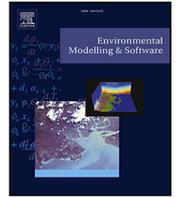




Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

ForestFit: An R package for modeling plant size distributions

Mahdi Teimouri^{a,*}, Jeffrey W. Doser^b, Andrew O. Finley^b

^a Department of Statistics, Faculty of Science and Engineering, Gonbad Kavous University, Gonbad Kavous, Iran

^b Department of Forestry, Michigan State University, East Lansing, MI, USA

ARTICLE INFO

Keywords:

Bayesian
Forest modeling
Grouped data
Parameter estimation
Mixture models
R package
Statistical distributions

ABSTRACT

Modeling the diameter distribution of trees in forest stands is a frequent application in ecology that supports key biologically and economically relevant management decisions. The choice of model used to represent the diameter distribution and how to estimate its parameters has received much attention in the forestry literature; however, accessible software that facilitates comprehensive comparison of the myriad modeling approaches is not available. To this end, we developed an R package called **ForestFit** that simplifies estimation of common probability distributions used to model tree diameter distributions, including the two- and three-parameter Weibull distributions, Johnson's SB distribution, Birnbaum–Saunders distribution, and finite mixture distributions. Frequentist and Bayesian techniques are provided for individual tree diameter data, as well as grouped data. Additional functionality facilitates fitting curves to height-diameter data. The package also provides a set of functions for computing probability distributions and simulating random realizations from common finite mixture models.

1. Introduction

The diameter distribution of trees in a stand is a relatively simple measure used to inform forest management decisions that have both biological and economical relevance (Bailey and Dell, 1973). Determining models and associated parameter estimation techniques to characterize an entire stand's diameter distribution based upon a sample of trees has been an important topic in forest biometrics over the last 50 years (Green et al., 1994; Merganič and Sterba, 2006). Using an appropriate model to describe the diameter distribution enables an evaluation of the stand growth as well as potential estimates of stand volume. Numerous probability distributions have been proposed and used to model the diameter distribution, such as the two-parameter Weibull distribution, three-parameter Weibull distribution (Bailey and Dell, 1973), Johnson's SB (JSB) distribution (Fonseca et al., 2009), and finite mixture distributions (Zasada and Cieszewski, 2005). Forest biometricians have developed a suite of methods for estimating the parameters of these models, including simple methods such as the method of moments and method of percentiles (Wang and Keats, 1995), maximum likelihood techniques (Gove and Fairweather, 1989), and Bayesian methods (Green et al., 1994).

Another important tree and stand characterization used by foresters is the diameter to height relationship. Because diameter is an inexpensive field measurement relative to acquiring tree height, the

forest biometrics literature is rich with models to predict height based on diameter (Temesgen et al., 2014). Due to the extensive suite of techniques in the forest biometrics literature to model diameter distributions and height-diameter relationships, there is a need for freely available software that facilitates streamlined model parameter estimation and comparison. To this end, we develop an open-source R package called **ForestFit** that provides a limited set of functions to fit a large variety of popular diameter distributions and height-diameter models. **ForestFit** is open-source software released under GPL2/GPL3 license and available on the Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/web/packages/ForestFit/index.html>. Importantly, the functions are designed to simplify comparison among models and parameter estimation techniques using a variety of model fit criteria, without requiring an in-depth knowledge of the statistical techniques used in the estimation process. **ForestFit** will have immediate use in forest science as well as other ecological and environmental fields where modeling size distributions is an important task (e.g., botany, fisheries, and hydrology).

In this paper, we describe the functionality of the **ForestFit** package and illustrate its features using the analysis of mixed ponderosa pine (*Pinus ponderosa*) and western Juniper (*Juniperus occidentalis*) forest plots located in the Blue Mountains near Burns, Oregon, USA (Kerns et al., 2017).

* Corresponding author.

E-mail address: teimouri@aut.ac.ir (M. Teimouri).

Table 1
Parameter estimation methods for two- and three-parameter Weibull distributions.

Parameters	Method	Citation
2	Generalized regression type 1	Kantar (2015)
2	Generalized regression type 2	Kantar (2015)
2	L-moment	Hosking (1990)
2	Maximum likelihood (ML)	Gove and Fairweather (1989)
2	Logarithmic moments	Teimouri et al. (2013)
2	Method of moments	Bailey and Dell (1973)
2	Percentiles	Wang and Keats (1995)
2	Rank correlation	Teimouri and Nadarajah (2012)
2	Least squares	Kantar (2015)
2	U-statistics	Sadani et al. (2019)
2	Weighted ML	Cousineau (2009)
2	Weighted least squares	Zhang et al. (2008)
3	ML	Green et al. (1994)
3	Modified moments	Cohen and Whitten (1982)
3	Modified ML	Cohen and Whitten (1982)
3	Moments	Cran (1988)
3	Maximum product spacing	Teimouri et al. (2013)
3	Weighted ML	Cousineau (2009)

2. ForestFit components

In this section we provide an overview of the features available in the **ForestFit** package along with relevant citations that provide a more rigorous statistical treatment.

2.1. Estimating parameters of the Weibull distribution

The Weibull distribution is perhaps the most conspicuous distribution used for modeling diameter distributions due to its ability to represent common shapes in diameter distribution data (Bailey and Dell, 1973; Gorgoso et al., 2007; Gorgoso-Varela and Rojo-Alboreca, 2014; Merganić and Sterba, 2006; Pretzsch, 2009; Stankova and Zlatanov, 2010). **ForestFit** offers several methods for estimating the parameters of two- and three-parameter Weibull distributions, which are listed in Table 1 along with relevant sources for further details regarding each estimation method.

2.2. Bayesian analysis

Computational improvements over the last 30 years have led to the increase in popularity of Bayesian techniques for estimating model parameters for diameter distributions (Green et al., 1994). In **ForestFit** we provide functionality for estimating the parameters of the three-parameter Weibull and four-parameter JSB distributions using the Bayesian paradigm. The JSB distribution is becoming increasingly popular for modeling diameter distributions in forestry because of its flexibility to model a wide variety of forest types (Fonseca et al., 2009; George and Ramachandran, 2011; Zhang et al., 2003) and is commonly applied in other fields, e.g., hydrology (Cugerone and De Michele, 2015) and climatology (Lu et al., 2008).

With the exception of the location parameter, **ForestFit** uses the same algorithm as Green et al. (1994) for estimating the shape and scale parameters of the three-parameter Weibull distribution.

2.3. Modeling diameter distributions using grouped data

Data collection protocols in forestry often lead to diameter observations that are recorded in groups or classes of diameters (e.g., < 5 m, between 5 m and 10 m, etc.) rather than as individual tree measurements. More specifically, suppose a data set has been classified into m separate groups with lower bounds r_0, r_1, \dots, r_m where $r_0 = \min\{x_1, \dots, x_n\}$ and $r_m = \max\{x_1, \dots, x_n\}$, with x_i being the individual diameter for tree i out of a total n trees in the sample. We define f_i , for $i = 1, \dots, m$, as the frequency or number of trees in the i th group. A schematic of grouped data is provided in Table 2.

Table 2
A schematic of grouped data.

Class number	Class	Frequency
1	$(r_0 - r_1]$	f_1
2	$(r_1 - r_2]$	f_2
3	$(r_2 - r_3]$	f_3
\vdots	\vdots	\vdots
$m - 1$	$(r_{m-2} - r_{m-1}]$	f_{m-1}
m	$(r_{m-1} - r_m]$	f_m

ForestFit provides estimation of diameter distribution parameters for the three-parameter Birnbaum–Saunders (BS) distribution, the three-parameter generalized exponential (GE) distribution, and the three-parameter Weibull distribution using grouped data through the use of the vectors $r = (r_0, r_1, \dots, r_m)'$ and $f = (f_1, f_2, \dots, f_m)'$.

2.4. Modeling diameter distributions using finite mixture models

Certain forest stands (e.g., uneven aged or biotically disturbed stands) exhibit multimodal diameter distributions that are not adequately represented by a single probability distribution. In such settings, forest biometricians and statisticians often use finite mixture distributions to characterize complex and multimodal diameter distributions (Zhang et al., 2001; Liu et al., 2002; Zasada and Cieszewski, 2005; Zhang and Liu, 2006; Liu et al., 2014). The probability density function (PDF) of a K-component mixture model has the form

$$g(x|\theta) = \sum_{k=1}^K \omega_k f(x|\theta_k), \quad (1)$$

where $\theta = (\omega', \theta'_1, \dots, \theta'_K)'$ in which $\theta_k = (\alpha_k, \beta_k)'$ is the parameter vector of the k th component with PDF $f(\cdot|\theta_k)$ and $\omega = (\omega_1, \dots, \omega_K)'$ is the vector of weight parameters. The weight parameters ω_k s are non-negative and sum to one, i.e., $\sum_{k=1}^K \omega_k = 1$. **ForestFit** enables $f(\cdot|\theta_k)$ to take the following forms for individual (i.e., ungrouped) data: Burr XII, Chen, Fisher, Fréchet, gamma, Gompertz, log-logistic, log-normal, Lomax, skew-normal, and two-parameter Weibull (PDFs are given in Appendix). For grouped data, the finite mixture distribution can be composed of gamma, log-normal, skew-normal, and two-parameter Weibull distributions.

Application of finite mixture models are increasing in popularity in environmental and ecological analyses, see, e.g., Hobbs and Hooten (2015) and Jaworski and Podlaski (2012). To facilitate testing and simulation using the common mixture models implemented in **ForestFit**, the package includes functions to compute the probability density function, cumulative density function, quantile function, and simulate random variables from the finite mixture models composed of the aforementioned probability distributions. Furthermore, **ForestFit** provides specific functions to compute the probability density function, cumulative density function, quantile function, and simulate random variables from the gamma shape mixture model that has received increased attention in forestry for characterizing diameter distributions (Venturini et al., 2008).

2.5. Estimating the parameters of height-diameter models

In addition to modeling diameter distributions, characterizing the height-diameter relationship is an important task for obtaining accurate growth and yield measurements to inform forest management (Temesgen et al., 2007). A variety of mathematical relationships have been proposed to model this relationship, the most common of which are implemented in **ForestFit** and listed with associated references in Table 3.

Table 3

Common mathematical relationships used to depict the height-diameter relationship in a forest stand. H is the height of a tree, D is the diameter of the tree, and $\beta_1, \beta_2, \beta_3$ are estimated parameters. 1.3 represents the minimum height of a mature tree in a forest stand in meters, and represents the height of a human at breast height.

Model	Reference	Formula
Chapman–Richards	Richards (1959)	$H = 1.3 + \beta_1 + \frac{\beta_2}{D + \beta_3}$
Gompertz	Winsor (1932)	$H = 1.3 + \beta_1 e^{-\beta_2 e^{-\beta_3 D}}$
Hossfeld IV	Zeide (1993)	$H = 1.3 + \frac{\beta_1}{1 + \frac{\beta_2}{D^{\beta_3}}}$
Korf	Flewelling and Jong (1994)	$H = 1.3 + \beta_1 e^{-\beta_2 D^{-\beta_3}}$
Logistic	Pearl and Reed (1920)	$H = 1.3 + \frac{\beta_1}{1 + \beta_2 e^{-\beta_3 D}}$
Prodan	Prodan (1968)	$H = 1.3 + \frac{D^2}{\beta_1 D^2 + \beta_2 D + \beta_3}$
Ratkowsky	Ratkowsky and Giles (1990)	$H = 1.3 + \beta_1 e^{-\frac{\beta_2}{D^{\beta_3}}}$
Sibbesen	Huang et al. (1992)	$H = 1.3 + \beta_1 D^{\beta_2} D^{-\beta_3}$
Weibull	Yang et al. (1978)	$H = 1.3 + \beta_1 \left(1 - e^{-\beta_2 D^{\beta_3}}\right)$

3. Worked examples

Here we illustrate key features available in **ForestFit**. All models and estimation methods assume the data points (e.g., diameter of an individual tree) are independent and identically distributed. For illustration, we use a set of 0.08 ha plots comprising mixed ponderosa pine and western Juniper located in the Malheur National Forest on the southern end of the Blue Mountains near Burns, Oregon, USA (Kerns et al., 2017). These data are freely available (<https://www.fs.usda.gov/rds/archive/catalog/RDS-2017-0041>) under the condition that users cite the reference (Kerns et al., 2017). Further, we include these data in **ForestFit** so the data can easily be obtained using the command `data(DBH)`. For illustrative purposes, we only use the diameter and height measurements at 1.3 m height of all live trees in plots 72, 57, 55, and 51, and refer to these values as `d72`, `d57`, `d55`, and `d51`, respectively. The heights of trees in plot 55 are referred to as `h55`. Below we provide the commands for loading the `DBH` data set and obtaining the data from plot 55.

```
library(ForestFit)
data(DBH)
d <- DBH[DBH[, 1] == 55, 10]
h <- DBH[DBH[, 1] == 55, 11]
d55 <- d[!is.na(d)]
h55 <- h[!is.na(h)]
```

3.1. Estimating parameters of the Weibull distribution

The function `fitWeibull` estimates parameters of the two- and three-parameter Weibull distribution and takes the form

```
fitWeibull(data, location, method, starts)
```

where `data` is a vector of data observations, `location` indicates whether to use a three-parameter Weibull distribution (`TRUE`) or a two-parameter Weibull distribution (`FALSE`), `method` denotes the estimation method to use, and `starts` is a vector of starting values for the parameters to be estimated. Available estimation methods and their code in **ForestFit** are shown in Table 4.

The code below calls `fitWeibull` to fit the three parameter Weibull distribution using maximum product spacing (`mps`) to the plot 72 data, with the function output printed below.

Table 4

Parameter estimation methods for the Weibull distribution and associated code used in the `method` argument of the `fitWeibull` function. MM refers to modified method, ML refers to maximum likelihood, and MML refers to modified maximum likelihood.

Method	Code
Generalized regression type 1	greg1
Generalized regression type	greg2
L-moment	lm
ML (2 parameters)	ml
Logarithmic moments	mlm
Method of moments	moment
Percentiles	pm
Rank correlation	rank
Least squares	reg
U-statistics	ustat
Weighted ML	wml
Weighted least squares	wreg
ML (3 parameters)	mle
MM Type 1	mm1
MM Type 2	mm2
MM Type 3	mm3
MML type 1	mm11
MML type 2	mm12
ML type 3	mm13
MML type 4	mm14
Maximum product spacing	mps
T-L moment	t1m

```
starts <- c(2, 2, 3)
fitWeibull(data = d72, location = TRUE,
           method = "mps", starts = starts)

$estimate
      alpha      beta      mu
[1,] 1.297011 18.57024 12.41853

$measures
      AIC      CAIC      BIC      HQIC
[1,] 238.7653 239.6542 243.0672 240.1676
      AD      CVM      KS
[1,] 0.9055053 0.1239423 0.1441135
      log.likelihood
[1,] -116.3826
```

The first element of the output list, `estimate`, contains the point estimates for the model parameters. The second element of the list, `measures`, contains a series of goodness-of-fit measures including Akaike Information Criterion (AIC) (Akaike, 1998), Consistent Akaike Information Criterion (CAIC) (Hurvich and Tsai, 1991), Bayesian Information Criterion (BIC) (Schwarz, 1978), Hannan–Quinn information criterion (HQIC) (Hannan and Quinn, 1979), Anderson–Darling (AD) (Anderson and Darling, 1952), Cramér–von Mises (CVM) (Cramér, 1928), Kolmogorov–Smirnov (KS) (Smirnov, 1948), and log-likelihood (`log.likelihood`). These measures allow for comparison between different models and estimation techniques to determine the best fitting model according to the desired criteria. Larger values of the log-likelihood indicate a better model fit, while smaller values of all other model comparison criteria indicate a better model fit.

3.2. Bayesian analysis

Bayesian inference for diameter distributions modeled using the JSB and Weibull distributions is provided via the `fitbayesJSB` and `fitbayesWeibull` functions, respectively. The `fitbayesJSB` function has the form

```
fitbayesJSB(data, n.burn = 8000, n.simul = 10000)
```

where `data` is a vector of observations, `n.burn` is an integer representing the number of burn-in iterations (i.e., the number of iterations the Gibbs sampler takes to reach convergence), and `n.simul` is the total number of Gibbs sampler iterations. By default, `fitbayesJSB` uses 10,000 total iterations and 8000 burn-in iterations. `fitbayesWeibull` has the same arguments. Below we call both functions using the diameters of all live trees in plot 72.

```
fitbayesJSB(d72, n.burn=8000, n.simul=10000)
```

```
$estimate
      delta      gamma      lambda      xi
[1,] 0.6897204 0.4077035 43.51213 11.68084

$measures
      AD      CVM      KS
[1,] 0.5059267 0.06804388 0.1235231
      log.likelihood
[1,] -113.4759
```

```
fitbayesWeibull(d72, n.burn=8000, n.simul=10000)
```

```
$estimate
      alpha      beta      mu
[1,] 1.632042 21.18533 10.28903

$measures
      AD      CVM      KS
[1,] 0.7636954 0.1028456 0.1378961
      log.likelihood
[1,] -116.7509
```

We see that for plot 72, all goodness-of-fit measures indicate the JSB model is a better fit for the data (Fig. 1). Additional examples using the `fitbayesJSB` and `fitbayesWeibull` functions with alternative data sets can be found in the manual pages of `ForestFit`.

3.3. Modeling diameter distributions using grouped data

Suppose diameter observations are given in m separate groups as shown in Table 2. `ForestFit` provides the function `fitgrouped` for estimating the parameters of the three-parameter BS and Weibull distributions using grouped data. This function takes the form

```
fitgrouped(r, f, family, method1, starts, method2)
```

where r is a length $m + 1$ vector of group boundaries as shown in Table 2, with the first element of r being the lower bound of the first group and all other m values representing the upper bound of the m groups. f is a vector of length m of the group frequencies. `family` represents the distribution used in the model, taking values `'weibull'` or `'birnbaum-saunders'`. `method1` is a string denoting the method of estimation. Here we enable three methods of estimation: approximated maximum likelihood (`'aml'`), the expectation-maximization (EM) (Dempster et al., 1977) algorithm (`'em'`), and maximum likelihood (`'ml'`). `starts` is a numeric vector containing the starting values for the shape, scale, and location parameters, respectively. Lastly, `method2` indicates the optimization method of the log-likelihood,

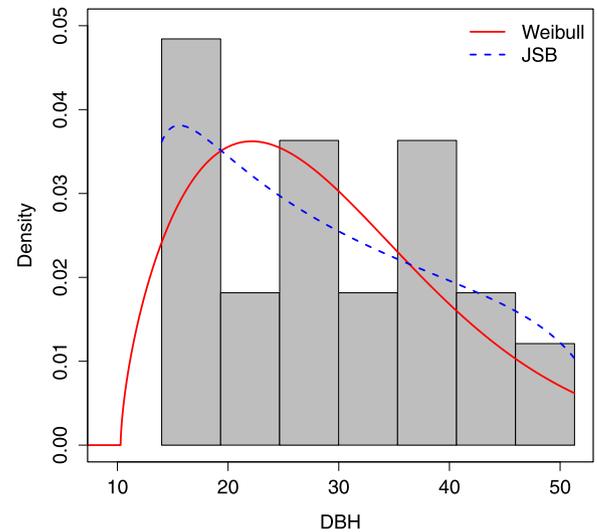


Fig. 1. Diameter distribution of trees in plot 72. fitted JSB and Weibull PDFs estimated using the Bayesian paradigm are superimposed.

taking values `'BFGS'`, `'CG'`, `'L-BFGS-B'`, `'Nelder-Mean'`, and `'SANN'`. If using the EM algorithm, the `stats` and `method2` arguments do not need to be specified. Below we manually group the diameter observations from plot 57 into six groups, and use the function `fitgrouped` on this grouped data set.

```
m <- 6
r <- seq(min(d57), max(d57), length=m+1)
D <- data.frame(table(cut(d57, r)))
fitgrouped(r = r, f = D$Freq,
           family = "birnbaum-saunders",
           method1 = "em")
```

```
$estimate
      alpha      beta      mu
[1,] 0.6234071 8.660411 8.453387

$measures
      AIC      CAIC      BIC      HQIC
[1,] 24.81261 25.03483 28.89871 26.4006
      AD Chi-square      CVM      KS
[1,] 8.401981 1.595622 4.393473 0.05225002
      log.likelihood
[1,] -10.4063
```

Fig. 2 displays the estimated BS PDF to plot 57's diameter distribution data.

3.4. Modeling diameter distributions of grouped data using finite mixture models

The `fitmixturegrouped` function provides an interface for fitting a wide-range of finite mixture distributions to grouped data and takes the form

```
fitmixturegrouped(family, r, f, K,
                  initial = FALSE, starts)
```

where r and f are vectors of the class boundaries and frequencies as previously described in Section 3.3. K is a single numeric value

denoting the number of component PDFs to include in the finite mixture model. All parameters in finite mixture models are estimated using the EM algorithm. The `family` argument is a character string denoting the distribution used in the mixture model. Currently available distributions are the gamma (`'gamma'`), log-normal (`'log-normal'`), skew-normal (`'skew-normal'`), and Weibull (`'weibull'`) families. The `initial` argument denotes whether or not the user wants to specify starting values for the EM algorithm. By default the argument is set to `FALSE` and thus the user does not need to specify a value for the `starts` argument. If `initial=TRUE`, the user must specify a sequence of initial values taking the form `starts = (ω, α, β)`, where the three vectors are of length `K` with elements ω_i being the weight of the i th component, and α_i and β_i are the associated parameters of the i th component. If `family = 'skew-normal'`, the user must specify starting values for the weights and three parameters for each component distribution (see Appendix). In the code that follows, we illustrate the `fitmixturegrouped` function by fitting a two-component skew-normal mixture model to plot 51's grouped data.

```
m <- 8; K <- 2
r <- seq(min(d51), max(d51), length=m+1)
D <- data.frame(table(cut(d51, r)))
f <- D$Freq
omega <- c(0.5, 0.5); alpha <- c(10, 40)
beta <- c(2, 2); lambda <- c(2, -2)
fitmixturegrouped(family = "skew-normal",
                  r = r, f = f,
                  K = K, initial = TRUE,
                  starts = c(omega, alpha,
                           beta, lambda))
```

```
$estimate
      weight      alpha      beta      lambda
[1,] 0.6296296 12.11366 1.504117 4.920386
[2,] 0.3703704 32.74526 8.369108 5.844095

$measures
      AIC      CAIC      BIC      HQIC
[1,] 27.82543 30.26021 41.74832 33.19494
      AD Chi-square      CVM      KS
[1,] 21.06318 1.060109 4.625016 0.03171587
      log.likelihood
[1,] -6.912714
```

Similar to previous functions using models with a single PDF, the output is a list where the first element (i.e., `estimate`) contains the estimated parameters of the first and second components of the model, and the second element (i.e., `measures`) contains the goodness-of-fit measures. Fig. 3 displays the estimated two-component skew-normal PDF.

3.5. Modeling diameter distributions using finite mixture models fitted to ungrouped data

As a companion to the grouped data `fitmixturegrouped` function described in Section 3.4, the `fitmixture` function fits finite mixture models to individual diameter distribution data, and takes the form

```
fitmixture(data, family, K, initial = FALSE, starts)
```

where `data` is a vector of individual diameter measurements and `family` denotes the family of probability distributions to use. Available families of PDFs are the BS (`'birnbaum-saunders'`), Burr XI

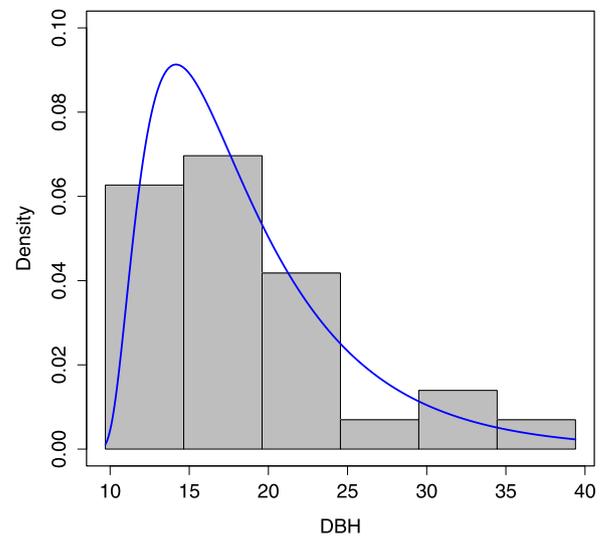


Fig. 2. Diameter distribution of trees in plot 57 classified into six groups. The estimated PDF of the BS distribution is superimposed.

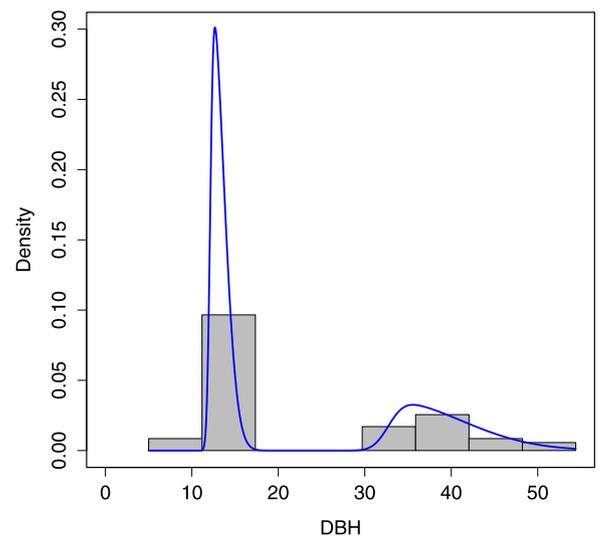


Fig. 3. Histogram of grouped diameter distribution data in plot 51. superimposed is the estimated PDF of the two-component skew-normal mixture model.

(`'burrxi'`), Chen (`'chen'`), Fisher (`'f'`), Fréchet (`'Frechet'`), gamma (`'gamma'`), generalized exponential (`'ge'`), Gompertz (`'gompertz'`), log-normal (`'log-normal'`), log-logistic (`'log-logistic'`), Lomax (`'lomax'`), skew-normal (`'skew-normal'`), and Weibull (`'weibull'`) families. The remaining arguments are analogous to those used in the `fitmixturegrouped` function. Below, the `fitmixture` function is called to fit a two-component log-normal mixture model to plot 51's diameter data.

```
fitmixture(d51, "log-normal", 2)
```

```
$estimate
```

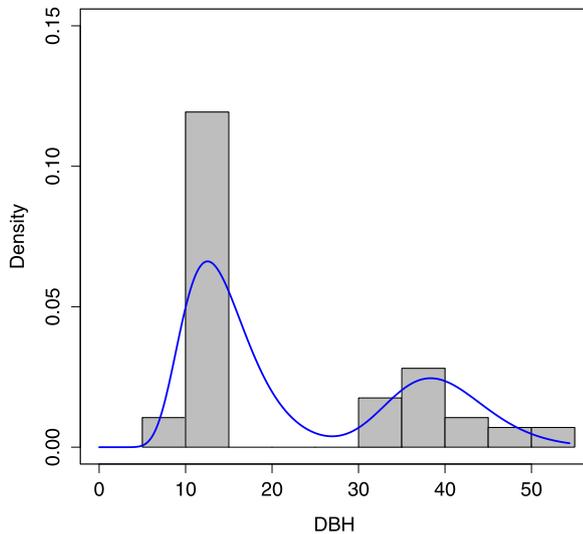


Fig. 4. Diameter distribution of plot 51. superimposed is the estimated PDF of the two-component log-normal mixture model.

```

weight  alpha  beta
[1,] 0.6522847 2.618974 0.2998769
[2,] 0.3477153 3.668380 0.1461719

$measures
      AIC      CAIC      BIC      HQIC
[1,] 419.7617 420.9381 429.9769 423.7317
      AD      CVM      KS
[1,] 5.790067 1.107157 0.3496867
log.likelihood
[1,] -204.8808

$cluster
[1] 1 1 1 2 2 2 2 2 1 2 1 2 1 1 1 1 1 1 1 1
[21] 1 1 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1
[41] 2 2 1 1 2 2 2 2 2 1 1 2 1 1 1 1 1

```

Interpretation of the first and second components of the output list is analogous to the interpretation provided in Section 3.4. The third list element (`cluster`) specifies the component of the mixture distribution from which each individual diameter measurement arises. Fig. 4 displays the fitted two-component log-normal PDF to the diameter distribution data of plot 51. Additional examples using the `fitmixture` function with alternative data sets can be found in the manual pages of `ForestFit`.

3.6. Distribution functions for finite mixture models

Analogous with R's standard functions for working with common probability distributions, we provide commands for computing the density, distribution function, quantile function, and random generation from the finite mixture models used in `ForestFit`. The functions take the following forms

```

dmixture(x, family, K, param)
pmixture(x, family, K, param)
qmixture(p, family, K, param)
rmixture(n, family, K, param)

```

where x is a vector of observations, n is the number of realizations to be simulated from the mixture model, p is a numeric value that satisfies

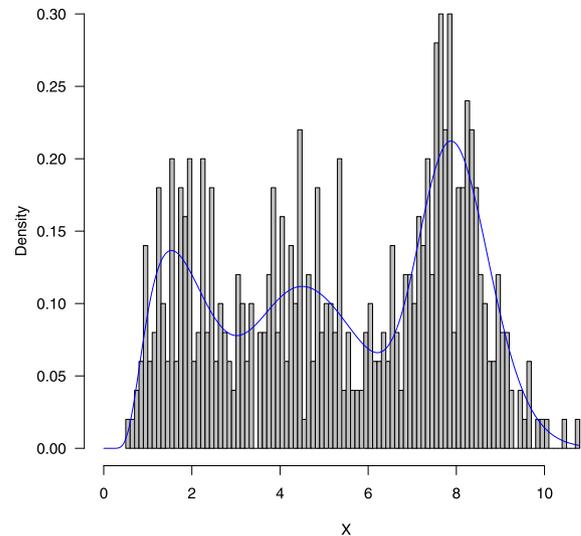


Fig. 5. Histogram of 500 random realizations from a three-component BS mixture model.

$0 < p < 1$, family is one of the families introduced in Section 2.4, K is the number of components, and `param` is the mixture model parameter vector that takes the same form as the `starts` argument of the `fitmixturegrouped` function described previously in Section 3.4. As an illustration, the code below generates 500 realizations from a three-component BS mixture model and displays the resulting distribution in Fig. 5:

```

n <- 500; K <- 3
weight <- c(0.4, 0.3, 0.3)
alpha <- c(0.1, 0.25, 0.5)
beta <- c(8, 5, 2)
param <- c(weight, alpha, beta)
X <- rmixture(n, "birnbaum-saunders",
             K, param)
XX <- seq(0, max(X), 0.01)
pdf <- dmixture(XX, "birnbaum-saunders",
               K, param)
hist(X, freq=FALSE, breaks=120, col='gray',
     las = 1, main = '', xlim = c(0, 11))
lines(XX, pdf, col='blue')

```

3.7. The gamma shape mixture (GSM) model

The gamma shape mixture model (GSM) introduced by Venturini et al. (2008) has received considerable attention in forestry for characterizing diameter distributions. Because of its popularity, we have implemented specific functions in `ForestFit` for estimating the parameters of the GSM distribution, computing the density function, computing the distribution function, and generating realizations for the GSM distribution. The function calls for the aforementioned tasks are

```

fitgsm(data, K)
dgsm(data, omega, beta, log = FALSE)
pgsm(data, omega, beta, log.p = FALSE, lower.tail = TRUE)
rgsm(n, omega, beta)

```

where `data` is a vector of observations, n is the number of realizations simulated from the GSM model, K is the number of components, `omega` is a vector of mixing parameters, `beta` is the rate parameter, `log`

indicates whether to compute the density function (`log = FALSE`) or the log-density function (`log = TRUE`), `log.p` indicates whether to compute the distribution function (`log.p = FALSE`) or the log-transformed distribution function (`log.p = TRUE`), and `lower.tail` indicates whether to compute the upper tail (`lower.tail = FALSE`) or lower tail of the distribution (`lower.tail = TRUE`). As an illustration, we generate 500 realizations from a ten-component GSM model and display the resulting distribution in Fig. 6.

```
n <- 500
K <- 10
omega <- rep(0.1, 10)
beta <- 0.25
X <- rgsm(n, omega, beta)
XX <- seq(0, max(X), 0.01)
out <- fitgsm(X, K)
pdf <- dgsm(XX, out$omega, out$beta)
hist(X, freq = FALSE, breaks = 30, col = 'gray',
     main = '', xlim = c(0, 80))
lines(XX, pdf, col = 'blue')
```

```
out

$beta
[1] 0.2114758

$omega
 [1] 6.767306e-02 1.585398e-01 2.282670e-01
 [4] 7.416988e-03 1.440731e-02 1.349225e-01
 [7] 3.415604e-01 4.696039e-02 2.524812e-04
[10] 4.967357e-08

$measures
      AIC      CAIC      BIC      HQIC
[1,] 4024.886 4025.336 4067.032 4041.424
      AD      CVM      KS
[1,] 0.1486677 0.01948765 0.01939521
      log.likelihood
[1,] -2002.443
```

The output list of the `fitgsm` function consists of three parts:

1. `beta`: estimated rate parameter.
2. `omega`: estimated vector of mixing parameters.
3. `measures`: sequence of goodness-of-fit measures including AIC, CAIC, HQIC, AD, CVM, KS and `log.likelihood`.

3.8. Modeling height-diameter relationships

The general function for fitting curves to paired observations of height and diameter is

```
fitcurve(h, d, model, starts)
```

where `h` and `d` are vectors of height and diameter observations, respectively. The `model` argument denotes the model to fit with options: Chapman-Richards (`'chapman-richards'`), Gompertz (`'gompertz'`), Hossfeld IV (`'hossfeldiv'`), Korf (`'korf'`), logistic (`'logistic'`), Prodan (`'prodan'`), Ratkowsky (`'ratkowsky'`), Sibbesen (`'sibbesen'`), and Weibull (`'weibull'`). Here we fit a Weibull model to the height-diameter relationship of plot 55.

```
starts <- c(18, 0.0005, 2)
fitcurve(h55, d55, "weibull", starts=starts)
```

```
$summary
      Estimate Std. Error  t value
beta1 25.01192975 2.400710045 10.418555
beta2  0.01670455 0.004924011  3.392468
beta3  1.15619358 0.124108288  9.316006

      Pr(>|t|)
beta1 1.258552e-14
beta2 1.289826e-03
beta3 6.630701e-13

$residuals
 [1] 0.8231685 -0.8206984  3.1817793
 [4] -1.4609972 -0.8469181  2.1519832
 [7] -1.6820165 -0.8469181 -0.3682378
[10] -0.8469181  2.0348396  0.5414598
[13] -0.6693033 -0.1354301 -2.7076580
[16]  0.2174916  1.4929903 -1.0507349
[19]  1.2306967  0.3741064 -1.1724307
[22] -2.0955590 -3.7551437  2.8916349
[25]  1.0950994 -2.4201337  2.3717872
[28] -0.2263893 -0.3799072 -0.8469181
[31] -0.8469181 -1.6912881 -0.2627979
[34]  1.0228089 -0.2643013  1.4774336
[37]  1.5548580 -2.5159687  0.1424660
[40] -0.1394704 -0.5379420 -2.2095351
[43]  1.1160033  0.2372021 -1.4451420
[46]  0.3684380  0.4441287  1.1916349
[49]  1.2717576 -0.9287836 -0.7846899
[52] -0.3624105  1.0094147  0.5018908
[55] -2.0315620  3.6684380  1.7044410
[58]  0.0893665

$`var-cov`
      beta1      beta2
beta1 2.374618859 0.0032551742
beta2 0.003255174 0.0000099897
beta3 -0.105817390 -0.0002393425
      beta3
beta1 -0.1058173899
beta2 -0.0002393425
beta3  0.0063462337

$`residual Std. Error`
[1] 1.557911
```

The output list consists of five parts:

1. `summary`: estimated parameters, standard error of the estimators, *t*-value and corresponding *p*-value.
2. `residuals`: residuals of the fitted curve.
3. `'var-cov'`: variance-covariance matrix of the model coefficient estimators.
4. `'residual Std. Error'`: residual standard error.
5. A scatterplot of the height-diameter relationship with the fitted values (Fig. 7)

4. Future development

We are currently working on software functionality that includes: (1) a broader set of goodness-of-fit measures for the Bayesian models; (2) a function that automates “optimal” model/method selection that searches among all methods in Table 4 with selection based on a

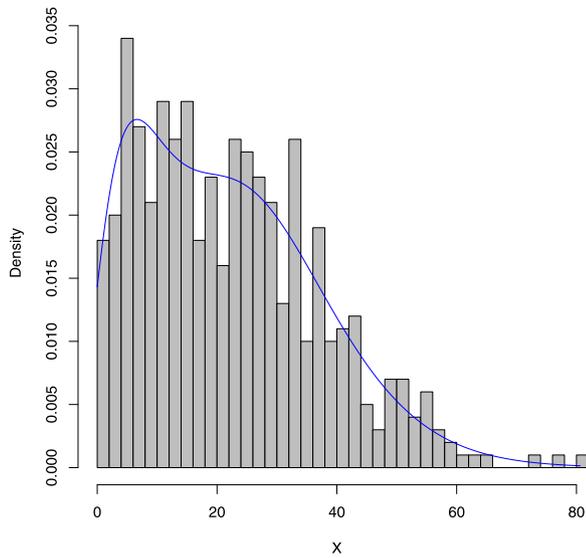


Fig. 6. Histogram of 500 realizations from a ten-component GSM model.

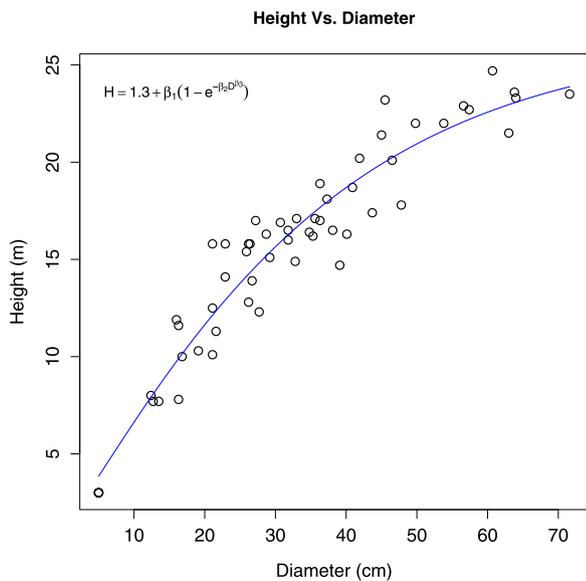


Fig. 7. Weibull model for the height-diameter relationship in plot 55.

set of user identified goodness-of-fit measures; (3) expanded set of mixture models, e.g., Venturini et al. (2008); (4) addition of random effects that accommodate grouped distribution and height-diameter data, e.g., stands, species, functional types, etc.

5. Conclusion

A common task in forestry is modeling the diameter distribution of trees in a given forest stand. In this work, we have developed and introduced an R package called **ForestFit** that provides numerous models and estimation techniques for modeling diameter distributions from both individual tree data and grouped data, as well as providing additional functionality for simulating finite mixture distributions and fitting curves to height-diameter data. To help facilitate comparison of the different model estimation methods provided by **ForestFit**, we include multiple goodness of fit measures that the user can use to determine the best model technique for their purposes. While this package was developed in the context of forestry, the models we fit and

simulate have numerous applications throughout other ecological and environmental fields, such as botany, fisheries, and hydrology. For example, the JSB distribution is used in hydrology to model rain drop size distribution (Cugerone and De Michele, 2015) and finite mixture distributions are broadly applied in environmental and ecological statistics (for numerous examples, see Hobbs and Hooten, 2015), suggesting **ForestFit** can have wide use across ecological and environmental fields. **ForestFit** is open-source software released under GPL2/GPL3 license on the CRAN.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Probability density functions used in ForestFit

- BS $f(x, \theta) = \frac{\sqrt{\frac{x-\mu}{\beta}} + \sqrt{\frac{\beta}{x-\mu}}}{2\alpha(x-\mu)} \phi\left(\frac{\sqrt{\frac{x-\mu}{\beta}} - \sqrt{\frac{\beta}{x-\mu}}}{\alpha}\right)$,
- Burr XII $f(x, \theta) = \alpha\beta x^{\alpha-1} (1+x^\alpha)^{-\beta-1}$,
- Chen $f(x, \theta) = \alpha\beta x^\alpha \exp\left\{x^\alpha - \beta \exp\{x^\alpha\} + \beta\right\}$,
- Fisher $f(x, \theta) = \frac{\Gamma\left(\frac{\alpha+\beta}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)\Gamma\left(\frac{\beta}{2}\right)} \left(\frac{\alpha}{\beta}\right)^{\frac{\alpha}{2}} x^{\frac{\alpha}{2}-1} \left[1 + \frac{\alpha}{\beta} x\right]^{-\frac{\alpha+\beta}{2}}$,
- Fréchet $f(x, \theta) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{-\alpha-1} \exp\left\{-\left(\frac{x}{\beta}\right)^{-\alpha}\right\}$,
- Gamma $f(x, \theta) = \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp\left\{-\frac{x}{\beta}\right\}$,
- GE $f(x, \theta) = \beta \exp\{-\beta(x-\mu)\} \left[1 - \exp\{-\beta(x-\mu)\}\right]^{\alpha-1}$,
- Gompertz $f(x, \theta) = \beta \exp\{\alpha x\} \exp\left\{\frac{\beta \exp\{\alpha x\} - 1}{\alpha}\right\}$,
- Johnson's SB $f(x, \theta) = \frac{\delta \lambda \exp\left\{-\frac{1}{2}\left[\gamma + \delta \log\left(\frac{x-\xi}{\lambda+\xi-x}\right)\right]^2\right\}}{\sqrt{2\pi(x-\xi)(\lambda+\xi-x)}^{-2}}$,
- Log-logistic $f(x, \theta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} \left[\left(\frac{x}{\beta}\right)^\alpha + 1\right]^{-2} \exp\left\{-\frac{1}{2}\left(\frac{\log x - \alpha}{\beta}\right)^2\right\}$,
- Log-normal $f(x, \theta) = \frac{\exp\left\{-\frac{1}{2}\left(\frac{\log x - \alpha}{\beta}\right)^2\right\}}{\sqrt{2\pi}\beta x}$,
- Lomax $f(x, \theta) = \frac{\alpha\beta}{(1+\alpha x)^{\beta+1}}$,
- Skew-normal $f(x, \theta) = 2\phi\left(\frac{x-\alpha}{\beta}\right)\Phi\left(\lambda\frac{x-\alpha}{\beta}\right)$,
- Weibull $f(x, \theta) = \frac{\alpha}{\beta} \left(\frac{x-\mu}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{x-\mu}{\beta}\right)^\alpha\right\}$,

where θ is the family parameter vector; $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density function and distribution function of the standard normal distribution, respectively. Noticeably, the three-parameter BS, gamma, GE, and Weibull distributions simplify to the two-parameter BS, gamma, GE, and Weibull distributions, respectively, when $\mu = 0$.

References

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp. 199–213. http://dx.doi.org/10.1007/978-1-4612-1694-0_15.

Anderson, T.W., Darling, D.A., 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann. Math. Stat. 23 (2), 193–212. <http://dx.doi.org/10.1214/aoms/1177729437>.

Bailey, R.L., Dell, T., 1973. Quantifying diameter distributions with the weibull function. Forest Sci. 19 (2), 97–104. <http://dx.doi.org/10.1093/forestscience/19.2.97>.

Cohen, C.A., Whitten, B., 1982. Modified maximum likelihood and modified moment estimators for the three-parameter weibull distribution. Comm. Statist. Theory Methods 11 (23), 2631–2656. <http://dx.doi.org/10.1080/03610928208828412>.

Cousineau, D., 2009. Nearly unbiased estimators for the three-parameter Weibull distribution with greater efficiency than the iterative likelihood method. Br. J. Math. Stat. Psychol. 62 (1), 167–191. <http://dx.doi.org/10.1348/000711007X270843>.

Cramér, H., 1928. On the composition of elementary errors. Scand. Actuar. J. 1928 (1), 13–74. <http://dx.doi.org/10.1080/03461238.1928.10416862>.

- Cran, G., 1988. Moment estimators for the 3-parameter weibull distribution. *IEEE Trans. Reliab.* 37 (4), 360–363. <http://dx.doi.org/10.1109/24.9839>.
- Cugeron, K., De Michele, C., 2015. Johnson sb as general functional form for raindrop size distribution. *Water Resour. Res.* 51 (8), 6276–6289. <http://dx.doi.org/10.1002/2014WR016484>.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 39 (1), 1–38.
- Flewelling, J.W., Jong, R.D., 1994. Considerations in simultaneous curve fitting for repeated height–diameter measurements. *Can. J. Forest Res.* 24 (7), 1408–1414. <http://dx.doi.org/10.1139/x94-181>.
- Fonseca, T.F., Marques, C.P., Parresol, B.R., 2009. Describing maritime pine diameter distributions with johnson's sb distribution using a new all-parameter recovery approach. *Forest Sci.* 55 (4), 367–373. <http://dx.doi.org/10.1093/forestscience/55.4.367>.
- George, F., Ramachandran, K., 2011. Estimation of parameters of johnson's system of distributions. *J. Modern Appl. Statist. Methods* 10 (2), 9. <http://dx.doi.org/10.22237/jmass/1320120480>.
- Gorgoso, J., González, J.Á., Rojo, A., Grandas-Arias, J., 2007. Modelling diameter distributions of betula alba l. stands in northwest spain with the two-parameter weibull function. *Forest Syst.* 16 (2), 113–123. <http://dx.doi.org/10.5424/srf/2007162-01002>.
- Gorgoso-Varela, J., Rojo-Alboreca, A., 2014. Use of gumbel and weibull functions to model extreme values of diameter distributions in forest stands. *Ann. Forest Sci.* 71 (7), 741–750. <http://dx.doi.org/10.1007/s13595-014-0369-1>.
- Gove, J.H., Fairweather, S.E., 1989. Maximum-likelihood estimation of weibull function parameters using a general interactive optimizer and grouped data. *Forest Ecol. Manag.* 28 (1), 61–69. [http://dx.doi.org/10.1016/0378-1127\(89\)90074-1](http://dx.doi.org/10.1016/0378-1127(89)90074-1).
- Green, E.J., Roesch Jr, F.A., Smith, A.F., Strawderman, W.E., 1994. Bayesian estimation for the three-parameter weibull distribution with tree diameter data. *Biometrics* 254–269. <http://dx.doi.org/10.2307/2533217>.
- Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 41 (2), 190–195. <http://dx.doi.org/10.1111/j.2517-6161.1979.tb01072.x>, URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1979.tb01072.x>, arXiv:<https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1979.tb01072.x>.
- Hobbs, N.T., Hooten, M.B., 2015. *Bayesian Models*. Princeton University Press.
- Hosking, J.R.M., 1990. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 52 (1), 105–124, URL <https://www.jstor.org/stable/2345653>.
- Huang, S., Titus, S.J., Wiens, D.P., 1992. Comparison of nonlinear height–diameter functions for major alberta tree species. *Can. J. Forest Res.* 22 (9), 1297–1304. <http://dx.doi.org/10.1139/x92-172>.
- Hurvich, C.M., Tsai, C.L., 1991. Bias of the corrected aic criterion for underfitted regression and time series models. *Biometrika* 78 (3), 499–509. <http://dx.doi.org/10.1093/biomet/78.3.499>.
- Jaworski, A., Podlaski, R., 2012. Modelling irregular and multimodal tree diameter distributions by finite mixture models: an approach to stand structure characterisation. *J. Forest Res.* 17 (1), 79–88. <http://dx.doi.org/10.1007/s10310-011-0254-9>.
- Kantar, Y.M., 2015. Generalized least squares and weighted least squares estimation methods for distributional parameters. *REVSTAT—Stat. J.* 13, 263–282. <http://dx.doi.org/10.1080/03610918.2011.611315>.
- Kerns, B.K., Westlind, D.J., Day, M.A., 2017. Season and interval of burning and cattle exclusion in the southern blue mountains, oregon: Overstory tree height, diameter and growth. *Forest Serv. Res. Data Arch.* <http://dx.doi.org/10.2737/RDS-2017-0041>.
- Liu, F., Li, F., Zhang, L., Jin, X., 2014. Modeling diameter distributions of mixed-species forest stands. *Scand. J. Forest Res.* 29 (7), 653–663. <http://dx.doi.org/10.1080/02827581.2014.960891>.
- Liu, C., Zhang, L., Davis, C.J., Solomon, D.S., Gove, J.H., 2002. A finite mixture model for characterizing the diameter distributions of mixed-species forest stands. *Forest Sci.* 48 (4), 653–661. <http://dx.doi.org/10.1093/forestscience/48.4.653>.
- Lu, Y., Ramirez, O., Rejesus, R., Knight, T., Sherrick, B., 2008. Empirically evaluating the flexibility of the johnson family of distributions: A crop insurance application. *Agric. Resour. Econ. Rev.* 37, <http://dx.doi.org/10.1017/S1068280500002161>.
- Merganič, J., Sterba, H., 2006. Characterisation of diameter distribution using the weibull function: method of moments. *Eur. J. Forest Res.* 125 (4), 427–439. <http://dx.doi.org/10.1007/s10342-006-0138-2>.
- Pearl, R., Reed, L.J., 1920. On the rate of growth of the population of the united states since 1790 and its mathematical representation. *Proc. Natl. Acad. Sci. USA* 6 (6), 275.
- Pretzsch, H., 2009. *Forest Dynamics, Growth and Yield: from Measurement to Model*. Springer Verlag, Berlin Heidelberg, <http://dx.doi.org/10.5860/choice.47-2562>.
- Prodan, M., 1968. The spatial distribution of trees in an area. *Allg. Forst Jagdztg.* 139, 214–217.
- Ratkowsky, D.A., Giles, D.E., 1990. *Handbook of nonlinear regression models*. Marcel Dekker, New York.
- Richards, F., 1959. A flexible growth function for empirical use. *J. Exp. Botany* 10 (2), 290–301.
- Sadani, S., Abdollahzad, K., Teimouri, M., Ranjbar, V., 2019. A new estimator for weibull distribution parameters: Comprehensive comparative study for weibull distribution.
- Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Statist.* 6 (2), 461–464. <http://dx.doi.org/10.1214/aos/1176344136>.
- Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.* 19 (2), 279–281. <http://dx.doi.org/10.1214/aoms/1177730256>.
- Stankova, T.V., Zlatanov, T.M., 2010. Modeling diameter distribution of austrian black pine (*Pinus nigra Arn.*) plantations: a comparison of the weibull frequency distribution function and percentile-based projection methods. *Eur. J. Forest Res.* 129 (6), 1169–1179. <http://dx.doi.org/10.1007/s10342-010-0407-y>.
- Teimouri, M., Hoseini, S.M., Nadarajah, S., 2013. Comparison of estimation methods for the weibull distribution. *Statistics* 47 (1), 93–109. <http://dx.doi.org/10.1080/02331888.2011.559657>.
- Teimouri, M., Nadarajah, S., 2012. A simple estimator for the weibull shape parameter. *Int. J. Struct. Stab. Dyn.* 12 (02), 395–402. <http://dx.doi.org/10.1109/24.9839>.
- Temegen, H., Hann, D.W., Monleon, V.J., 2007. Regional height–diameter equations for major tree species of southwest oregon. *West. J. Appl. Forestry* 22 (3), 213–219.
- Temegen, H., Zhang, C., Zhao, X., 2014. Modelling tree height–diameter relationships in multi-species and multi-layered forests: A large observational study from north-east china. *Forest Ecol. Manag.* 316, 78–89. <http://dx.doi.org/10.1016/j.foreco.2013.07.035>.
- Venturini, S., Dominici, F., Parmigiani, G., 2008. Gamma shape mixtures for heavy-tailed distributions. *Ann. Appl. Stat.* 2 (2), 756–776. <http://dx.doi.org/10.1214/07-AOAS156>.
- Wang, F., Keats, J., 1995. Improved percentile estimation for the two-parameter weibull distribution. *Microelectron. Reliab.* 35 (6), 883–892. [http://dx.doi.org/10.1016/0026-2714\(94\)00168-N](http://dx.doi.org/10.1016/0026-2714(94)00168-N).
- Winsor, C.P., 1932. The gompertz curve as a growth curve. *Proc. Natl. Acad. Sci. USA* 18 (1), 1. <http://dx.doi.org/10.1073/pnas.18.1.1>.
- Yang, R.C., Kozak, A., Smith, J.G., 1978. The potential of weibull-type functions as flexible growth curves. *Can. J. Forest Res.* 8 (4), 424–431. <http://dx.doi.org/10.1139/x78-062>.
- Zasada, M., Cieszewski, C.J., 2005. A finite mixture distribution approach for characterizing tree diameter distributions by natural social class in pure even-aged scots pine stands in poland. *Forest Ecol. Manag.* 204 (2–3), 145–158. <http://dx.doi.org/10.1016/j.foreco.2003.12.023>.
- Zeide, B., 1993. Analysis of growth equations. *Forest Sci.* 39 (3), 594–616. <http://dx.doi.org/10.1093/forestscience/39.3.594>.
- Zhang, L., Gove, J.H., Liu, C., Leak, W.B., 2001. A finite mixture of two weibull distributions for modeling the diameter distributions of rotated-sigmoid, uneven-aged stands. *Can. J. Forest Res.* 31 (9), 1654–1659. <http://dx.doi.org/10.1139/cjfr-31-9-1654>.
- Zhang, L., Liu, C., 2006. Fitting irregular diameter distributions of forest stands by weibull, modified weibull, and mixture weibull models. *J. Forest Res.* 11 (5), 369–372. <http://dx.doi.org/10.1007/s10310-006-0218-7>.
- Zhang, L., Packard, K.C., Liu, C., 2003. A comparison of estimation methods for fitting weibull and johnson's sb distributions to mixed spruce fir stands in northeastern north america. *Can. J. Forest Res.* 33 (7), 1340–1347. <http://dx.doi.org/10.1139/x03-054>.
- Zhang, L., Xie, M., Tang, L., 2008. On weighted least squares estimation for the parameters of weibull distribution. In: *Recent Advances in Reliability and Quality in Design*. Springer, pp. 57–84. http://dx.doi.org/10.1007/978-1-84800-113-8_3.