Fitting truncated generalized beta distributions to stem density data from permanent sample plots in Quebec, Canada

Abstract

Statistical models predicting stem diameter distributions have found many applications in forestry. Our objective is to develop a methodology that can be used to derive a stem diameter distribution model for any combination of species and cover type in Quebec, Canada, using readily-available data from the government-run permanent sample plot inventory program. We test 25 truncated distributions from the generalized beta family to a large dataset of stems inventoried from permanent fixed-area plots in the province of Quebec, Canada, using a non-linear least-squares parameter-fitting algorithm. We describe a two-stage parameter-fitting methodology that produces improved estimates of parameter estimation error and parameter correlation for input data with bounded domain. We report best-fit distribution, best-fit parameter estimates (with standard error on parameter estimates), and AICc for each of 30 subdatasets covering the entire province of Quebec (representing all combinations of 10 species groups and 3 cover types). Best-fit results are clearly dominated by the four distributions in the generalized gamma family.

1 Introduction

Stem diameter distributions (i.e. stand tables) have long played an important role in forestry (Bailey and Dell, 1973; Hyink and Moser, 1983). Published models tend to be specific to a given combination of species, stand structure, geographic area, and inventory sampling method. No stem diameter distribution models have been published to date for the province of Quebec, Canada. Furthermore, no generalized methodology has been published to model stem diameter distributions from permanent sample plot (PSP) data, documenting how to correctly estimate best-fit parameter uncertainty and correlations for the common case where observed diameter data has *a priori* bounded domain (e.g. only merchantable stems of a certain minimum diameter are inventoried and trees never grow beyond a certain maximum diameter). The present study fills these gaps in the literature.

The most commonly-used statistical model used to describe stem diameter seems to be the Weibull distribution (Bailey and Dell, 1973; Liu et al., 2002; Cao, 2004; Coomes and Allen, 2007). Other models include the gamma (Nelson, 1964), exponential (Meyer and Stevenson, 1943) and S_B (Johnson, 1949) distributions. The Weibull, gamma and exponential distributions are all derivatives of the generalized gamma distribution, which is itself a member of the of generalized beta family of statistical distributions.

We fit 25 truncated distributions from the generalized beta family to a large dataset of stems from government-compiled permanent fixed-area plots in the province of Quebec, Canada. We describe a two-stage distribution-fitting methodology that correctly handles parameter estimation error and correlations for input data with bounded domain. We present best-fit distributions for 30 combinations of species group and cover type.

Our best-fit distribution results cover all combinations of species and cover

types in Quebec, and could be used directly. Alternatively, our methodology can be easily replicated using readily-available PSP data, for example to derive models for different aggregations of species and cover type, or for a different geographic extent of plot data used as input. The two-stage parameter-fitting methodology is potentially applicable to any context where trunctated data is fitted to standard-form statistical distributions.

The remainder of this paper is organized as follows. We describe our methodology in §2. Results are presented in §3, followed by discussion in §4.

2 Methods

Ducey and Gove (2015) document three parent distributions in the generalized beta family that can be used to derive several other distributions. These parent distributions are the generalized beta distribution of the first kind (GB1), the generalized beta distribution of the second kind (GB2), and the generalized gamma distribution (GG). The probability density functions (PDF) of GB1 and GB2 distributions have the following forms (adapted from Ducey and Gove, 2015)

$$\mathrm{GB1}(x;a,b,p,q) = \frac{|a|x^{ap-1}\left[1-(x/b)^a\right]^{q-1}}{b^{ap}B(p,q)}, \qquad 0 < x^a < b^a, b > 0, p > 0, q > 0 \tag{1}$$

and

$$\mathrm{GB2}(x;a,b,p,q) = \frac{|a|x^{ap-1}x^{q-1}}{b^{ap}B(p,q)\left[1-(x/b)^a\right]^{p+q}}, \qquad a > 0, b > 0, p > 0, q > 0 \tag{2}$$

defined for $x > 0$, where $\mathrm{B}(p,q)$ represents the beta function (not to be confounded with the beta, or generalized beta, distributions), which is given by

$$\mathrm{B}(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt. \tag{3}$$

The PDF of the generalized gamma GG distribution has the following form

$$\mathrm{GG}(x;a,\beta,p) = \frac{ax^{ap-1}e^{-\left(\frac{x}{\beta}\right)^a}}{\beta^{ap}\Gamma(p)}, \qquad a>0, \beta>0, q>0 \tag{4}$$

defined for $x > 0$, where $\Gamma(p)$ represents the gamma function (not to be confounded with the gamma, or generalized gamma, distributions), which is given by

$$\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx. \tag{5}$$

We can define the PDFs for 22 different distributions in the generalized beta family in terms of one of the three parent distributions, as follows (adapted from Ducey and Gove, 2015)

$$\mathrm{IB1}(x;b,p,q) = \mathrm{GB1}(x;-1,b,p,q) \tag{6}$$

$$\mathrm{UG}(x;b,d,q) = \lim_{a\to\infty} \mathrm{GB1}(x;a,b,d/a,q) \tag{7}$$

$$\mathrm{B1}(x;b,p,q) = \mathrm{GB1}(x;1,b,p,q) \tag{8}$$

$$\mathrm{B2}(x;b,p,q) = \mathrm{GB2}(x;1,b,p,q) \tag{9}$$

$$\mathrm{SM}(x;a,b,q) = \mathrm{GB2}(x;a,b,1,q) \tag{10}$$

$$\mathrm{Dagum}(x;a,b,p) = \mathrm{GB2}(x;a,b,p,1) \tag{11}$$

$$\mathrm{Pareto}(x;b,p) = \mathrm{GB1}(x;-1,b,p,1) \tag{12}$$

$$\mathrm{P}(x;b,p) = \mathrm{GB1}(x;1,b,p,1) \tag{13}$$

$$\mathrm{LN}(x;\mu,\sigma) = \lim_{a\to 0} \mathrm{GG}(x;a,(\sigma^2a^2)^{1/a},(a\mu+1)/(\sigma^2a^2)) \tag{14}$$

$$\mathrm{GA}(x;\beta,p) = \mathrm{GG}(x;1,\beta,p) \tag{15}$$

$$\mathrm{W}(x;a,\beta) = \mathrm{GG}(x;a,\beta,1) \tag{16}$$

$$\mathrm{F}(x;u,v) = \mathrm{GB2}(x;1,v/u,u/2,v/2) \tag{17}$$

$$\mathrm{L}(x;b,q) = \mathrm{GB2}(x;1,b,1,q) \tag{18}$$

$$\mathrm{IL}(x;b,p) = \mathrm{GB2}(x;1,b,p,1) \tag{19}$$

$$\mathrm{Fisk}(x;a,b) = \mathrm{GB2}(x;a,b,1,1) \tag{20}$$

$$\mathrm{U}(x;b) = \mathrm{GB1}(x;1,b,1,1) \tag{21}$$

$$\tfrac{1}{2}\mathrm{N}(x;0,\sigma) = \mathrm{GG}(x;2,\sigma^2,1/2) \tag{22}$$

$$\chi^2(x;p) = \mathrm{GG}(x;1,2,p) \tag{23}$$

$$\mathrm{EXP}(x;\beta) = \mathrm{GG}(x;1,\beta,1) \tag{24}$$

$$\mathrm{R}(x;\beta) = \mathrm{GG}(x;2,\beta,1) \tag{25}$$

$$\tfrac{1}{2}\mathrm{t}(x;df) = \mathrm{GB2}(x;2,\sqrt{df},1/2,df/2) \tag{26}$$

$$\mathrm{LL}(x;b) = \mathrm{GB2}(x;1,b,1,1) \tag{27}$$

We use a weighted non-linear least squares (NLLS) algorithm to fit target distributions to PSP inventory data binned into 26 size classes of uniform width W.

The objective function value of the NLLS problem minimizes the sum of squares of the residual terms

$$Z\left(f(x;\hat{\boldsymbol{P}})\right) = \min \sum_{i\in\{I|\hat{y}_i>0\}} e\left(f(x_i;\boldsymbol{P}),\hat{y}_i\right)^2 \tag{28}$$

with

$$e\left(f(x_i;\boldsymbol{P}),\hat{y}_i\right) = w_i\left[f(x_i;\boldsymbol{P}) - \hat{y}_i\right] \tag{29}$$

where x_i is the diameter corresponding to the center of bin $i \in I$, $f(x_i;\boldsymbol{P})$ is the value of the PDF of the target distribution at $x_i \in \boldsymbol{X}$ (given a vector of parameters $\boldsymbol{P}$). $\hat{y}_i \in \hat{\boldsymbol{Y}}$ represents the estimated stem density in bin i, which corresponds to the average of plot-wise stem density measurements.

Note that residual terms are scaled by a weight factor $w_i = 1-\min(E_{\hat{y}_i}\hat{y}_i^{-1}, 1)$, which dampens the impact of $\hat{y}_i$ on Z as a function of the relative margin of error $E_{\hat{y}_i}\hat{y}_i^{-1}$. We cap relative margin of error at 1 (negative values of w_i would have the effect of *rewarding* large residual value $f(x_i;\boldsymbol{P})-\hat{y}_i$, which would make NLLS algorithm results unnecessarily difficult to interpret). Thus, w_i converges to 1 as relative margin of error approaches 0, and $w_i = 0$ if $E_{\hat{y}_i}\hat{y}_i^{-1} \geq 1$. Note that if sampling error is high enough for all bins (due to insufficient sample size), such that $w_i = 0, \forall i \in I$, the objective function value is 0 regardless of values of input data vector $\hat{\boldsymbol{Y}}$ and the NLLS optimisation problem becomes meaningless.

The margin of error corresponds to the product $t\sigma_{\hat{y}_i}$ of the critical t value (with $\alpha = 0.05$ and $|\hat{\boldsymbol{Y}}| - 1$ degrees of freedom) and bin-wise sampling error

$$\sigma_{\hat{y}_i} = \sqrt{\frac{\sum_{j\in J}\left(y_{ij} - \hat{y}_i\right)^2}{|\hat{\boldsymbol{Y}}| - 1}} \tag{30}$$

where y_{ij} corresponds to the observed stem density in bin i in sample plot j (Schreuder et al., 2004).

We normalize our binned data, such that $\sum_{i \in I} W\hat{y}_i = 1$. The domain of input data is bounded, such that $a \leq x_1 - w/2$ and $x_{|I|} + w/2 \leq b$, where $a > 0$. Our dataset intentionally includes only merchantable stems, i.e. $a = 9$, and contains very few stems of diameter greater than 24 inches (61 cm), i.e. $b = 61$.

The integral of the standard forms of the PDFs described above over the interval $[0, \infty]$ is 1 for any given vector of input parameters $\boldsymbol{P}$, that is

$$\int_0^\infty f(x; \boldsymbol{P}) dx = 1. \tag{31}$$

Fitting the standard forms of f to the normalized binned data will generally produce poor fits, as the sum of residuals will be positively biased due to bounded domain (i.e. $\sum_{i \in I} e_i > 1$), with quality of fit inversely proportional to $b - a$. We can obtain a better fit using an augmented PDF $f'(x; \boldsymbol{P'}) = sf(x; \boldsymbol{P})$. The global scaling parameter s effectively relaxes the unity constraint on the integral of f'. Thus, using f', we obtain similar quality fits for any scaling of bin value vector $\hat{\boldsymbol{Y}}$.

The variance $\sigma^2_{\hat{p}_j}$ of best-fit parameter estimator $\hat{p}_j \in \hat{\boldsymbol{P}}$ corresponds to element j of the diagonal of the covariance matrix. The covariance matrix, which is automatically calculated by most software implementations of the NLLS algorithm, corresponds to the inverse of the negative of the expected values of the Hessian matrix $-E[H(\hat{\boldsymbol{P}})]$, where the Hessian $H(\hat{\boldsymbol{P}})$ is the matrix of second derivatives of the likelihood function $\mathcal{L}$ with respect to $\hat{\boldsymbol{P}}$. Standard error $\sigma_{\hat{p}_j} = \sqrt{\sigma^2_{\hat{p}_j}}$ of parameter $\hat{p}_j \in \hat{\boldsymbol{P}}$ corresponds to the square root of the variance.

Note that variance estimates are only correct asymptotically. In practice, fitting algorithms will use numerical approximations of Hessian matrix values.

Quality of finite approximations of the second derivatives of $\mathcal{L}$ will tend to be proportional to sample size $|\hat{\boldsymbol{Y}}|$, inversely proportional to distance from parameter constraint boundaries, and inversely proportional to the number of parameters $|\hat{\boldsymbol{P}}|$.

Parameter estimation error for augmented function $f'(x; \boldsymbol{P'})$ can be improved, without deteriorating fit quality, by solving the fitting problem in two stages. In the first stage, we determine $\hat{\boldsymbol{P}}'$ by solving for $Z(f'(x; \hat{\boldsymbol{P}}'))$. For the best-case scenario, where $f'(x; \boldsymbol{P'})$ is fitted to an infinitely large sample $\hat{\boldsymbol{Y}}$ randomly drawn from $f'(x; \hat{\boldsymbol{P}}')$, the estimated value of scaling parameter $\hat{s} \in \hat{\boldsymbol{P}}'$ will completely eliminate the bias in the sum of residuals $\sum_{i \in I} e(f(x_i; \hat{\boldsymbol{P}}), \hat{y}_i)$, such that $\int_a^b f(x; \bar{\boldsymbol{P}}')dx = \sum_{i \in I} W \hat{y}_i$.

In the second stage, we solve for $Z(f''(x; \hat{\boldsymbol{P}}, \hat{s}))$, where f'' corresponds to our augmented distribution f' with the scaling parameter value fixed at $s = \hat{s}$ (i.e. only the original vector of parameters $\boldsymbol{P}$ is optimized by the fitting algorithm).

The shape distributions from both stages are equivalent, such that

$$Z(f'(x; \hat{\boldsymbol{P}}')) \simeq Z(f''(x; \hat{\boldsymbol{P}}, \hat{s})). \tag{32}$$

However, error vector $\boldsymbol{\sigma}_{\hat{P}}$ and parameter covariance (which can be estimated from off-diagonal elements of the covariance matrix) estimated in the second stage will tend to be more reliable.

Our computational experiment dataset consists of 52 192 stems extracted from a database of PSP data, collected from public forests in Quebec (Canada). This data was collected by the *Ministère de la forêt, de la faune et des parcs* (MFFP) as part of the official government inventory program[1], which operates

[1]Detailed information on the PSP inventory program under which our test data was collected is available from the MFFP web site (`http://www.mffp.gouv.qc.ca/forets/inventaire/`).

on a 10-year cycle.

Data was collected throughout the province of Quebec, using 37 foot (11.28 meter) radius circular fixed-area plots. The full dataset contains 1 685 233 stems, sampled from 12 570 permanent sample plot locations. However, this includes repeated measures from four decennial inventory cycles, collected from 7 different PSP networks. We filtered data to include only stems from the most recent inventory cycle, which ensures that we are not tallying repeated measures on the same plots. We further filtered data to include only stems from the largest of the seven PSP networks (codename *BAS1*), which ensures uniform data-collection standards for all stems.

Our ultimate goal (i.e. beyond the scope of this paper) is to link a long-term wood supply optimization model with a short-term fibre-procurement optimization model. Thus, we are interested in modelling diameter distribution of merchantable stems in mature (operable), undisturbed stands. We therefore applied a series of other filters to our stem dataset to exclude plots in disturbed or immature stands, unmerchantable stems (with diameter less than 3.54 inches [9 cm]), very large stems (with diameter greater than 24 inches [61 cm]), and dead or otherwise unmerchantable stems.

A Jupyter Notebook containing Python code implementing these filters and detailed explanations is available from the corresponding author upon request. Although we do not have permission to distribute the PSP dataset, it is possible to request a copy from the *Ministère des forêts, de la faune et de parcs.*

We segmented the 52 192 stems in our filtered PSP dataset into 30 sub-datasets, representing combinations of 10 species groups and 3 cover types. More detailed information on species groups is provided in an appendix. For each of 30 sub-datasets $d \in D$, we applied our two-stage fitting method on 25 target distributions $f \in F$, i.e. GB1, GB2, and GG parent distributions, and

the 22 special cases of these distributions defined in (6) through (27).

We used the small-sample form of the Akaike information criterion (AICc) to evaluate goodness-of-fit for each combination of $d \in D$ and $f \in F$. AICc is given by

$$\text{AICc} = AIC + \frac{2K(K+1)}{n-K-1} \tag{33}$$

with

$$\text{AIC} = 2K - n \ln\left(\frac{\chi^2}{n}\right) \tag{34}$$

where $K = |\mathbf{P}| + 1$ (i.e. the cardinality of the parameter vector $\mathbf{P}$, plus the μ parameter of the implicit i.i.d. Gaussian error distribution of input data vector $\hat{\mathbf{Y}}$), $n = |\hat{\mathbf{Y}}|$, as recommended in Burnham and Anderson, 2002), and χ^2 is the sum of squared residuals given by

$$\chi^2 = \sum_{i \in I} e\left(f(x_i; \hat{\boldsymbol{P}}), \hat{y}_i\right)^2 \tag{35}$$

For each sub-dataset $d \in D$, we ranked distributions $f \in F$ in decreasing order of AICc, and reported best-fit distribution, best-fit parameter values (with standard error estimates on parameter values) for first and second stages, and second-stage AICc.

3 Results

Figures 1 and 2 show best-fit distributions plotted against empirical input data distribution, binned by diameter class. The name of the best-fit distribution is identified in the legend for each subplot.

Table 1 reports estimated parameter values, standard error on parameter estimates, and second-stage AICc for best-fit distributions.

4 Discussion

As predicted, second-stage parameter standard error estimates are systematically lower than first-stage error estimates. This is attributable to fixing of the s parameter in the second stage.

Four distributions (GG, GA, W, EXP) dominate our best-fit model selection experiment, taking first place for 28 out of 30 combinations of species group and cover type. The χ^2 and B1 distributions had the lowest AICc value for the other two cases.

This confirms previous results in the forestry literature reporting success using GG, GA, W, and EXP distributions to model stem diameter distribution from stem tally data. If analytic resources are highly constrained, we recommend limiting the list of candidate distributions these four. Note that in most cases many of the other 21 distributions were rejected altogether, either because the NLLS algorithm did not converge, or because fitting results were deems unstable (due to extremely high parameter values, or extremely high or otherwise unreliable parameter error estimates). GA, W, and EXP distributions are all derivatives of the GG distribution, with one or more of its three parameters fixed to a value of 1. It is not surprising that the GG distribution generally had slightly better fit (i.e. lower χ^2), however results show that AICc recommends selecting a more parsimonious model in most (but not all) cases.

Overall, fit results are very good, as indicated by relatively low second-stage parameter standard error estimates. This observation can be confirmed by visual inspection of fit results in Figures 1 and 2. Relatively small sample sizes in some combinations of species group and cover type yielded binned datasets with more erratic values (including empty bins, which were excluded from NLLS algorithm input data before fitting). Naturally, best-fit parameter standard error and AIC values are higher for these datasets.

Our input dataset includes a large number of stems, inventoried throughout the province of Quebec. Our best-fit distribution results could be used to forecast diameter distribution for mature stands in Quebec, or in other locations with similar forests. For researchers looking for more customized fits, our two-stage methodology can easily be replicated on publicly-available inventory data.

5 Acknowledgements

This study was supported by funding from the *FORAC Research Consortium*.

References

Bailey, R. L. and Dell, T. (1973). Quantifying diameter distributions with the weibull function. *Forest Science*, 19(2):97–104.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag New York.

Cao, Q. V. (2004). Predicting parameters of a weibull function for modeling diameter distribution. *Forest science*, 50(5):682–685.

Coomes, D. A. and Allen, R. B. (2007). Mortality and tree-size distributions in natural mixed-age forests. *Journal of Ecology*, 95(1):27–40.

Ducey, M. J. and Gove, J. H. (2015). Size-biased distributions in the generalized beta distribution family, with applications to forestry. *Forestry*, 88(1):143–151.

Hyink, D. M. and Moser, J. W. (1983). A generalized framework for projecting

forest yield and stand structure using diameter distributions. *Forest Science*, 29(1):85–95.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):149–176.

Liu, C., Zhang, L., Davis, C. J., Solomon, D. S., and Gove, J. H. (2002). A finite mixture model for characterizing the diameter distributions of mixed-species forest stands. *Forest Science*, 48(4):653–661.

Meyer, H. A. and Stevenson, D. D. (1943). The structure and growth of virgin beech-birch-maple-hemlock forests in northern pennsylvania. *J. Agric. Res*, 67(2).

Nelson, T. C. (1964). Diameter distribution and growth of loblolly pine. *Forest Science*, 10(1):105–114.

Schreuder, H. T., Ernst, R., and Ramirez-Maldonado, H. (2004). Statistical techniques for sampling and monitoring natural resources. Technical Report RMRS-GTR-126, USDA Forest Service, Rocky Mountain Research Station.

Appendix

Table 2 lists common and Latin names of species in the species groups used to segment our PSP data.

Figure 1: Best-fit distributions are shown with a solid line. Empirical distributions (binned by 0.79 inch [2 cm] diameter class) are shown with gray circles. Bin-wise sampling error is shown with light gray error bars. Species group is fixed for a given row of subfigures, and cover type is fixed for a given column of subfigures.

Figure 2: (Continued from Figure 1) Best-fit distributions are shown with a solid line. Empirical distributions (binned by 0.79 inch [2 cm] diameter class) are shown with gray circles. Bin-wise sampling error is shown with light gray error bars. Species group is fixed for a given row of subfigures, and cover type is fixed for a given column of subfigures.

Table 1: Best-fit distributions for each combination of species group and cover type. We report estimated parameter values and standard error for first- and second-stage fits, and second-stage AIC.

Table 2: Mapping of species group names to species common and Latin names. Alternate names are shown in parentheses.