

Finite mixture model for the application in forestry

Fuxiang Liu

Department of statistics at School of economics, Harbin University of Commerce, Heilongjiang
Province, PR China
liufuxiang2009@163.com

Keywords: finite mixture model, Weibull function, Diameter-class distribution, mixed-species forest stand.

Abstract: Finite mixture model (FMM) has been used extensively to fit such distribution in different study fields including medicine, biology, fisheries, forestry, environmental science, engineering, and economics. This study compares three diameter distribution models to fit mixed-species forest stands using four example plots with two or three species components in Daxing'an Mountain P.R. China. Our results indicated that the FMM models are more flexible to describe highly skewed and irregular diameter distributions for the whole plot, as well as provide the acceptable estimation for each species component and the mixing proportions. Thus, the FMM models can be a useful tool for effectively managing mixed-species forest stands.

1. Introduction

A frequency distribution composed of two or more component distributions is defined as a “mixture” or “compound” distribution, involving a finite number of components. FMM has been used extensively to fit such distribution in different study fields including medicine, biology, fisheries, environmental science, engineering, and economics (e.g., MacDonald and Pitcher 1979, Zasada and Cieszewski 2005, Zhang and Liu 2006). FMM simultaneously estimates the parameters of each component distribution in the mixture, as well as the mixing probabilities of component membership (Chandra 1977, Everitt and Hand 1981, Titterton et al. 1985). A variety of statistical distribution functions have been used as the component distribution (Hasselblad 1969, Shaked 1980), such as exponential (Bartholomew 1969), beta (Bremner 1978), logistic (Shah 1963), gamma (Ashton 1971), normal (Behboodan 1970), Weibull (Falls 1970), etc. Statistical methods have been applied to estimate the parameters of mixture distributions, ranging from moment method (Blischke 1962), maximum likelihood (Falls 1970), graphical technique (Kao 1959), to Bayes estimation (Padgett and Tsokos 1978).

Forest researchers and practitioners were used to consider that FMM was too complicated to use and apply. With recent and rapid improvement and advances on computing technology and software packages such as R package (Leisch 2004), Stata (Debo 2008), SAS (SAS institute, Inc. 2010) and others (Haughton 1997), it is now relatively easy to fit a diameter-class distribution by FMM. However, the studies on FMM in the forestry and ecology literature were most focused on comparing the model fitting of “whole stand” by different methods (e.g., Zhang et al. 2001, Zhang and Liu 2006). Little attention has been paid to the fitting of component distributions in the mixture

distribution. Therefore, the main purpose of this study was to fit mixed two-species and three-species forest stands, using four example plots, by three methods: (1) a single Weibull function for fitting the whole stand, (2) a single Weibull function for fitting each species component separately, and (3) a FMM model for fitting the 2 or 3 species components simultaneously. The performance of the 3 methods was compared with the fitting of the whole stand, the fitting of each species component, and the estimation of the mixing proportions for the species components.

2. Theoretical Background

Suppose a mixture distribution consisting of k components ($j = 1, 2, \dots, k$) with a random variable x of interest under study (e.g., tree diameter). The distribution of the j th individual component is described by a specific pdf, $f_j(x)$, and the general pdf, $f(x, p)$, for the mixture distribution can be expressed as follows:

$$f(x, p) = \sum_{j=1}^k p_j f_j(x) = p_1 f_1(x) + p_2 f_2(x) + \dots + p_k f_k(x) \quad (1)$$

Where $p = (p_1, p_2, \dots, p_{k-1})$ is the vector of $k-1$ mixing proportions of the individual components in the mixture and must satisfy the constraints: $0 < p_j < 1$ and $p_k = 1 - \sum_{j=1}^{k-1} p_j$. Note the overall mixture distribution $f(x, p)$ may be a mixture of different component distributions or a mixture of the same component distribution. In this study, the 3-parameter Weibull distribution is selected as the component distribution $f_j(x)$:

$$f_j(x) = \left(\frac{c_j}{b_j} \right) \left(\frac{x - a_j}{b_j} \right)^{c_j - 1} \exp \left[- \left(\frac{x - a_j}{b_j} \right)^{c_j} \right], \quad x \geq a_j, a_j \geq 0, b_j > 0, c_j > 0 \quad (2)$$

Where a_j , b_j and c_j are the location, scale and shape parameters of the j th individual component distribution, respectively. The cumulative distribution function (cdf) of the j th individual component distribution is

$$F_j(x) = 1 - \exp \left[- \left(\frac{x - a_j}{b_j} \right)^{c_j} \right] \quad (3)$$

Then, the compound cumulative distribution function is defined as

$$F(x, p) = \sum_{j=1}^k p_j F_j(x) = p_1 F_1(x) + p_2 F_2(x) + \dots + p_k F_k(x) \quad (4)$$

The maximum likelihood approach is commonly applied to fit a FMM model, due to its attractive statistical properties and relatively easy to use in practice (Chandra 1977). The joint likelihood density function is as follows

$$L = \prod_{i=1}^n f(x, p) = \prod_{i=1}^n [p_1 f_1(x) + p_2 f_2(x) + \dots + p_k f_k(x)] \quad (5)$$

The natural logarithm of the likelihood function ($\log L$) is expressed by

$$\log L = \sum_{i=1}^n \log [f(x, p)] = \sum_{i=1}^n \log [p_1 f_1(x) + p_2 f_2(x) + \dots + p_k f_k(x)] \quad (6)$$

The partial derivatives of $\log L$ are taken with respect to each of the model and mixing parameters of the mixture distribution. These partial derivatives are set equal to zero, and then

solved by a numerical iterative algorithm such as the Newton-Raphson algorithm to yield the ML estimates (McLachlan and Peel 2000).

2.1. Example Plots

The four fixed sample plots were set up in 2000 and re-measured in 2005 and 2010. This study used the data of 2005. The plots were 0.06 ha in size. Trees with diameter at the breast height (dbh) ≥ 5 cm were measured on the sample plots with an accuracy of 1 mm. The descriptive statistics of tree diameters for the four example plots and each species in the plot were shown in Table 1. All trees were grouped into 2-cm diameter classes. The ages of the four plots were between 60 and 70 years old with basically reversed J-shaped diameter distributions. More trees were gathered in small diameter classes (5– 10 cm). The features of the four example plots were (1) Plot 1 had 63% Dahurian larch (5 – 25 cm) and 37% white birch (5 – 15 cm). The two species had similar frequency distributions (reversed J-shaped), as well as similar to the distribution of the whole plot; (2) Plot 2 had 67% Dahurian larch (5 – 17 cm) and 33% white birch (5 – 37 cm). The frequency distribution of Dahurian larch was highly skewed to the right, while the frequency distribution of white birch was uniform across the diameter classes; (3) Plot 3 had three tree species. Dahurian larch (41%, 5 – 23 cm) had a relatively normal frequency distribution, white birch (52%, 5 – 19 cm) had a skewed distribution to the right due to a few large-sized trees, and Mongolian pine (7%, 5 – 19 cm) had a uniform distribution across the diameter classes; and (4) Plot 4 also had 3 species: 42% Dahurian larch (5 – 27 cm), 25% white birch (5 – 16 cm), and 34% Mongolian pine (5 – 31 cm). Dahurian larch had a skewed to the right distribution, while both white birch and Mongolian pine had relatively uniform distributions but different ranges of tree sizes.

Table 1. Descriptive statistics of tree diameters in the example plots.

Plot	Species	Number of Trees	Observed Proportion	Mean	Std	Min	Max
1	Total	61		9.02	3.9109	5	23
	Larch	38	0.63	9.95	4.507	5	23
	White birch	22	0.37	7.55	1.8022	5	13
2	Total	111		10.18	6.5074	5	37
	Larch	74	0.67	7.41	1.9096	5	15
	White birch	37	0.33	15.73	8.576	5	37
3	Total	100		9.84	3.4604	5	23
	Larch	41	0.41	8.15	2.943	5	19
	White birch	52	0.52	11.12	3.2913	5	23
	Mongolian pine	7	0.07	7.56	1.45	5	19
4	Total	89		12.74	6.6999	5	31
	Larch	37	0.42	11.62	5.479	5	27
	White birch	22	0.25	9.64	2.9931	5	15
	Mongolian pine	30	0.34	16.4	8.2203	5	31

2.2. Modeling Methods

In this study, we chose the 3-parameter Weibull function (equation [2]) as the component density function. The location parameter a was fixed at 5 cm (i.e., the minimum measured tree diameter) and the scale parameter b and shape parameter c were estimated from the data.

2.2.1. Model Fitting of Whole Plot

Use 20-point type for the title, aligned to the center, linespace exactly at 14-point with a bold and

italic font style and initial letters capitalized. No formulas or special characters of any form or language are allowed in the title.

Words like “is”, “or”, “then”, etc. should not be capitalized unless they are the first word of the title.

The parameter estimates of the three fitting methods and the standard errors (in parentheses) of the parameters for the four example plots were presented in Table 2. It was evident, according to the model AIC, that the Method 1 (FMM) fitted each example plot better than the Method 2 (a single Weibull) to fit the whole plot, except for Plot 3 (Table 2). However, the Method 1 ($-2\log L = 493.4$) actually fitted the Plot 3 better than the Method 2 ($-2\log L = 498.3$), but the AIC is designed to penalize the number of estimated model parameters. Because the number of the FMM model parameters ($m = 8$) was much larger than those of the single Weibull model parameters ($m = 2$) it appeared that the AIC of the Method 1 (509.4) was larger than that of the Method 2 (502.3). Note there was no AIC value for the Method 3 because it fitted each species separately and no model fitting statistics was available for the whole plot.

Table 2. Parameter estimates (standard error)

Method 1									
Plot	b_1	c_1	b_2	c_2	b_3	c_3	p_1	p_2	AIC
1	2.395	0.5954	8.3111	0.6048			0.6424		292.8
	(0.3069)	(0.1157)	(0.5536)	(0.3283)			(0.4413)		
2	2.0881	0.5775	12.4659	0.6755			0.6456		562.6
	(0.1284)	(0.06948)	(0.2406)	(0.1781)			(1.4309)		
3	7.2498	0.5439	2.3807	0.5468	0.8674	1.828822	0.3396	0.4136	509.4
	(2.0595)	(0.4586)	(0.8340)	(0.09926)	(0.03490)	(0.09926)	(12.7683)	(6.2704)	
4	4.3579	0.1141	11.1591	0.1723	21.8590	0.7031	0.5633	0.1761	539.9
	(0.3276)	(0.02809)	(0.1555)	(0.1326)	(0.3276)	(0.09881)	(1.2657)	(0.4325)	

3. Conclusions

To evaluate the model fitting and prediction performance the bias, RMSE, and goodness-of-fit likelihood-ratio test were computed and shown for each method and each example plot in Table 3. The model prediction and model residuals across the diameter classes are displayed for each example plot in Figures 1, respectively. The Plot 1 had a balanced reversed-J shape diameter distribution for the whole plot (Fig. 1(a)). Based on the likelihood-ratio test all three methods adequately fitted this plot ($p\text{-value} > 0.05$), although the Method 1 had much smaller bias and RMSE than those of other two methods.

Both Plots 3 and 4 had 3 species components and the diameter distributions of the whole plot were more irregular in shape, either bimodal (Plot 3, Fig 1(c)) or rotated-sigmoid (Plot 4, Fig 1(d)). For both plots, the Method 1 (FMM) was the only one fitting these irregular distributions well, while both Methods 2 and 3 failed to adequately describe them according to the likelihood-ratio tests ($p\text{-value} < 0.05$) and yielded much larger biases and RMSEs. Basically, the Methods 2 and 3 produced typical reversed-J shape curves for these two plots, thus underestimated for the diameter class 6 cm and overestimated for larger diameter classes.

Our model fitting and comparison results indicated that the mixture Weibull models were more flexible to fit regular and irregular diameter distributions Our results confirmed the findings in previous studies (e.g., Zhang et al. 2001, Liu et al. 2002, Zhang and Liu 2006).

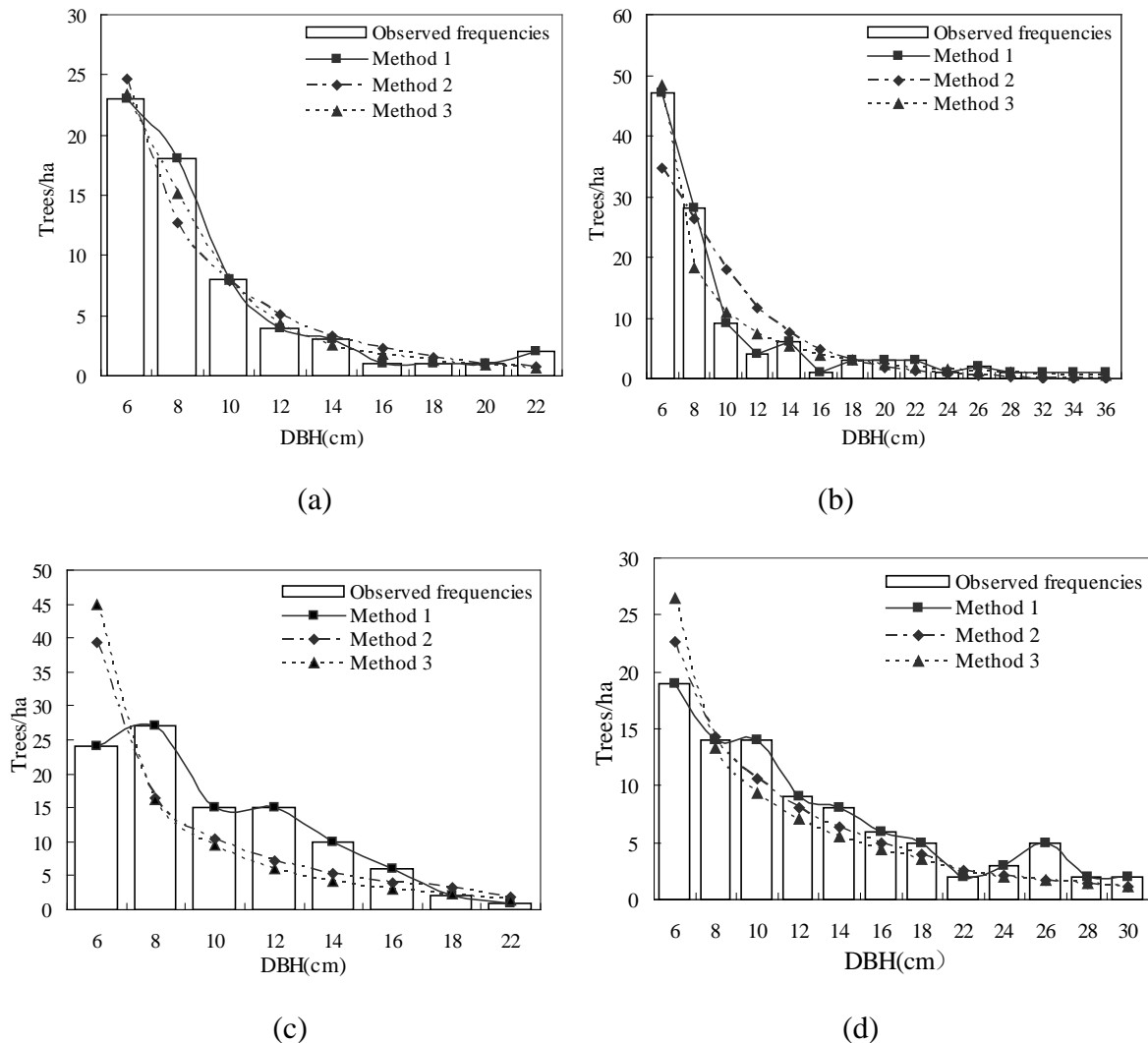


Figure 1. Comparison of three model fitting methods for the example plots. (a) Plot 1, (b) Plot 2, (c) Plot 3 and (d) Plot 4.

Acknowledgements

This research was financially supported by Doctor Startup Fund of Harbin University of Commerce (14rw12) and Program for Philosophy and Social Science for the Planning of Heilongjiang province (16JYE01)

References

- [1] Ashton, W.D. (1971). Distribution for gaps in road traffic. *J. Inst. Math.* 7:37-46.
- Bailey, R.L. and Dell, T.R. (1973). Quantifying diameter distributions with the Weibull function. *For. Sci.* 19:97-104.
- [2] Bartholomew, D.J. (1969). Sufficient conditions for a mixture of exponentials to be a probability density function. *Ann. Math. Stat.* 40:2183-2188.
- [3] Behboodan, J. (1970). On a mixture of normal distributions. *Biometrika* 57:215-217.
- [4] Bremner, J.M. (1978). Mixtures of beta distributions algorithm AS 123. *App. Stat.* 27:104-09.

- [5] Chandra, S. (1977). On the mixtures of probability distributions. *Scand. J. Stat.* 4:105-112.
- [6] Deb, P. (2008). Finite mixture models. Summer North American Stata Users' Group Meetings 2008 7, Stata Users Group, revised 28 Aug 2008. Accessed on November 1, 2013. www.stata.com/meeting/snasug08/deb_fmm_slides.pdf
- [7] Falls, L.W. (1970). Estimation of parameters in compound Weibull distributions. *Technometrics* 12:399-407.
- [8] Hafley, A.M. and Schreuder, H.T. (1977). Statistical distributions for fitting diameter and height data in even-aged stands. *Can. J. For. Res.* 4:481-487.
- [9] Liu, C., Zhang, L., Davis, C.J., Solomon, D.S. and Gove, J.H. (2002). A finite mixture model for characterizing the diameter distribution of mixed-species forest stands. *For. Sci.* 48:653-661.
- [10] Maltamo, M., Kangas, A., Uuttera, J., Tornaiainen, T. and Saramaki, J. (2000). Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogeneous Scots pine stands. *For. Ecol. Manag.* 133:263–274.
- [11] McLachlan, G. and Peel, D. (2000). *Finite mixture models*. John Wiley & Sons, New York.
- [12] Podlaski, R. (2010a). Two-component mixture models for diameter distributions in mixed-species, two-age cohort stands. *For. Sci.* 56:379–390.
- [13] Podlaski, R. (2010b). Diversity of patch structure in Central European forests: are tree diameter distributions in near-natural multilayered *Abies-Fagus* stands heterogeneous? *Ecol. Res.* 25:599–608.
- [14] SAS Institute, Inc. (2010). *SAS/STAT user's guide, version 9.3*. SAS Institute, Cary, NC.
- [15] Stankova, T.V. and 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. For. Res.* 129:1169-1179.