NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

18-13

Correction factors for unbiased, efficient estimation and prediction of biomass from log-log allometric models

David Clifford, Noel Cressie, Jacqueline R. England, Stephen H. Roxburgh and Keryn I. Paul

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Correction factors for unbiased, efficient estimation and prediction of biomass from log-log allometric models

David Clifford[a,b], Noel Cressie[c], Jacqueline R. England[a,d], Stephen H. Roxburgh[a,e], Keryn I. Paul[a,e]

[a] Commonwealth Scientific and Industrial Research Organisation (CSIRO) Sustainable Agriculture Flagship, PO Box 2583, Brisbane, QLD 4001, Austraila;

[b] CSIRO Computational Informatics, PO Box 2583, Brisbane, QLD 4001, Australia;

[c] National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia;

[d] CSIRO Ecosystem Sciences, Private Bag 10, Clayton South VIC 3169, Australia;

[e] CSIRO Ecosystem Sciences, GPO Box 1700, ACT 2601, Australia.

Corresponding author:

David Clifford

Telephone: +61 7 3833 5532

Email: David.Clifford@csiro.au

Highlights:

- We review nine alternatives for correcting bias in log-log allometrics.
- We use simulations to evaluate the ability of these to estimate average biomass.
- We evaluate their ability to predict biomass of new trees.
- Methods not commonly used in forest science performed best.

Abstract

Allometric relationships are commonly used to estimate average biomass of trees of a particular size and to predict biomass of individual trees based on an easily measured covariate variable such as stem diameter. They are typically power relationships which, for the purpose of data fitting, are transformed using natural logarithms to convert the model to its linear equivalent. Implementation of these equations to estimate the relationships and to predict biomass of new trees on the natural (i.e., actual) scale requires back-transforming the logarithmic predictions. Because these transformations involve non-linearity, care must be taken during this step to avoid bias. Several correction factors have been proposed in the literature for removing the gross bias in estimates, but their performance as predictors of biomass has not yet been examined. This is a very important problem, and here we review nine such correction factors in terms of their abilities to estimate biomass and predict biomass for new trees. We compare their performance by examining their bias and variability based on large datasets of above-ground biomass and stem diameter for eight species of harvested trees and shrubs in the genera *Eucalyptus* and *Acacia* (n = 102-365 individuals per species). We found that good estimates of average biomass turned out to be good predictors of biomass for new trees. The linear model fitted has log of the above-ground biomass as the response variable and log of the stem diameter as the covariate. The only exactly unbiased estimate among those considered was the uniform minimum variance unbiased (UMVU) estimate, which involves evaluating a confluent hypergeometric function to obtain its correction factor. Three alternative correction factors that are easy to compute also performed well. One of these minimises mean squared error and was found to result in low bias, low prediction bias, the lowest mean squared error, and the lowest mean squared prediction error among all correction factors examined.

Keywords: allometry, *Eucalyptus*, *Acacia*, above-ground biomass, destructive sampling, stem diameter

Introduction

Forests are an important component of the global carbon cycle through the flux and storage of carbon in plant biomass and soil. Accurate quantification of forest biomass is therefore important for understanding carbon stocks in existing forest ecosystems and the potential for greenhouse mitigation from reforestation and afforestation. A relatively easy, non-destructive evaluation of biomass can be obtained from above-ground measurements.

Allometric relationships are commonly used in this regard. For trees, such relationships are typically power equations of the form, $y = a.x^b$, which relate biomass y to a covariate x such as stem diameter at breast height. A transformation using natural logarithms converts this equation to its linear equivalent, $\ln(y) = \beta_0 + \beta_1.\ln(x)$, where $\beta_0 = \ln(a)$ and $\beta_1 = b$. Typically, a stochastic version of this, namely $\ln(y) = \beta_0 + \beta_1.\ln(x) + \varepsilon$, is fitted using the standard regression assumptions that the error ε has zero mean, constant variance, and is normally distributed. Allometric models may also be based on additional covariates (e.g., Kuyah et al. 2012), and the formulae given in this paper accommodate multivariate regression.

Other workers (Parresol 2001; Lambert et al. 2005) have found that modelling the error structure on the original data scale can on occasions give results as good as or even better than applying a transformation. Another alternative to the simple power-law model used here is the use of weighted non-linear (or combined) allometric models, for which there is a considerable literature (e.g., Brown et al. 1989; Bi and Hamilton 1998; Parresol 1999; Ritson and Sochacki 2003; Bi et al. 2004; Morote et al. 2012). However, to our knowledge, the accuracy of these various approaches to overcoming traditional problems associated with back transformation of allometric relationships has not been explored in detail and, consequently, log-log models remain the most common form of allometric models used in forest science.

For practical use, any model predictions and model estimates computed on the logarithmic scale must be back-transformed to the original, plant-biomass scale. Because this transformation is non-linear, and there is variability in the observed data around the fitted relationship, a simple 'naive' exponential-based transformation will generate bias (e.g., Finney 1941). Consequently, correction factors are typically calculated to remove this bias when back-transforming.

For tree allometrics, several estimates of average biomass have been commonly used, including the residual maximum likelihood (REML) estimate, also known as the Baskerville estimate (Baskerville 1972), Duan's smearing estimate (Duan 1983), and the Snowdon correction factor or ratio estimate (Snowdon 1991). Snowdon's ratio estimate has also had an impact in the forestry literature (e.g. Búrquez and Martínez-Yrízar 2011) although, to our knowledge, its statistical efficiency has not been previously tested. Smith (1993) and Hui et al. (2010) review many of the

earlier correction factors in the field of allometry. Within statistics, several other estimates have been proposed (e.g., Finney 1941; Bradu and Mundlak 1970; El-Shaarawi and Viveros 1997; Shen and Zhu 2008), but they have not been assessed in terms of their prediction properties for the purpose of tree allometrics.

The ultimate goal of applying allometrics is the prediction of biomass for new trees, and so these correction factors need to be evaluated in terms of their prediction performance. Optimal predictors are those that minimise the mean squared prediction error (MSPE). Therefore, the aim of this paper is to compare the bias and variability of several possible predictors of forest biomass, including those commonly used in log-log allometric relationships. Our analysis is based on large datasets of above-ground biomass and stem diameter for trees and shrubs in the genera *Eucalyptus* and *Acacia*.

Methodology

Statistical formulation of the problem of estimating biomass

We write our regression model for log-biomass using matrix notation, such as may be found in Shen and Zhu (2008). Matrix notation is helpful as the matrix formulae are the same whether we have one, two, or more covariates. The regression-model setup states that the log-biomass response Y for a collection of n trees is related to the covariates X via the equation:

$$Y = X\beta + \varepsilon$$

where ε is an n-dimensional vector of independent and identically distributed mean-zero normal random variables with variance σ^2; X is a matrix of dimension n by (p+1); $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ is the (p+1)-dimensional vector of fixed effects; $Y = (Y_1, ..., Y_n)^T$ is the n-dimensional vector of log-biomass data; and T denotes the matrix-transpose operation. This model has m = n-(p+1) degrees of freedom, and the covariates encoded within X include constant term and p other variables that may include such variables as log height, log basal area, age, or binary dummy variables associated with species, for example.

The ordinary-least-squares (OLS) estimate for β is $\hat{\beta} = (X^TX)^{-1} X^T Y$, and this has a normal distribution, $N(\beta, \sigma^2 (X^TX)^{-1})$. Based on the assumptions made so far, the OLS estimate for β is unbiased. The residual sum of squares (RSS) is related to a chi-squared random variable; that is, $RSS = Y^T[I - X(X^TX)^{-1}X^T]Y$ has a scaled chi-squared distribution, $\sigma^2 \chi_m^2$, with expected value $E(RSS) = m\sigma^2$. Note also that $\hat{\beta}$ and RSS are independent, a result that plays a key role in deriving the expected values of biomass on the original scale.

The maximum likelihood (ML) estimate for σ^2, $\widehat{\sigma^2}_{ML} = RSS/n$, is biased. The REML estimate for σ^2, $\widehat{\sigma^2}_{REML} = RSS/m = s^2$, is unbiased (Patterson and Thompson 1971; Harville 1977). The ML estimate of β is the same as the OLS estimate $\hat{\beta}$, and it is unbiased.

For a new tree whose covariate value is x_0, the model states that the log-biomass is $Y_0 = x_0^T\beta + \varepsilon_0$. Thus, Y_0 has a normal distribution, $N(x_0^T\beta, \sigma^2)$, with mean $x_0^T\beta$ and variance σ^2. When transformed back to the original scale, the biomass is $\exp(Y_0) = \exp(x_0^T\beta + \varepsilon_0)$, which has mean $\mu(x_0) = \exp(x_0^T\beta + \frac{1}{2}\sigma^2)$ and variance $(\exp(\sigma^2) - 1)\exp(2x_0^T\beta + \sigma^2)$.

Allometrics is an attempt to solve two related tasks. The first is the estimation of the average biomass of trees whose covariate value is x_0, that is, the estimation of the (constant) value $\mu(x_0)$. The second is the prediction of the biomass for a specific new tree whose covariate value is x_0, that is the prediction of the (random) value $\exp(Y_0)$. We perform estimation and prediction based on our fitted regression model, and hence our estimates and predictors are both random quantities with means and variances. The difference between estimation of a fixed quantity (e.g., the average biomass of trees with covariate x_0) and the prediction of a random quantity (e.g., the biomass of a specific tree with covariate x_0) is subtle. Of course the numerical magnitude of the estimates and predictors are the same; the only difference lies in their variability. Predicting a random variable generally leads to greater variability.

Estimates of biomass

Because we use regression to estimate expected log-biomass, a naive estimate for the expected biomass is a direct back-transformation of that value to the original scale. The value $x_0^T\hat{\beta}$ is an unbiased estimate of the constant $x_0^T\beta$, which is equal to $E(Y_0 | x_0)$. However, $\exp(x_0^T\hat{\beta})$ is not an unbiased estimate of $\exp(x_0^T\beta + \frac{1}{2}\sigma^2) = \mu(x_0) = E(\exp(Y_0) | x_0)$, the expected biomass of a tree with covariate value x_0. We call $x_0^T\hat{\beta}$ the 'naive' estimate because it is the first estimate one might try, but it is biased. All corrections for this bias that we have found within the scientific literature involve multiplicative correction factors, C_{Est}, so that the estimates are all of the form:

$$\hat{?}_{Est}(x_0) = C_{Est} \cdot \exp(x_0^T\hat{\beta})$$

Then $\hat{?}_{Naive}(x_0)$ corresponds to the special case of $C_{Naive} = 1$.

One commonly used estimate of this form is based on the OLS estimate for β and the REML estimate for σ^2 plugged into $\mu(x_0) = \exp(x_0^T\beta + \frac{1}{2}\sigma^2)$. As such, the correction factor for this REML-based estimate is:

$$C_{REML} = \exp(\tfrac{1}{2}s^2)$$

and this straightforward approach can go a long way towards correcting the bias of the naive estimate; it has been used in a forestry setting by Baskerville (1972).

Two commonly used estimates in forestry for biomass estimation are the ratio estimate and the smearing estimate. The ratio estimate was first proposed by Snowdon (1991), and its correction factor is:

$$C_{Ratio} = \frac{1^T \exp(Y)}{1^T \exp(X\hat{\beta})}$$

The smearing estimate was first described by Duan (1983), and its correction factor is:

$$C_{Smear} = \frac{1^T \exp(Y - X\hat{\beta})}{n}$$

One of the problems with the REML-based estimate is that it does not take into account the uncertainty in the estimate s^2 for σ^2. Finney (1941) proposed an estimate that addressed this shortcoming when estimating a term such as $\exp(\mu + \frac{1}{2}\sigma^2)$, which is a slight simplification of what we aim to estimate here; see also Heien (1968). In our context, the correction factor for Finney's estimate is:

$$C_{Finney} = \exp(\tfrac{1}{2}s^2) \cdot \left(1 - s^2\left(\frac{s^2+2}{4n}\right) + s^4\left(\frac{3s^4+44s^2+84}{96n^2}\right)\right)$$

This correction factor, like the others we have seen so far, is a constant value and does not vary with x_0, the covariate of the tree whose biomass we wish to estimate. In reality, the variability in predicted log-biomass values changes with the covariates due to uncertainty in the estimate of β, which affects our estimate of the mean log-biomass. Notice that $\mathrm{var}(x_0^T\hat{\beta}) = x_0^T(X^TX)^{-1}x_0\,\sigma^2$, which we shorten to $v(x_0).\sigma^2$, as this expression is used in the four remaining correction factors. The additional source of uncertainty due to $\hat{\beta}$ was taken into account by Bradu and Mundlak (1970), who derived the uniform minimum variance unbiased (UMVU) estimate for $\mu(x_0)$. The UMVU estimate has correction factor:

$$C_{UMVU}(x_0) = {}_0F_1\left(\frac{m}{2}; \frac{m(1-v(x_0))}{4}s^2\right)$$

where ${}_0F_1$ is a confluent hypergeometric function, and recall that $m = n-(p+1)$ is the number of degrees of freedom associated with the regression. For simple linear regression the number of parameters is $p+1=2$, and hence $m=n-2$. Software for evaluating confluent hypergeometric functions is available in the GNU Scientific Library (GSL); see Galassi et al. (2009). Hankin (2006) provides software for linking the GSL library to the R programming environment (R Core Team, 2012), which may be a more accessible way for statisticians to evaluate this estimate.

Perhaps because of a perceived difficulty in evaluating the confluent hypergeometric function, other researchers have looked for easy-to-evaluate estimates that approximate the UMVU estimate. El-Shaarawi and Viveros (1997) proposed an estimate (EV) with correction factor:

$$C_{EV}(x_0) = \exp\left(\frac{1-v(x_0)}{2}s^2 - \frac{1}{4m}s^4 - \frac{1}{6m^2}s^6\right)$$

Shen and Zhu (2008) proposed two alternative estimates that are respectively designed to minimise mean squared error (MM) and minimise bias (MB) within a particular class of estimates. The correction factors for these estimates are:

$$C_{MM}(x_0) = \exp\left(\frac{ms^2}{2(m+2+3nv(x_0))+ 3s^2}\right)$$

and

$$C_{MB}(x_0) = \exp\left(\frac{ms^2}{2(m+nv(x_0))+ s^2}\right)$$

We have included an appendix to this paper that contains the R code required to fit a log-log allometric model and make predictions using each of these correction factors (see Appendix 1). We include code to model the volume of 31 cherry trees using diameter and height as covariates. This classic statistical dataset is available through the SMIR package (Aitken et al., 2012) and may be familiar to many statisticians due to its inclusion in Ryan et al. (1976).

Prediction of biomass

We also use regression to predict log-biomass of new trees, and our predictor can be back-transformed to form a predictor of the biomass of these trees. Consider a new tree with biomass $\exp(Y_0)$ and covariate x_0. Then the value $x_0^T\hat{\beta}$ is an unbiased predictor of $Y_0 = x_0^T\beta + \varepsilon_0$, since both random quantities have the same expected value, namely $x_0^T\beta$. However, $\exp(x_0^T\hat{\beta})$ is not an unbiased predictor of $\exp(x_0^T\beta + \varepsilon_0)$, the biomass of this new tree, since we saw earlier that the expected value of biomass for this tree is $\exp(x_0^T\beta + \frac{1}{2}\sigma^2)$.

Now, any estimate of biomass can also be used as a predictor of biomass. The distinction between the two lies in their errors. The estimation error is:

$$\hat{?}_{Est}(x_0) - \exp(x_0^T\beta + \tfrac{1}{2}\sigma^2),$$

whereas the prediction error is:

$$\hat{?}_{Est}(x_0) - \exp(x_0^T\beta + \varepsilon_0),$$

which is more variable due to the second source of variability, namely ε_0. The averages of these quantities are the bias and prediction bias, respectively.

Datasets

Datasets of total above-ground biomass (kg dry matter (DM)) and stem diameter (cm) for eight species of trees and shrubs were used. The shrub species were *Acacia calamifolia* Sweet ex Lindl., *A. hakeoides* A. Cunn. ex Benth., and *A. pycnantha* Benth.; and the tree species were *Eucalyptus loxophleba* Benth., *E. melliodora* A.Cunn. ex Schauer, *E. occidentalis* Endl., *E. spathulata* Hook., and *E. viminalis* Labill. The species were harvested from a range of revegetation sites in southern Australia; a detailed description of the measurement methods is given in Paul et al. (2012). There was an average of 163 individuals per species, ranging from 102 to 365 individuals. Stem diameters for trees were diameter at breast (130cm) height (DBH) and for shrubs were diameter at 10 cm

height (D10). After the diameters were recorded, the trees/shrubs were harvested to measure biomass as described in Paul et al. (2012).

We fitted simple-linear-regression models for each species using Y = log (above-ground biomass) as the response variable and x = log (stem diameter) as the covariate (Table 1). Log (stem diameter) explains over 90% of the variation in log (above-ground biomass) in each dataset (Table 1). Our estimates of error standard deviations (s) for each species range from 0.25 to 0.45. Assuming σ lies within this range, the magnitude of biases we are adjusting for are between 3 and 10% of biomass. Figure 1 is a plot of the data for the *E. viminalis* dataset, showing the strength of the linear relationship between predictor and response on the log-log scale. Regression diagnostics were performed, and no evidence was found to reject the model assumptions of normality and constant variance of the errors.

Comparison of estimates

To compare the different estimates, we conducted a simulation study calibrated to look like the datasets specified in the previous section. It is necessary to perform a simulation study to evaluate the performance of the correction factors because true biomass values are known and the model assumptions are met. A comparison based on the single datasets will not allow us to validate the variability of the estimates and predictions.

For each tree/shrub species we simulated a new dataset for the given set of stem diameters using a simple-linear-regression model with the regression coefficients and error standard deviations listed in Table 1. We then proceeded to estimate these regression coefficients, β_0 and β_1, and the error standard deviation, σ, from the simulated data. Next, we computed the nine biomass estimates at each diameter value in that dataset. Finally, these estimates were compared with the known expected biomass $\mu(x_0)$, for each x_0, which was computed using the known regression coefficients listed in Table 1; the error of the biomass estimate, namely the difference between the estimated biomass and the known biomass, was obtained. We carried out 10,000 simulations, and the bias (average error), standard deviation (SD) of the errors, and mean squared error ($MSE = SD^2 + bias^2$) for the estimates, obtained by averaging across the 10,000 simulations, were recorded for each tree/shrub (i.e., for each x_0). We further averaged these summary statistics across individuals for each species. This enabled comparisons of the estimates across a range of stem diameters within each species as well as more general comparisons across species.

Comparison of predictors

The estimates of expected biomass, C_{Est} .exp($x_o^T\hat{\beta}$), are often used to predict biomass of new trees, so we evaluated the performance of these predictors by comparing their predicted values to additional simulated biomass values Y_0 with covariate x_0. These additional biomass values were simulated using the true allometric parameters given in Table 1. The prediction bias and MSPE, obtained by averaging the prediction errors and squared prediction errors across the 10,000 simulations, were recorded for each x_0 and then further averaged across individuals for each species.

Results

The magnitude and direction of bias are important. Because of the relationship between bias, variance, and mean squared error, it is natural to make comparisons of the correction factors in terms of the bias^2, MSE, (prediction bias)^2, and MSPE values averaged across all trees/shrubs for each dataset for nine different correction factors; see Table 2. The directions of bias are examined in Figure 2B. The need for a bias correction is evident through the consistently poor performance of the Naive correction factor, both in terms of estimation and in terms of prediction. For all estimates, the Bias^2 value is a minor component of the MSE, except for the Naive estimate. The MSE values for the Naive estimates are always larger than the others. An examination of MSE and MSPE values shows that Snowdon's Ratio correction factor also performs poorly. In what follows, we focus primarily on the remaining seven correction factors.

The overall performance of the seven correction factors as estimates and predictors can be evaluated using Table 2. The differences in performance of the correction factors are subtle and patterns in performance are more evident in graphical displays. For simplicity, here we present plots for the *E. viminalis* dataset only (Figure 2), allowing comparison of the MSE and MSPE values (Figure 2A) and the Bias and Prediction Bias (Figure 2B) for the seven correction factors. The pattern and order of correction factors found for this dataset are similar to those for the other seven datasets (see Appendix 2). The correction factor that consistently gives the smallest MSE and MSPE values in all cases is the MM correction factor. In terms of bias, the MM correction factor is on par with, or slightly larger in magnitude than, the MB, EV, and UMVU correction factors. In all examples, there is very little difference between the MB, EV, and UMVU correction factors. The bias of the MM correction factor indicates that it slightly underestimates biomass. The Ratio, Finney, and Smear correction factors have higher MSE and MSPE values in each example, and their biases indicate that they slightly overestimate biomass each time.

Figure 3 presents plots of the bias (Figure 3A), the relative standard deviation (Figure 3B) and the relative MSE (Figure 3C) of the estimates as a function of diameter for the *E. viminalis* dataset. The

biases of four of the seven correction factors move away from zero as the stem diameters increase, however the bias of the MB, EV, and UMVU correction factors stay close to zero (Figure 3A). Again we observe no visible difference between the MB, EV, and UMVU correction factors in terms of bias at any diameter, and any observed difference between the UMVU's bias and zero is due to the fact that Figure 3A is based on a simulation.

Using the UMVU estimate of biomass as our basis for comparison, we examine the standard deviations of our estimates relative to the standard deviation of the UMVU estimate (Figure 3B). The REML, Finney, and Smear estimates are more variable than the MB, EV, and UMVU estimates. The MM estimate is less variable than the UMVU estimate.

Finally, we examine the MSE to obtain an overall evaluation of the estimates relative to the MSE of the UMVU estimate (Figure 3C). MSE acts as a natural overall method for comparing these estimates as it combines both bias (Figure 3A) and variability (Figure 3B) into a single measure. The MM estimate has consistently lower MSE values at each diameter compared with the UMVU estimate (Figure 3C). The patterns shown in Figure 3 are also apparent for the other species.

Discussion

Our study resulted in two main findings. First, only one estimate, the UMVU estimate, was truly unbiased, but the EV and MB estimates gave almost identical performance whilst avoiding the need to evaluate a confluent hypergeometric function (Figures 2, 3). Second, the MM correction factor removed the bulk of the gross bias and performed better in terms of prediction; the MM correction factor had bias of slightly larger magnitude but also had the lowest MSPE compared to other correction factors (Figure 3). MM underestimated biomass but the magnitude of this bias was small, making up less than 1% of the MSPE.

These results have implications for the application of tree allometrics, as the correction factors found to perform best here are not currently in general use. Three of the estimates tested here, the REML, Smearing, and Snowdon's Ratio estimates, have been commonly used in estimation of forest biomass. Previous work, including Lambert et al. (2005), who also cited criticisms raised by Flewelling and Pienaar (1981) and Hepp and Brister (1982), has criticised the REML estimate about the magnitude of its bias, especially for small sample sizes. In agreement with this, we found that this correction factor overestimated biomass, more so for larger individuals (Figure 3A). A comparison of estimates using simulated sampling studies of *Pinus radiata* datasets (Snowdon 1991) found that, in most cases, the Ratio estimate gave less biased and/or more accurate estimates of total biomass than the REML and Finney estimates. Further, the Ratio estimate was found to have lower bias and increased accuracy at the plot level than the Smearing and REML estimates when

validated against whole-plot harvests of woody vegetation in the Sonoran Desert (Búrquez and Martínez-Yrízar 2011). Based on its low bias relative to these two other correction factors, Snowdon's Ratio correction factor has been commonly applied in the field of tree allometrics (e.g., Montagu et al. 2005; Paul et al. 2008). Here we found that although Snowdon's Ratio estimate performed better than several of the other estimates in terms of bias (see Table 2), as a predictor it performed relatively poorly. Based on our investigations, we suggest that four of the correction factors considered here, and which are not currently used in forest science, provide better alternatives in terms of their ability to predict biomass.

Although we assess correction factors in terms of their performance as estimates and predictors of total above-ground biomass of trees and shrubs using allometric equations, this assessment also applies to other logarithmic regressions including, for example, allometric equations for the prediction of leaf area (Marshall and Waring, 1986). Much of the previous work (e.g., Smith 1993) has focussed on correcting the bias associated with estimates of the average response variable (in our case, tree biomass) from log-log allometric equations. However, as prediction of the response variable (in our case, tree biomass) is the ultimate goal here, we have also evaluated correction factors in terms of their ability to predict.

In this study, we tested datasets covering only relatively small trees (all stem diameters <42 cm, with the exception of *E. occidentalis*, with three of the 118 stem diameters lying in the range 42-79 cm), where the commonly used log-log allometric equations are relatively robust, provided they are based on a sufficient number of individuals. However, when relatively large trees with relatively high variability in biomass due to factors such as hollows or decayed wood are included, linear allometric equations developed on transformed data may not give due weighting to larger trees, which hold most of the biomass for a given site (e.g., see Brown et al. 1989; Roxburgh et al. 2006; Kuyah et al. 2012). In such cases, the application of other allometric models may provide a better alternative. For example, other workers (Parresol 2001; Lambert et al. 2005) have found that modelling the error structure on the original data scale gives results as good as or even better than applying a transformation. Also, weighted non-linear (or combined) allometric models have been used recently by numerous workers (e.g., Brown et al. 1989; Parresol 1999; Bi et al. 2004; Morote et al. 2012). Further work is required to test the accuracy of these various approaches to overcoming traditional problems associated with back transformation of allometric relationships.

The common use of logarithmic regression in allometry has been questioned by several researchers in situations where the ultimate goal is not prediction but the description of the relationship between variables (e.g., Warton et al. 2006, Packard et al. 2010). Warton et al. (2006) examined allometrics through a bivariate model of biomass and diameter; that is, their model takes

errors in the measurement of diameters into account. More recently, Warton et al. (2012) released software that offers two alternatives to OLS regression for finding the regression line of best fit, namely 'major axis' and 'standardised major axis'. Warton et al. (2006) explicitly highlight that regression is preferred when prediction is the purpose of line-fitting.

Conclusions

When unbiased estimates are required, the UMVU correction factor is an obvious choice, and the MB and EV correction factors give almost identical performance whilst avoiding the need to evaluate a confluent hypergeometric function.

When unbiased estimates and predictions are not strictly required, there is one alternative correction factor, the MM correction factor, which should be considered. As a predictor it removes the bulk of the gross bias and it has superior performance in terms of MSPE. The MM correction factor has slightly larger bias than the MB, EV and UMVU correction factors, but it has the lowest MSPE among all nine correction factors that we considered. Its bias is slightly negative, but small; the square of its bias makes up less than 1% of the MSPE. Therefore, when predicting biomass of new trees, we recommend the use of the MM correction factor.

Acknowledgements

We thank the CSIRO Sustainable Agriculture Flagship for financial support of this work. The datasets used in this study were collected as part of a broader project on improving carbon accounting in mixed-species environmental plantings, funded largely by the Department of Climate Change and Energy Efficiency with additional financial support provided by the Victorian Department of Sustainability and Environment. Collaborators in the collection and analysis of biomass datasets included Simon Murphy, Jaymie Norris, Trevor Hobbs, Craig Neumann, Zoe Read, Laura Kmoch, Geoff McArthur, Tom Fairman, Rob Law, Benjamin Finn, Mark Brammar, and Geoff Minchin. For reviewing the manuscript, we thank Natalie Kelly, A.O. Nicholls, and two anonymous referees.

References

Aitkin, M., Francis, B., Hinde, J. and Darnell, R., 2012. SMIR: Companion to Statistical Modelling in R. R package version 0.02. http://CRAN.R-project.org/package=SMIR

Baskerville, G.L., 1972. Use of logarithmic regression in the estimation of plant biomass. Can. J. For. Res. 2, 49-53.

Bi, H., Hamilton, F., 1998. Stem volume equations for native tree species in southern New South Wales and Victoria. Aust. For. 61, 275-286.

Bi, H., Turner, J., Lambert, M.J., 2004. Additive biomass equations for native eucalypt forest trees of temperate Australia. Trees 18, 467–479.

Bradu, D., Mundlak, Y., 1970. Estimation in lognormal linear models. J. Am. Stat. Assoc. 65, 198-211.

Brown, S., Gillespie, A.J.R., Lugo, A.E., 1989. Biomass estimation methods for tropical forests with applications to forest inventory data. For. Sci. 35, 881–902.

Búrquez, A., Martínez-Yrízar, A., 2011. Accuracy and bias on the estimation of aboveground biomass in the woody vegetation of the Sonoran Desert. Botany 89, 625-633.

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. J. Am. Stat. Assoc. 78, 605-610.

El-Shaarawi, A.H., Viveros, R., 1997. Inference about the mean in log-regression with environmental applications. Environmetrics 8, 569-582.

Finney, D.J., 1941. On the distribution of a variate whose logarithm is normally distributed. Suppl. to the J. R. Stat. Soc., 7, 155-161.

Flewelling, J.W., Pienaar, L.V., 1981. Multiplicative regression with lognormal errors. For. Sci. 27, 281-289.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi. F., 2011. GNU Scientific Library Reference Manual, third ed. http://www.gnu.org/software/gsl/.

Hankin, R.K.S., 2006. Special functions in R: Introducing the GSL package. R News, 6.

Harville, D., 1977. Maximum-likelihood approaches to variance-component estimation and to related problems. J. Am. Stat. Assoc. 72, 320-338.

Heien, D.M., 1968. A note on log-linear regression. J. Am. Stat. Assoc. 63, 1034-1038.

Hepp, T.E., Brister, G.H., 1982. Estimating crown biomass in loblolly pine plantations in the Carolina flatwoods. For. Sci. 28, 115–127.

Hui, C., Terblanche, J.S., Chown, S.L., McGeoch, M.A., 2010. Parameter landscapes unveil the bias in allometric prediction. Methods Ecol. Evol. 1, 69-74.

Kuyah, S., Dietz, J., Muthuri, C., Jamnadass, R., Mwangi, P., Coe, R., Neufeldt, H., 2012. Allometric equations for estimating biomass in agricultural landscapes: I. Aboveground biomass. Agric. Ecosyst. Environ. 158, 216-224.

Lambert, M.C., Ung, C.H., Raulier, F., 2005. Canadian national tree aboveground biomass equations. Can. J. For. Res. 35, 1996-2018.

Marshall, J.D., Waring, R.H., 1986. Comparison of Methods of Estimating Leaf-Area Index In Old-Growth Douglas-Fir. Ecology, 67, 975-979.

Montagu, K.D., Dűttmer K., Barton, C.V.M., Cowie, A.L., 2005. Developing general allometric relationship for regional estimates of carbon sequestration—an example using Eucalyptus pilularis from seven contrasting sites. For. Ecol. Manage. 204, 113–127.

Morote, G.F.A, López Serrano, F.R., Andrés, M., Rubio, E., González Jiménez, J.L., de las Heras, J., 2012. Allometries, biomass stocks and biomass allocation in the thermophilic Spanish juniper woodlands of Southern Spain. For. Ecol. Manage. 270, 85–93.

Packard, G.C., Birchard, G.F., Boardman, T.J., 2011. Fitting statistical models in bivariate allometry. Biol. Rev. 86, 549-563.

Parresol, B.R., 1999. Assessing tree and stand biomass: A review with examples and critical comparisons. For. Sci. 45, 573-593.

Parresol, B. R., 2001. Additivity of nonlinear biomass equations. Can. J. For. Res. 31, 865-878.

Patterson, H., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545-554.

Paul, K.I., Jacobsen, K., Koul, V., Leppert, P., Smith, J., 2008. Predicting growth and sequestration of carbon by plantations growing in regions of low-rainfall in southern Australia. For. Ecol. Manage. 254, 205-216.

Paul, K., Roxburgh, S., Raison, J., Larmour, J., England, J., Murphy, S., Norris, J., Ritson, P., Brooksbank, K., Hobbs, T., Neumann, C., Lewis, T., Read, Z., Clifford, D., Kmoch, L., Rooney, M., Freudenberger, D., Jonson, J., Peck, A., Giles, R., Bartle, J., McAurthur, G., Wildy, D., Lindsay, A., Preece, N., Cunningham, S., Powe, T., Carter, J., Bennett, R., Mendham, D., Sudmeyer, R., Rose, B., Butler, D., Cohen, L., Fairman, T., Law, R., Finn, B., Brammar, M., Minchin, G., van Oosterzeeand, P., A. Lothian, 2012. Improved estimation of biomass accumulation by environmental plantings and mallee plantings using FullCAM. Report for Department of Climate Change and Energy Efficiency. CSIRO Sustainable Agriculture Flagship, Canberra. 93 pp.

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ritson, P., Sochacki, S., 2003. Measurement and prediction of biomass and carbon content of Pinus pinaster trees in farm forestry plantations, south-western Australia. For. Ecol. Manage. 175, 103-117.

Roxburgh, S.H., Wood, S.W., Mackey, B.G., Woldendorp, G., Gibbons, P. 2006. Assessing the carbon sequestration potential of forested ecosystems: a case study from temperate Australia. J. Appl. Ecol. 43, 1149-1159.

Ryan, T. and Joiner, B. and Ryan, B., 1976. Minitab Students Handbook, Duxbury Press, North Scituate, Mass

Shen, H., Zhu, Z., 2008. Efficient mean estimation in log-normal linear models. J. Stat. Plan. Inference 138, 552-567.

Smith, R.J., 1993. Logarithmic transformation bias in allometry. Am. J. Phys. Anthropol. 90, 215-228.

Snowdon, P., 1991. A ratio estimator for bias correction in logarithmic regressions. Can. J. For. Res. 21, 720-724.

Warton, D.I., Wright, I.J., Falster, D.S., Westoby, M., 2006. Bivariate line-fitting methods for allometry. Biol. Rev. 81, 259-291.

Warton, D.I., Duursma, R.A., Falster, D.S., Taskinen, S., 2012. Smatr 3 - an R package for estimation and inference about allometric lines. Methods Ecol. Evol. 3, 257-25

List of Figures

Figure 1:

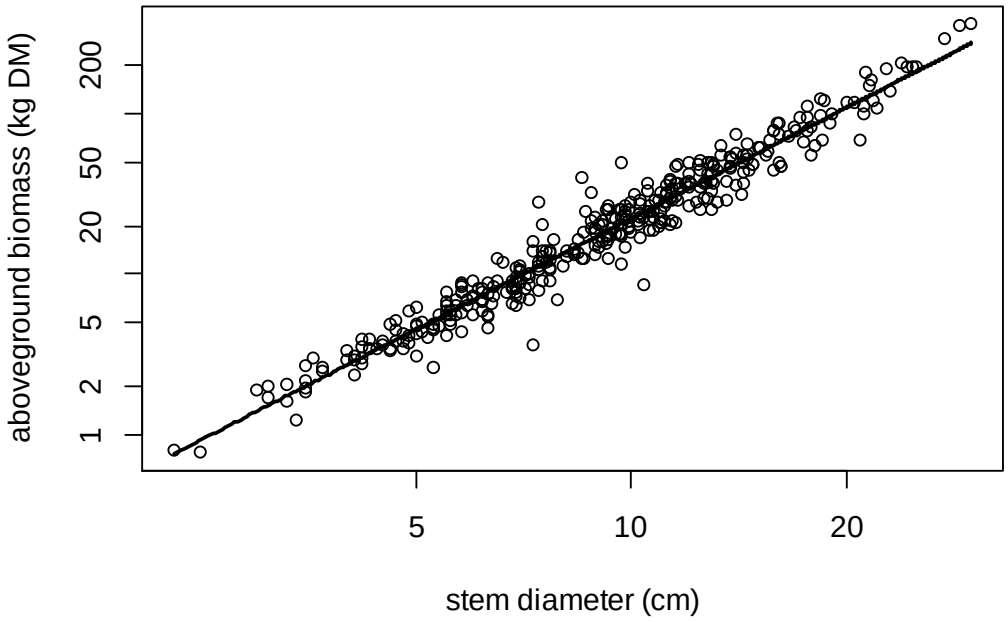
aboveground biomass (kg DM)
1
2
5
20
50
200
5
10
20
stem diameter (cm)

Figure 2:

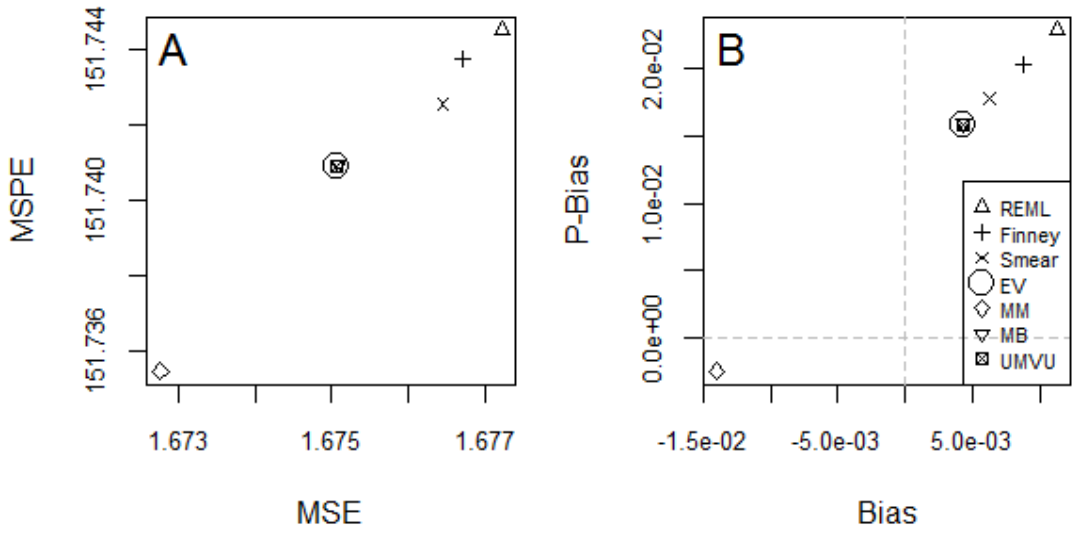
A
MSPE
151.744
151.740
151.736
1.673
1.675
1.677
MSE
B
P-Bias
2.0e-02
1.0e-02
0.0e+00
-1.5e-02
-5.0e-03
5.0e-03
Bias
REML
Finney
Smear
EV
MM
MB
UMVU

Figure 3:

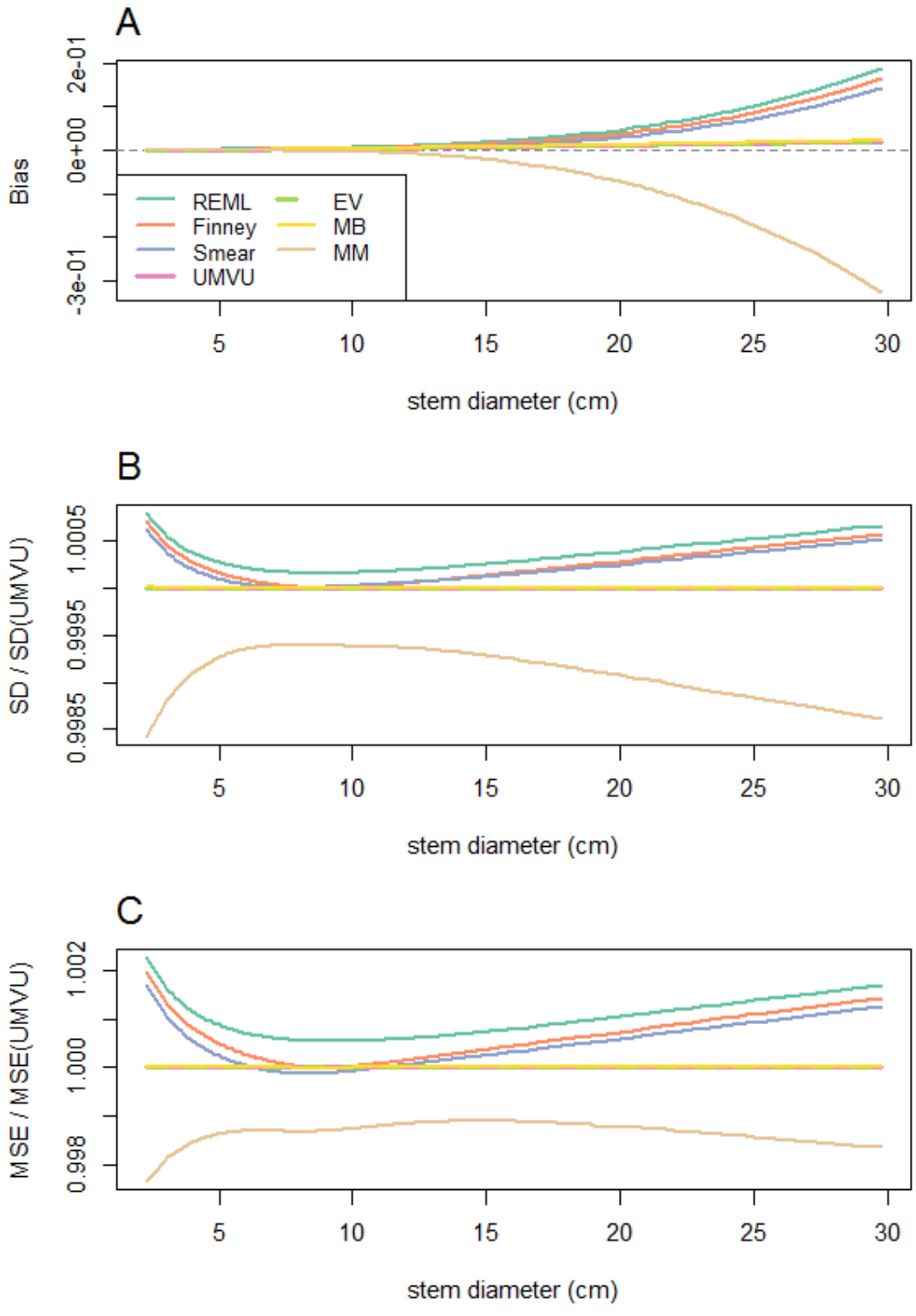
A
Bias
2e-01
0e+00
-3e-01
REML
Finney
Smear
UMVU
EV
MB
MM
5
10
15
20
25
30
stem diameter (cm)
B
SD / SD(UMVU)
1.0005
0.9995
0.9985
5
10
15
20
25
30
stem diameter (cm)
C
MSE / MSE(UMVU)
1.002
1.000
0.998
5
10
15
20
25
30
stem diameter (cm)

Table 1: Summary statistics for the diameters and biomass values and the regression on the log-log scale for each of the eight datasets, including the number of shrubs and trees (n), the minimum (Min), maximum (Max), mean, and standard deviation (StDev) for the diameters and biomass, and least squares estimates for intercept and slope, the root mean squared error (s), the number of degrees of freedom (m=n-2), and the proportion of variation of the response explained by the predictor (R^2) for each species. DBH is stem diameter at breast (130cm) height; D10 is stem diameter at 10 cm.

		Stem diameter	Diameter Summary (cm)				Biomass Summary (kg DM)				Regression Summary				
	n		Min	Max	Mean	StDev	Min	Max	Mean	StDev	Intercept $\hat{\beta}_0$	Slope $\hat{\beta}_1$	s	m=n-2	R^2
Acacia calamifolia	128	D10	1.0	15.9	6.0	3.2	0.113	87	12.2	14.9	-2.23	2.41	0.360	126	0.939
A. hakeoides	113	D10	0.5	21.3	6.8	4.0	0.059	123	10.6	15.4	-2.10	2.10	0.358	111	0.946
A. pycnantha	102	D10	1.0	23.8	9.2	4.7	0.114	130	26.4	27.0	-2.37	2.36	0.450	100	0.927
Eucalyptus loxophleba	104	DBH	0.2	29.0	8.4	5.0	0.122	191	30.9	37.0	-0.779	1.82	0.366	102	0.930
E. melliodora	169	DBH	1.8	39.2	7.7	6.0	0.608	698	27.5	71.6	-1.73	2.12	0.326	167	0.939
E. occidentalis	118	DBH	2.3	79.0	11.7	11.5	1.040	6710	178.0	765.0	-2.12	2.43	0.240	116	0.979
E. spathulata	206	DBH	2.3	41.3	9.7	5.3	2.270	966	60.2	90.8	-1.30	2.22	0.251	204	0.954
E. viminalis	365	DBH	2.3	29.8	10.0	4.9	0.769	365	32.0	44.5	-2.19	2.30	0.242	363	0.954

Table 2: Mean squared error (MSE), squared bias (Bias^2), mean square prediction error (MSPE), and squared prediction-bias (P-Bias^2) values for each estimate, averaged across all trees for each dataset, where MSE = Variance + Bias^2.

		A. calamifolia	*A. hakeoides*	*A. pycnantha*	*E. loxophleba*	*E. melliodora*	*E. occidentalis*	*E. spathulata*	*E. viminalis*
MSE	Naive	2.60	1.690	22.0	14.70	23.0	2180	39.6	3.57
	Ratio	1.54	1.080	11.1	9.38	21.3	3920	34.8	1.90
	REML	1.29	0.871	8.83	7.77	16.3	2040	28.9	1.68
	Finney	1.29	0.869	8.79	7.76	16.3	2040	28.9	1.68
	Smear	1.29	0.868	8.78	7.75	16.3	2040	28.9	1.68
	EV	1.28	0.865	8.74	7.73	16.2	2020	28.8	1.68
	MM	1.28	0.860	8.63	7.70	16.1	2010	28.7	1.67
	MB	1.28	0.866	8.74	7.73	16.2	2030	28.8	1.68
	UMVU	1.28	0.865	8.74	7.73	16.2	2020	28.8	1.68
		A. calamifolia	*A. hakeoides*	*A. pycnantha*	*E. loxophleba*	*E. melliodora*	*E. occidentalis*	*E. spathulata*	*E. viminalis*
Bias^2	Naive	0.618	0.391	6.93	3.64	1.55	17.4	3.54	0.798
	Ratio	7.12E-05	3.53E-06	5.37E-04	1.16E-03	3.89E-05	8.51E-02	1.32E-04	2.86E-05
	REML	3.11E-04	2.15E-04	5.97E-03	5.77E-05	1.64E-03	7.45E-02	1.58E-03	1.28E-04
	Finney	1.15E-04	7.14E-05	2.16E-03	1.61E-04	1.04E-03	5.46E-02	8.99E-04	7.63E-05
	Smear	1.61E-05	7.17E-06	3.71E-04	9.96E-04	5.95E-04	3.84E-02	4.13E-04	3.95E-05
	EV	2.96E-06	2.58E-06	4.38E-04	1.16E-03	1.08E-05	2.21E-04	6.28E-05	1.85E-05
	MM	1.54E-03	1.08E-03	1.77E-02	2.07E-02	5.42E-03	0.236	4.45E-03	1.96E-04
	MB	3.37E-06	2.85E-06	4.48E-04	1.14E-03	1.83E-05	6.14E-06	7.17E-05	1.88E-05
	UMVU	3.05E-06	2.65E-06	4.45E-04	1.15E-03	1.10E-05	2.16E-04	6.30E-05	1.85E-05

MSPE		*A. calamifolia*	*A. hakeoides*	*A. pycnantha*	*E. loxophleba*	*E. melliodora*	*E. occidentalis*	*E. spathulata*	*E. viminalis*
	Naive	58.108	36.637	408.13	297.78	413.32	24688	956.22	153.55
	Ratio	56.975	36.083	397.71	292.42	412.2	26459	952.63	152
	REML	56.747	35.879	395.03	290.75	407.2	24579	946.2	151.74
	Finney	56.745	35.877	394.99	290.74	407.17	24577	946.19	151.74
	Smear	56.744	35.875	395.01	290.72	407.18	24575	946.16	151.74
	EV	56.742	35.872	394.95	290.71	407.04	24561	946.07	151.74
	MM	56.738	35.861	394.85	290.69	406.88	24546	945.92	151.74
	MB	56.742	35.872	394.95	290.71	407.05	24562	946.07	151.74
	UMVU	56.742	35.872	394.95	290.71	407.04	24561	946.07	151.74

P-Bias^2		*A. calamifolia*	*A. hakeoides*	*A. pycnantha*	*E. loxophleba*	*E. melliodora*	*E. occidentalis*	*E. spathulata*	*E. viminalis*
	Naive	0.622	0.386	6.86	3.65	1.53	16.1	3.49	0.777
	Ratio	3.27E-05	3.32E-06	1.39E-03	1.42E-03	1.87E-04	1.93E-02	6.02E-04	2.85E-04
	REML	2.22E-04	3.37E-04	8.34E-03	1.61E-05	2.30E-03	0.181	2.79E-03	5.23E-04
	Finney	6.40E-05	1.48E-04	3.67E-03	2.65E-04	1.58E-03	0.149	1.85E-03	4.11E-04
	Smear	1.65E-06	4.07E-05	1.11E-03	1.23E-03	1.01E-03	0.122	1.12E-03	3.18E-04
	EV	9.99E-07	2.82E-05	1.22E-03	1.42E-03	1.15E-04	1.90E-02	4.40E-04	2.51E-04
	MM	1.76E-03	8.53E-04	1.41E-02	2.17E-02	4.38E-03	0.111	2.88E-03	5.97E-06
	MB	7.85E-07	2.90E-05	1.24E-03	1.39E-03	1.38E-04	2.41E-02	4.63E-04	2.53E-04
	UMVU	9.53E-07	2.84E-05	1.24E-03	1.41E-03	1.16E-04	1.90E-02	4.41E-04	2.51E-04