	−0.003	−2.497	2.400	11.989	4.087	16.308	0.970
	Branch	0.145	−22.611	5.012	43.752	10.991	75.943	0.916
	Foliage	1.010	−7.445	2.135	35.323	4.498	49.657	0.787
	Total	2.851	−1.064	16.936	8.586	38.119	11.821	0.982
3SLS	Stem wood	1.773	−1.699	14.213	10.340	31.352	14.354	0.980
	Stem bark	0.038	−2.397	2.403	11.966	4.080	16.195	0.970
	Branch	−0.477	−23.539	4.973	44.008	10.728	77.406	0.920
	Foliage	1.010	−7.444	2.135	35.322	4.498	49.655	0.787
	Total	2.344	−1.135	16.627	8.550	36.963	11.830	0.984
FIML	Stem wood	1.387	3.597	13.803	11.167	29.179	14.983	0.982
	Stem bark	−0.097	−5.708	2.431	13.406	4.102	19.092	0.970
	Branch	0.593	−33.118	5.167	53.428	10.099	100.560	0.929
	Foliage	0.167	1.310	2.026	34.976	3.386	45.339	0.879
	Total	2.051	1.664	16.450	8.952	34.734	12.524	0.985
SUR2	Stem wood	2.584	−1.576	14.599	10.281	33.016	14.131	0.978
	Stem bark	0.086	−2.875	2.501	12.627	4.299	17.735	0.967
	Branch	−0.043	−21.553	4.986	43.201	10.940	74.714	0.917
	Foliage	−0.398	−20.103	2.109	40.342	4.147	67.313	0.819
	Total	2.229	−1.593	17.838	9.590	39.378	13.698	0.981

Note: *E*, mean prediction error; *E*%, percent mean prediction error; MABE, mean absolute error; MABE%, percent mean absolute error; RMSE, root-mean-square error; RMSE%, percent root-mean-square error; pseudo R^2.

$$(9)\quad \begin{array}{c} \\ \text{Wood} \\ \text{Bark} \\ \text{Branch} \\ \text{Foliage} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Wood} & \text{Bark} & \text{Branch} & \text{Foliage} \end{array} \\ \begin{pmatrix} 1 & 0.129 & 0.205 & 0.002 \\ & 1 & -0.093 & 0.038 \\ & & 1 & -0.197 \\ & & & 1 \end{pmatrix} \end{array}$$

In the same way, the equation system (Models 1–4) was fitted to the slash pine data using the weighted 3SLS. The weighting functions $DBH^{3.819}\ HT^{1.883}$, $DBH^{3.835}$, $DBH^{5.033}$, and $DBH^{6.061}\ HT^{-3.377}$ were used for stem wood, stem bark, branch, and foliage biomass equations, respectively, to address heteroscedasticity problem (Appendix Fig. B2). The estimated cross-correlation matrix among biomass component equations with the weighted 3SLS was

$$(10)\quad \begin{array}{c} \\ \text{Wood} \\ \text{Bark} \\ \text{Branch} \\ \text{Foliage} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Wood} & \text{Bark} & \text{Branch} & \text{Foliage} \end{array} \\ \begin{pmatrix} 1 & 0.117 & 0.211 & -0.001 \\ & 1 & -0.094 & 0.048 \\ & & 1 & -0.238 \\ & & & 1 \end{pmatrix} \end{array}$$

The equation system was also fitted using the weighted FIML. The weighting functions $DBH^{3.734}\ HT^{1.265}$, $DBH^{3.456}$, $DBH^{5.041}$, and $DBH^{5.690}\ HT^{-2.807}$ were used for stem wood, stem bark, branch, and foliage biomass models, respectively. Heteroscedasticity that existed in each of the biomass equation was well addressed by the weighted FIML (Appendix Fig. B3). The estimated cross-correlation matrix among biomass component equations with the weighted FIML was

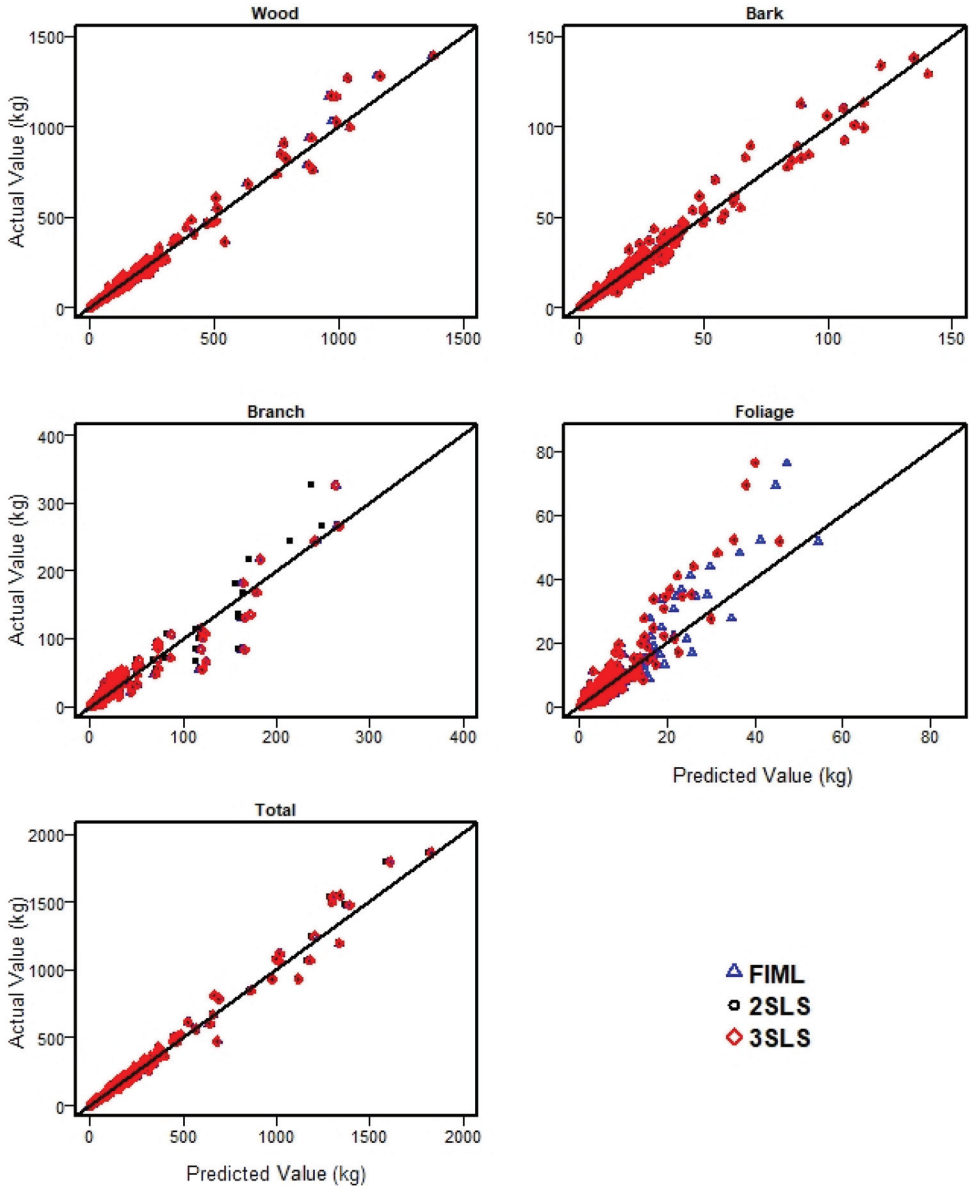
$$(11)\quad \begin{array}{c} \\ \text{Wood} \\ \text{Bark} \\ \text{Branch} \\ \text{Foliage} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Wood} & \text{Bark} & \text{Branch} & \text{Foliage} \end{array} \\ \begin{pmatrix} 1 & 0.115 & 0.188 & 0.051 \\ & 1 & -0.081 & 0.045 \\ & & 1 & -0.295 \\ & & & 1 \end{pmatrix} \end{array}$$

Evaluation of 2SLS, 3SLS, and FIML

Fit statistics for evaluating 2SLS, 3SLS, and FIML methods for fitting the system of eqs. 1–4 can be found in Table 2. The R^2 value was ≥0.98 for stem wood and total biomass, 0.97 for stem bark component, and 0.92–0.93 for branch component for all 2SLS, 3SLS, and FIML estimation methods. The FIML resulted in higher values of R^2 for total biomass and biomass components, except for stem bark component where all three estimation methods had the same value of R^2. For foliage component, FIML had R^2 = 0.879 and both 2SLS and 3SLS had R^2 = 0.787. For branch component, the values of *E*% were most substantially negative, less than –22.6% with all three estimation methods, while the *E* values were positive with 2SLS and FIML (0.15 and 0.59 kg, respectively) and negative with 3SLS (–0.48 kg). The values of *E*% for foliage component were –7.4% for both 2SLS and 3SLS, and 1.3% for FIML, while all *E* values for foliage component were positive with 1.0 kg for both 2SLS and 3SLS and 0.2 kg for FIML. The values of *E*% for stem wood and total biomass were negative (ranging from –1.7% to –1.1%) with 2SLS and 3SLS and positive with the FIML estimation (3.6% and 1.7%), while the *E* values for stem wood and total biomass were positive (ranging from 1.7 to 2.8 kg) with all estimation methods. The *E*% for stem bark component ranged from –5.7% to –2.4%, while all *E* values were close to zero (from –0.097 to 0.038 kg) for all three estimation methods. The values of MABE% and RMSE% were largest for foliage and branch biomass components and smaller for wood, bark, and total biomass predictions with all estimation methods. The values of MABE and RMSE, however, were naturally largest for stem wood and total biomass, and substantially smaller for bark and foliage biomass predictions with all estimation methods.

Figure 2 shows the relationship between the predicted and actual biomass components with 2SLS, 3SLS, and FIML estimation. As might be expected from Table 2, the performance of the three methods appears to be very close. Symbols for FIML and 3SLS are essentially superimposed at the level of resolution visible in Fig. 2 for nearly all predictions. 2SLS predictions are somewhat different in some cases especially for the branch and foliage biomass components. FIML and 3SLS tended to provide somewhat

Fig. 2. Relationships between predicted and actual stem wood, bark, branch, and foliage biomass components with prediction model equations fitted by two-stage least squares (2SLS), three-stage least squares (3SLS), and full information maximum likelihood (FIML) for slash pine. [Colour online.]



better predictions of the foliage and branch biomass components for large trees than 2SLS. However, all the three estimation methods underestimated foliage biomass for larger trees (DBH > 30 cm).

The performances of 2SLS, 3SLS, and FIML could be ranked simply by variation accounted for (i.e., R^2), by error statistics computed on the mass scale (*E*, MABE, RMSE), by error statistics computed on the percentage scale (*E*%, MABE%, RMSE%), or by combinations of these statistics. Based only on R^2, FIML was best for total biomass and biomass components, except for stem bark component where all three estimation methods had the same performance. Compared with 2SLS and 3SLS (R^2 = 0.787), FIML seemed to substantially improve foliage component prediction (R^2 = 0.879). Based only on R^2, the overall ranking of the three estimation methods followed the order of FIML > 3SLS > 2SLS, with the sum of the ranks being 5, 9, and 11, respectively (Appendix Table B1). Using *E*, MABE, RMSE, and R^2 to rank the three estimation methods, the FIML was best for stem wood, foliage, and total tree biomass, while the 3SLS was best for branch and 2SLS best for stem bark component estimation, leading to the overall ranking: FIML > 3SLS > 2SLS, with the sum of the ranks being 28, 39, and 45, respectively (Appendix Table B1). Using *E*%, MABE%, RMSE%, and R^2 to rank the three estimation methods, FIML was best for foliage component, 3SLS best for bark and branch components, 2SLS best for stem wood component, and both 2SLS and 3SLS better for total biomass. This resulted in the overall ranking: 3SLS > 2SLS > FIML, with no large difference in the sum of the ranks being 36, 40, and 42, respectively (Appendix Table B1). If considering all criteria, FIML was best for stem wood and foliage components and had the same performance as 3SLS for total biomass, 3SLS was also best for stem bark component and 2SLS best for branch component. The FIML and 3SLS had almost the same overall performance and were better than the 2SLS. The order of overall ranking followed: FIML > 3SLS > 2SLS, with the sum of the ranks being 65, 66, and 71, respectively (Appendix Table B1).

Table 3. Leave-one-out (LOO) cross-validation statistics for component biomass of slash pine trees from the model equations developed in this study fitted with weighted two-stage least squares (2SLS), three-stage least squares (3SLS), and full information maximum likelihood (FIML).

Method	Biomass	*E*	*E*%	MABE	MABE%	RMSE	RMSE%	R^2
2SLS	Stem wood	1.806	−1.699	14.493	10.445	32.212	14.495	0.979
	Stem bark	−0.002	−2.900	2.540	12.778	4.362	17.900	0.966
	Branch	0.148	−22.733	5.082	44.025	11.264	76.358	0.912
	Foliage	0.943	−16.548	2.353	42.240	4.901	67.176	0.747
	Total	2.896	−1.180	18.271	9.606	41.034	13.440	0.980
3SLS	Stem wood	1.982	−1.701	14.474	10.448	32.244	14.530	0.979
	Stem bark	0.044	−2.815	2.542	12.750	4.375	17.830	0.966
	Branch	−0.471	−23.714	5.075	44.409	11.113	78.078	0.914
	Foliage	0.875	−16.683	2.302	41.723	4.705	66.610	0.766
	Total	2.340	−1.315	18.017	9.596	39.960	13.559	0.984
FIML	Stem wood	1.994	−1.124	14.616	10.410	32.887	14.225	0.978
	Stem bark	0.030	−2.590	2.549	12.730	4.379	17.681	0.965
	Branch	−0.305	−25.029	5.039	45.243	10.870	80.694	0.918
	Foliage	0.495	−15.959	2.178	40.297	4.137	64.382	0.819
	Total	2.213	−0.886	18.069	9.639	40.294	13.483	0.980

Note: *E*, mean prediction error; *E*%, percent mean prediction error; MABE, mean absolute error; MABE%, percent mean absolute error; RMSE, root-mean-square error; RMSE%, percent root-mean-square error; pseudo R^2.

LOO cross-validation statistics for evaluating 2SLS, 3SLS, and FIML methods for fitting the system of eqs. 1–4 can be found in Table 3. The performances of these three fitting methods were also ranked based on the cross-validation statistics (Appendix Table B2). Compared with the ranks based on the fit statistics, there were some little changes in the ranks of 2SLS, 3SLS, and FIML for some biomass components based on the LOO cross-validation statistics. Both the fit and LOO cross-validation statistics suggested that FIML improved foliage component prediction in terms of either R^2, or its combinations with other error statistics on the mass scale, on the percentage scale, or on both the scales (Table 3 and Appendix Table B2). Using R^2 or (*E*, MABE, RMSE, R^2) associated with LOO cross-validation results, the FIML and 3SLS had the same or almost the same overall performance and were better than the 2SLS (Appendix Table B2). Based on LOO cross-validation statistics *E*%, MABE%, RMSE%, and R^2, the order of overall ranking followed: FIML > 2SLS > 3SLS, with the sum of the ranks being 33, 40, and 42, respectively (Appendix Table B2). If considering all criteria, the LOO cross-validation statistics led to the same order of overall ranking: FIML > 3SLS > 2SLS, as the fit statistics did.

Discussion

Inspection of Tables 2 and 3 indicates that most of the *E*% values are negative, while most of the corresponding *E* values are positive. The difference in the signs between *E*% and *E* is largely due to the mathematical formula for *E*%, eq. 5*b*. For a given level of the explanatory variables in the prediction model, the negative deviations (actual values are smaller than the predicted) have smaller denominators than positive deviations (actual values are larger than the predicted) of equal absolute value. These negative values of deviation divided by actual value are larger than corresponding positive terms at that level of the explanatory values. Thus, negative terms tend to "outweigh" positive terms in the computation of *E*% although it is possible to obtain positive values of *E* as reported in Tables 2 and 3. A model equation that is performing well would be expected to have a negative but small *E*% value and a small *E* value. Furthermore, there was no single system to predict biomass that was best for all components and total tree biomass, as demonstrated in the current study and others (Zhao et al. 2015, 2019; Dong et al. 2015). So, the better practice for evaluating different systems of biomass equations is to compare their overall predictive performances based on an array of statistics in absolute units and percentages for each biomass component and total tree biomass, using a ranking system as we did in the current study.

Zhao et al. (2019) developed three systems of biomass equations for slash pine using the same dataset used in the current study. Two systems were developed using the aggregation approach, one of which included component biomass equations and a total biomass equation that were jointly fitted using weighted SUR (SUR1), and the other included component equations only that were fitted using weighted SUR (SUR2). The third system followed a disaggregative approach and involved multiplying total biomass predictions by the estimated component ratios from the Dirichlet regression model for component ratios (DRM). The right-hand sides of equations in these systems included only exogenous variables. Based on *E*%, MABE%, RMSE%, and R^2, Zhao et al. (2019) found that the system associated with SUR2 had the best overall performance, when compared to the SUR1 and DRM systems. For comparison purposes in this study, we also calculated the *E*, MABE, and RMSE for the SUR2 system (see Table 2). In terms of the overall prediction performance, the new system developed in this study with FIML or 3SLS was marginally superior to the SUR2 system.

Recall, the SUR2 system includes only component biomass equations that are a power function of DBH and HT (Zhao et al. 2019). The system developed in the study (Models 1–4) also consisted of component biomass equations, in which stem wood and branch component equations are a power function of DBH and HT but stem bark and foliage component equations include an endogenous right-hand-side variable. Neither system included a total biomass prediction model in the estimation process, and total biomass predictions were obtained by adding predictions of individual biomass components. Zhao et al. (2019) found that the SUR2 was better for predicting total biomass than the SUR1 even though SUR1 uses total biomass as a dependent variable. The SUR2 was even better for predicting total biomass than the separately developed total biomass model (Zhao et al. 2019). It is especially revealing that addition of biomass component predictions from 2SLS, 3SLS, and FIML performed better than the SUR2, in terms of an array of statistics utilized (Table 2). This again confirmed that it is not necessary to include a total biomass equation in a system of biomass equations.

There were large differences in foliage biomass prediction among these approaches (Fig. 2, and figure 5 in Zhao et al. 2019). The 2SLS, 3SLS, and FIML methods tend to underpredict foliage

biomass for large trees while the SUR1, SUR2, and DRM methods tend to overpredict foliage biomass for large trees. For easier comparison, the actual versus predicted biomass components and total biomass from the current FIML system and the previous SUR2 system were shown in Appendix Fig. B4. For several of large values of foliage biomass, the FIML symbol is above and the SUR2 symbol is below the 1:1 line, but the FIML symbol is closer to the 1:1 line. Thus, FIML tended to provide somewhat better predictions of foliage biomass for larger trees than SUR2, as indicated by substantially better values of E, E%, MABE, MABE%, RMSE, RMSE%, and R^2 provided by FIML (Table 2). Possibly the link between branch and foliage predictions used by 2SLS, 3SLS, and FIML enabled improvements in foliage biomass prediction for this slash pine dataset.

The heteroscedasticity problem that always exists in biomass model residuals could be addressed by having each equation with its own weighting function as we did in this study and others (Zhao et al. 2015, 2019; Dong et al. 2015). In addition to this problem, mathematical relationships between biomass equation themselves and relationships between the error terms of biomass equations determine how to develop and estimate the system of biomass equations. For the system of biomass component equations that are all defined as a power function of DBH and HT, when it was fitted using weighted nonlinear least squares estimation (WNLSE) with the same weighted functions used in SUR2 (Zhao et al. 2019), the estimated cross-correlation matrix (Appendix Matrix B1) indicated high correlations between stem wood and bark residuals (0.322), branch and foliage residuals (0.421), stem wood and branch residuals (0.205), and between stem wood and foliage residuals (0.208). The presence of correlation between error terms of biomass equations and the large sample size (N = 306) enable the SUR approach to achieve more efficient estimation compared with WNLSE (Zhao et al. 2019). In the current study, we took advantage of the close relationships between stem wood and stem bark, and between branch and foliage components (Fig. 1) and proposed a system of component equations including endogenous right-hand-side variables. When this new system was naively fitted using WNLSE, the estimated cross-correlation matrix across the component equations (Appendix Matrix B2) indicated a substantial reduction in correlations between stem wood and bark residuals, between stem wood and foliage residuals, and between branch and foliage residuals, compared with Appendix Matrix B1. This result highlighted the importance of mathematical relationships between equations (i.e., model structure) on correlation between the errors of the equations.

For comparison purpose, we reported the cross-correlation matrix across equations estimated using weighted 2SLS (Matrix 9), although the 2SLS approach did not consider such correlations. All the estimated cross-correlation matrices for 2SLS (Matrix 9), 3SLS (Matrix 10), and FIML (Matrix 11) are very similar for most elements. Although the correlation between wood and foliage residuals is positive for 2SLS and FIML and negative for 3SLS, these correlations are quite small being less than 0.051 in absolute value. The correlations among other biomass components are all the same sign for 2SLS, FIML, and 3SLS and very similar in magnitude. Unlike the 2SLS approach, the 3SLS and FIML approaches estimates all coefficients simultaneously by taking into account such correlations across equations and are expected to improve the efficiency of the estimation. However, in the current study the 3SLS and FIML did not appear to achieve a noticeable reduction in standard errors of parameter estimates except for parameter b_{33} (Table 1) compared to 2SLS. This is due to small correlations between the errors of equations.

When estimates from 3SLS and FIML are compared, many parameters in the model system had similar estimates (Table 1). The performance of 3SLS and FIML was particularly similar. This might be expected with a sample size of over 300 trees because 3SLS and FIML are asymptotically equivalent for a large size sample. If the sample size is relatively small, the 3SLS is a compelling choice compared to the FIML. In the presence of correlations of error terms of biomass equations, the 3SLS and FIML with large samples should achieve more efficient estimates compared with the 2SLS.

Conclusions

Unlike the previously published biomass equations developed with the aggregation and disaggregation strategies, which include only exogenous variables on the right-hand sides, a different system of biomass component equations has been proposed for slash pines using an econometric approach. By taking advantage of relationships between the biomass components, the new system included endogenous right-hand-side variables in stem bark and foliage equations, in which parameters were estimated using the weighted nonlinear 2SLS, 3SLS, and FIML. 3SLS and FIML were extremely close each other in performance, and both appeared to provide somewhat better predictions of branch and foliage biomass for large trees and may have better properties in extrapolations beyond the fitting data for large trees for these biomass components. The overall predictive performances followed the order of FIML $>$ 3SLS $>$ 2SLS. The new system performed well or better than previously published biomass equation systems developed using an aggregative approach and fitted to the same data using weighted nonlinear seemingly unrelated regressions (SUR). Our results demonstrated that relationships between the biomass components can be used to enhance biomass predictions. The econometrics such as 3SLS and FIML, which allow the endogenous (dependent) variables in any of the system model equations to be used as explanatory variables in other system model equations, expand the possibilities for biomass component modeling. It is likely that these approaches to the development of integrated systems of biomass component prediction model equations could be successfully applied to data from other southern pine species as well as many other tree species groups.

Acknowledgements

This research was partially funded by a joint venture agreement (19-JV-11330145-058) between USDA Forest Service and University of Georgia (UGA). This work was also supported by McIntire Stennis Project (OKL0-3063 at Oklahoma State University and GEOZ-0180-MS at UGA) founded by the USDA National Institute of Food and Agriculture. We thank the Plantation Management Research Cooperative (PMRC) technical staff for their hard work in field sampling and data collection and thank PMRC members for permitting access and destructive sampling on their lands.

References

Affleck, D.L.R., and Diéguez-Aranda, U. 2016. Additive nonlinear biomass equations: a likelihood-based approach. For. Sci. **62**(2): 129–140. doi:10.5849/forsci.15-126.

Belsley, D.A. 1988. Two- or three-stage least squares? Comput. Sci. Econ. Manage. **1**: 21–30. doi:10.1007/BF00435200.

Borders, B.E. 1986. A compatible system of growth and yield equations for slash pine fitted with restricted three-stage least squares. For. Sci. **32**: 185–201. doi:10.1093/forestscience/32.1.185.

Clutter, J.L., Fortson, J.C., Pienaar, L.V., Brister, G.H., and Bailey, R.L. 1983. Timber management: a quantitative approach. John Wiley and Sons, New York.

Dong, L., Zhang, L., and Li, F. 2015. A three-step proportional weighting system of nonlinear biomass equations. For. Sci. **60**(1): 34–45. doi:10.5849/forsci.13-193.

Furnival, G.M., and Wilson, R.W. 1971. Systems of equations for predicting forest growth and yield. *In* Statistical ecology. Vol. 3. *Edited by* G.P. Patil, E.C. Pielou, and W.E. Waters. Pennsylvania State University Press, University Park, Pa. pp. 43–55.

Jenkins, J.C., Chojnacky, D.C., Heath, L.S., and Birdsey, R.A. 2003. National scale biomass estimators for United States tree species. For. Sci. **49**(1): 12–35. doi:10.1093/forestscience/49.1.12.

Judge, G.G., Griffiths, W.E., Hill, R., Lütkepohl, H., and Hill, T. 1985. The theory and practice of econometrics. 2nd ed. John Wiley & Sons, New York.
Lynch, T.B., and Clutter, M.L. 1998. A system of equations for prediction of plywood veneer yield and yield by grade for loblolly pine plywood bolts. For. Prod. J. **48**: 80–88. doi:10.1093/forestscience/49.1.12.
Murphy, P.A., and Sternitzke, H.S. 1979. Growth and yield estimation for loblolly pine in the west Gulf. Research Paper SO-154. USDA Forest Service Southern Experiment Station.
Parresol, B.R. 2001. Additivity of nonlinear biomass equations. Can. J. For. Res. **31**(5): 865–878. doi:10.1139/x00-202.
Radtke, P.J., Walker, D.M., Weiskittel, A.R., Frank, J., Coulston, J.W., and Westfall, J.A. 2015. Legacy tree data: a national database of detailed measurements for volume weight and physical properties. *Edited by* S.M. Stanton and G.A. Christensen. *In* New Directions in Inventory Techniques & Applications Forest Inventory & Analysis (FIA) Symposium 2015, Pushing Boundaries. pp. 25–30. PNW- GTR-931.
Rothenberg, T.J., and Leenders, C.T. 1964. Efficient estimation of simultaneous equation systems. Econometrica, **32**: 57–76. doi:10.2307/1913734.
Sabatia, C.O., Lynch, T.B., and Will, R.E. 2008. Tree biomass equations for naturally regenerated shortleaf pine. South. J. Appl. For. **32** (4): 163–167. doi:10.1093/sjaf/32.4.163.
SAS Institute Inc. 2011. SAS/ETS® 9.3 User's Guide. SAS Institute Inc., Cary, N.C.
Schumacher, F.X., and Hall, F.S. 1933. Logarithmic expression of timber-tree volume. J. Agric. Res. **47**: 719–734.
Tang, S., Zhang, H., and Xu, H. 2000. Study on establish and estimate method of compatible biomass model. Sci. Silv. Sin. **36** (Sp.1): 19–27. [In Chinese with English abstract.]
Tang, S., Li, Y., and Wang, Y. 2001. Simultaneously equations, error-in-variable models, and model integration in systems ecology. Ecol. Modell. **142**: 285–294. doi:10.1016/S0304-3800(01)00326-X.
Theil, H. 1971. Principles of econometrics. John Wiley & Sons, New York.
Zellner, A. 1962. An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. J. Am. Stat. Assoc. **57**: 348–368. doi:10.1080/01621459.1962.10480664.
Zhao, D., and Kane, M. 2017. New variable-top merchantable volume and weight equations derived directly from cumulative relative profiles for loblolly pine. For. Sci. **63**(3): 261–269. doi:10.5849/FS-2016-076.
Zhao, D., Kane, M., Markewitz, D., Teskey, R., and Clutter, M. 2015. Additive tree biomass equations for midrotation loblolly pine plantations. For. Sci. **61**(4): 613–623. doi:10.5849/forsci.14-193.
Zhao, D., Kane, M., Teskey, R., and Markewitz, D. 2016. Modeling above-ground biomass components and volume-to-weight conversion ratios for loblolly pine trees. For. Sci. **62**(5): 463–473. doi:10.5849/forsci.15-129.
Zhao, D., Westfall, J., Coulston, J., Lynch, T.B., Bullock, B.P., and Montes, C.R. 2019. Additive biomass equations for slash pine trees: comparing three approaches. Can. J. For. Res. **49**(1): 27–40. doi:10.1139/cjfr-2018-0246.

Appendix A: Potential bias in component ratio estimation with ordinary least squares

Let the value of a tree biomass component Y_C be expressed as the product of the total tree biomass Y_T multiplied by the ratio R_C of that component biomass to the total tree biomass:

$$(A1) \qquad Y_C = Y_T R_C$$

This equation is trivially correct for the observed values of total, component, and component ratios for a particular tree. The usual biomass components of interest include wood, bark, branch, and foliage components which add to total tree biomass Y_T. We now express Y_T and Y_C as true mean models f and g that are functions of a vector of independent variables $\boldsymbol{X}$, a vector of true parameter values $\boldsymbol{\beta}$ and true error terms ε_T and ε_C such that $E(\varepsilon_T) = E(\varepsilon_C) = 0$:

$$(A2) \qquad Y_T = f(\boldsymbol{\beta}_T, \boldsymbol{X}) + \varepsilon_T$$

$$(A3) \qquad R_C = g(\boldsymbol{\beta}_C, \boldsymbol{X}) + \varepsilon_C$$

The two equations above are trivially correct because the equalities depend on true mean models and true error terms. Nonlinear regression analyses would typically attempt to find good approximations for the functional forms f and g together with good parameter estimates $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\beta}}_C$. Now we want to estimate the expected value of the component as computed from the component ratio approach:

$$(A4) \qquad E(Y_C) = E(Y_T \times R_C) = E\{[f(\boldsymbol{\beta}_T, \boldsymbol{X}) + \varepsilon_T] \times [g(\boldsymbol{\beta}_C, \boldsymbol{X}) + \varepsilon_C]\}$$

$$(A5) \qquad \begin{aligned} E(Y_C) &= E[f(\boldsymbol{\beta}_T, \boldsymbol{X})g(\boldsymbol{\beta}_C, \boldsymbol{X}) + f(\boldsymbol{\beta}_T, \boldsymbol{X})\varepsilon_C + g(\boldsymbol{\beta}_C, \boldsymbol{X})\varepsilon_T + \varepsilon_T\varepsilon_C] \\ &= f(\boldsymbol{\beta}_T, \boldsymbol{X})g(\boldsymbol{\beta}_C, \boldsymbol{X}) + f(\boldsymbol{\beta}_T, \boldsymbol{X})E(\varepsilon_C) + g(\boldsymbol{\beta}_C, \boldsymbol{X})E(\varepsilon_T) + E(\varepsilon_T\varepsilon_C) \end{aligned}$$

Recall that $E(\varepsilon_T) = E(\varepsilon_C) = 0$, so $\text{cov}(\varepsilon_T, \varepsilon_C) = E(\varepsilon_T\varepsilon_C) - E(\varepsilon_T)E(\varepsilon_C) = E(\varepsilon_T\varepsilon_C)$, leading to

$$(A6) \qquad E(Y_C) = f(\boldsymbol{\beta}_T, \boldsymbol{X})g(\boldsymbol{\beta}_C, \boldsymbol{X}) + \text{cov}(\varepsilon_T, \varepsilon_C)$$

Therefore, even if the true mean and their true parameter values were known for total biomass $f(\boldsymbol{\beta}_T, \boldsymbol{X})$ and the ratio of the desired component to total $g(\boldsymbol{\beta}_C, \boldsymbol{X})$, there would be a bias equal to the covariance of the error terms $\text{cov}(\varepsilon_T, \varepsilon_C)$ for estimating expected amounts of components by using the product of the true mean models $f(\boldsymbol{\beta}_T, \boldsymbol{X})g(\boldsymbol{\beta}_C, \boldsymbol{X})$.

Now we wish to investigate the more realistic situation in which the parameters in the true mean models f and g are not known but estimated independently by ordinary least squares (OLS). Consider the following covariance between predictions from product of the estimated mean models for total biomass and component ratio:

$$(A7) \qquad \begin{aligned} \text{cov}(\hat{Y}_T, \hat{R}_C) &= E(\hat{Y}_T \times \hat{R}_C) - E(\hat{Y}_T)E(\hat{R}_C) \\ &= E(\hat{Y}_T \times \hat{R}_C) - f(\boldsymbol{\beta}_T, \boldsymbol{X})g(\boldsymbol{\beta}_C, \boldsymbol{X}) \end{aligned}$$

$$(A8) \qquad \Rightarrow \qquad f(\boldsymbol{\beta}_T, \boldsymbol{X})g(\boldsymbol{\beta}_C, \boldsymbol{X}) = E(\hat{Y}_T \times \hat{R}_C) - \text{cov}(\hat{Y}_T, \hat{R}_C)$$

Substituting eq. A8 into eq. A6 leads to

$$(A9) \qquad E(Y_C) = E(\hat{Y}_T \times \hat{R}_C) - \text{cov}(\hat{Y}_T, \hat{R}_C) + \text{cov}(\varepsilon_T, \varepsilon_C)$$

Thus, when multiplying the estimated mean models to estimated expected $E(Y_C)$ there is a bias of

$$(A10) \qquad \text{Bias} = \text{cov}(\hat{Y}_T, \hat{R}_C) - \text{cov}(\varepsilon_T, \varepsilon_C)$$

Superficially we may think the two covariances above would tend to cancel each other, but these covariances are conditional on a level of the independent variable vector $\boldsymbol{X}$, and so $\text{cov}(\hat{Y}_T, \hat{R}_C)$ may not be large. This covariance arises from estimation of the parameter vectors $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\beta}}_C$ and the fact that these parameter estimates are correlated assuming that they are based on the same sample trees. On the other hand, $\text{cov}(\varepsilon_T, \varepsilon_C)$ is usually assumed to be constant for all levels of $\boldsymbol{X}$ because the variance–covariance matrix for regression problems is usually assumed to be constant. Further there is no guarantee that the two covariances in the equation above have the same sign so instead of tending to cancel each other the opposite situation may hold.

Appendix B

All biomass component equations were defined as power functions of DBH and HT. These equations were fitted using weighted nonlinear least squares estimation (WNLSE), with the same weighted functions used in SUR2 (Zhao et al. 2019). The

estimated cross-correlation matrix among biomass component equations was

$$(B1)\quad \begin{array}{c} \\ \text{Wood} \\ \text{Bark} \\ \text{Branch} \\ \text{Foliage} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Wood} & \text{Bark} & \text{Branch} & \text{Foliage} \end{array} \\ \begin{pmatrix} 1 & 0.322 & 0.205 & 0.208 \\ & 1 & -0.032 & 0.023 \\ & & 1 & 0.421 \\ & & & 1 \end{pmatrix} \end{array}$$

The system of biomass component equations proposed in the current study (Models 1–4) included endogenous right-hand-side variables. When these equations were fitted using WNLSE with the same weighted functions used in 2SLS, the estimated cross-correlation matrix among biomass component equations was

$$(B2)\quad \begin{array}{c} \\ \text{Wood} \\ \text{Bark} \\ \text{Branch} \\ \text{Foliage} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Wood} & \text{Bark} & \text{Branch} & \text{Foliage} \end{array} \\ \begin{pmatrix} 1 & 0.079 & 0.208 & 0.041 \\ & 1 & -0.102 & 0.064 \\ & & 1 & -0.212 \\ & & & 1 \end{pmatrix} \end{array}$$

Given a system of biomass component equations (Models 1–4), three fitting methods (2SLS, 3SLS, and FIML) were ranked based on E, E%, MABE, MABE%, RMSE, RMSE%, and R^2 for each biomass component and total biomass in Table 2 for the fit statistics and in Table 3 for the LOO cross-validation statistics. The attributes were equally weighted. Rank one was used for the best method and three for the poorest. Appendix Tables B1 and B2 show the sum of the ranks of the fitting methods.

Table B1. Sum of the ranks, and ranks based on the rank sum (in brackets) of the three fitting methods, based on the fit statistics in Table 2.

		Biomass component and total biomass					
Criteria used for ranking	Fitting method	Wood	Bark	Branch	Foliage	Total biomass	Rank sum
R^2	2SLS	2	1	3	2	3	11 (3)
	3SLS	2	1	2	2	2	9 (2)
	FIML	1	1	1	1	1	5 (1)
E, MABE, RMSE, R^2	2SLS	10	6	9	8	12	45 (3)
	3SLS	9	7	7	8	8	39 (2)
	FIML	4	8	8	4	4	28 (1)
E%, MABE%, RMSE%, R^2	2SLS	6	7	9	11	7	40 (2)
	3SLS	9	4	8	8	7	36 (1)
	FIML	8	10	10	4	10	42 (3)
All criteria	2SLS	14	12	12	17	16	71 (3)
	3SLS	16	10	13	14	13	66 (2)
	FIML	11	17	17	7	13	65 (1)

Table B2. Sum of the ranks, and ranks based on the rank sum (in brackets) of the three fitting methods, based on the leave-one-out (LOO) cross-validation statistics in Table 3.

		Biomass component and total biomass					
Criteria used for ranking	Fitting method	Wood	Bark	Branch	Foliage	Total biomass	Rank sum
R^2	2SLS	1	1	3	3	2	10 (3)
	3SLS	1	1	2	2	1	7 (1)
	FIML	2	2	1	1	2	8 (2)
E, MABE, RMSE, R^2	2SLS	5	4	11	12	11	43 (2)
	3SLS	6	8	9	8	5	36 (1)
	FIML	11	10	4	4	7	36 (1)
E%, MABE%, RMSE%, R^2	2SLS	7	9	6	11	7	40 (2)
	3SLS	9	8	8	9	8	42 (3)
	FIML	5	6	10	4	8	33 (1)
All criteria	2SLS	11	12	14	20	16	73 (3)
	3SLS	15	15	15	15	12	72 (2)
	FIML	14	14	13	7	13	61 (1)

Fig. B1. Residual plots (left: A1–D1) for each biomass component fitted using 2SLS without weight functions, showing significant heteroscedasticity; Pearson residual plots (right: A2–D2) for each biomass component fitted using 2SLS with weight functions.

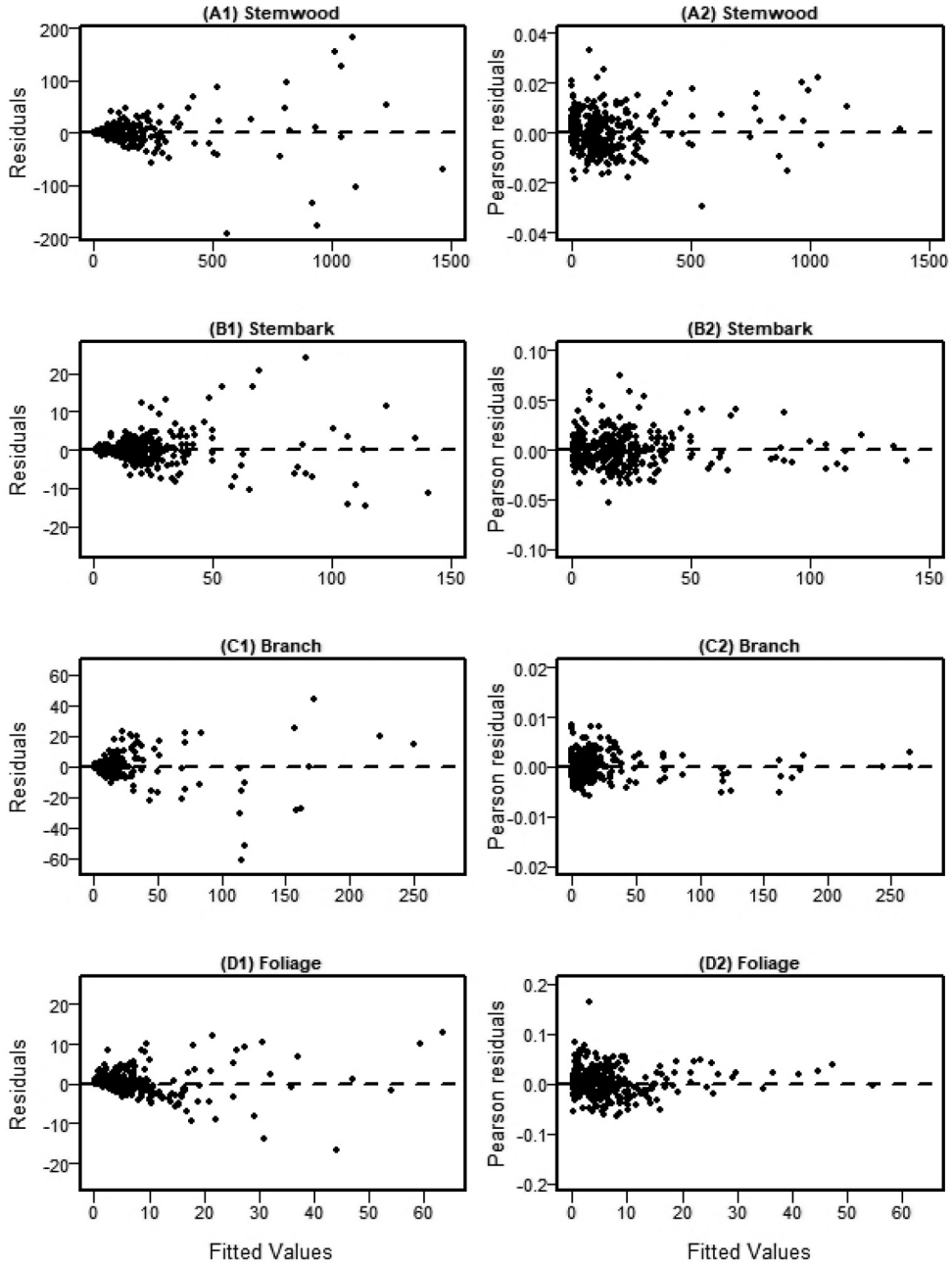


Fig. B2. Residual plots (left: A1–D1) for each biomass component fitted using 3SLS without weight functions, showing significant heteroscedasticity; Pearson residual plots (right: A2–D2) for each biomass component fitted using 3SLS with weight functions.

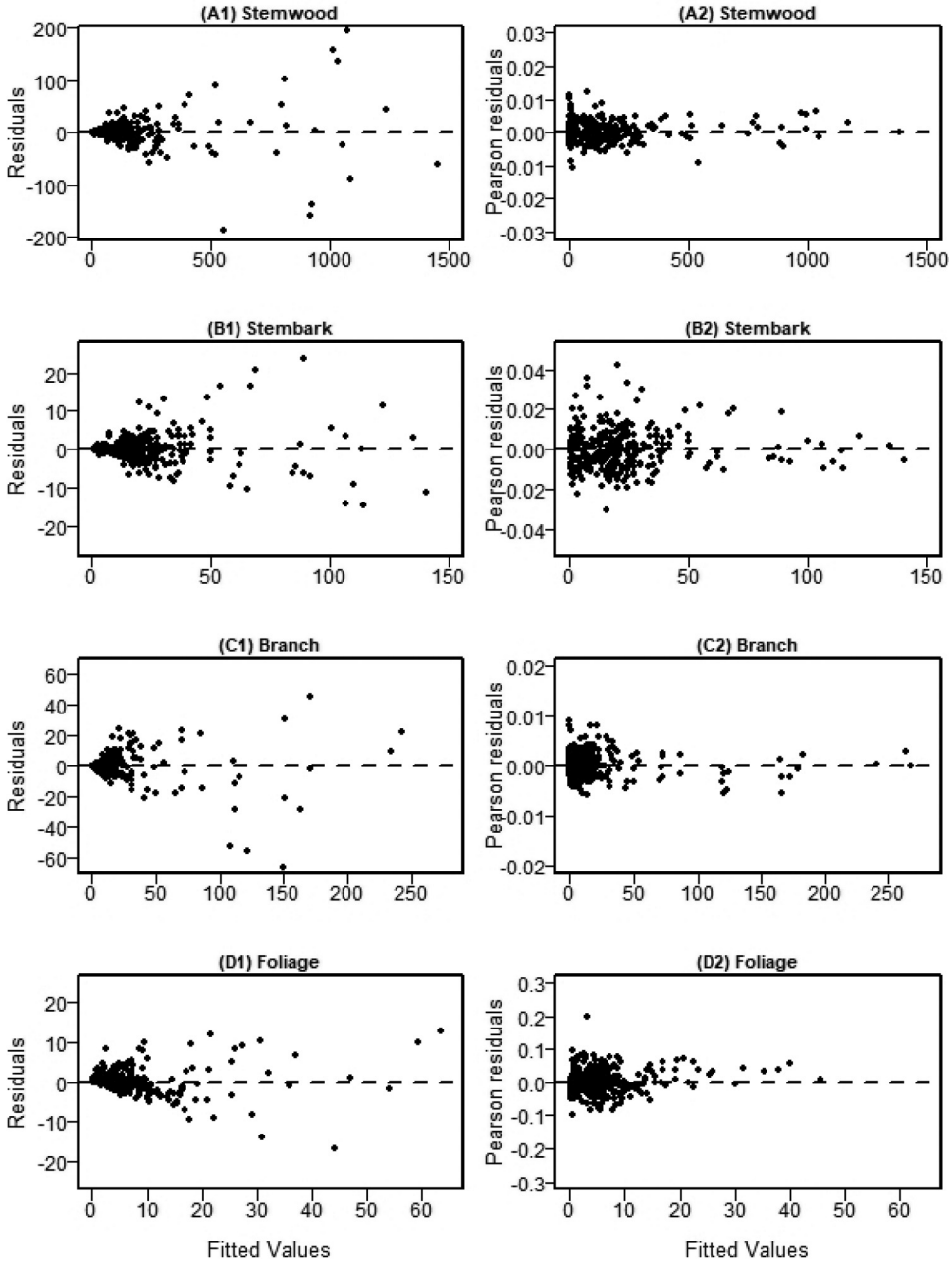


Fig. B3. Residual plots (left: A1–D1) for each biomass component fitted using FIML without weight functions, showing significant heteroscedasticity; Pearson residual plots (right: A2–D2) for each biomass component fitted using FIML with weight functions.

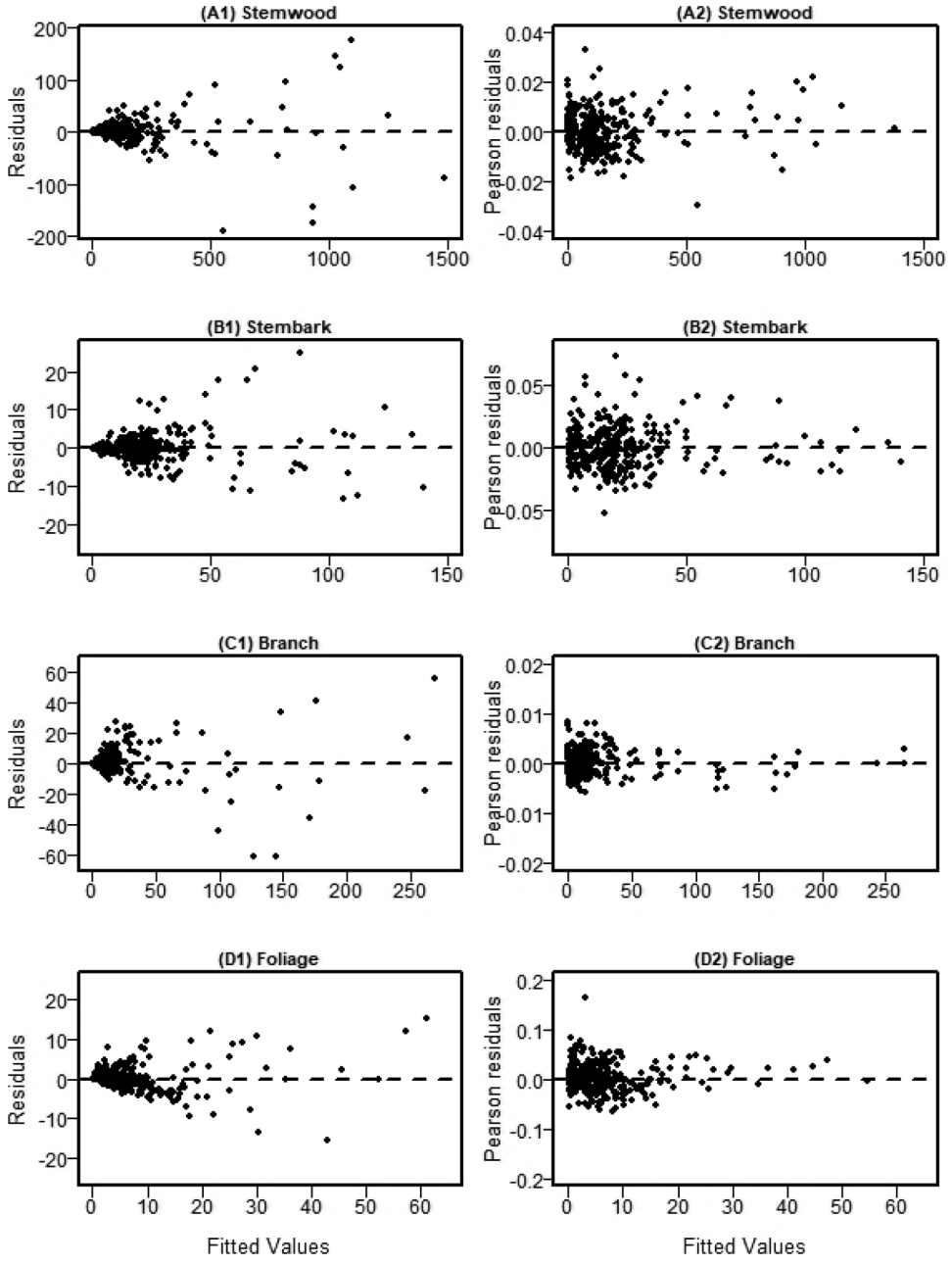


Fig. B4. Comparisons of stem wood, stem bark, branch, foliage, and total tree aboveground biomass predictions from the currently developed system with FIML and a previously developed aggregative system fitted with SUR2 (Zhao et al. 2019). [Colour online.]