Developing a new model for predicting the diameter distribution of oak forests using an artificial neural network

Shisheng Long[1], Siqi Zeng[1], Guangxing Wang[2]✉

Abstract The parameters of the probability density function (PDF) may be estimated using the parameter prediction method (PPM) and the parameter recovery method (PRM). However, these methods can suffer from accuracy issues. We developed and evaluated the prediction accuracy of two PPMs (stepwise regression model and dummy variable model) and an artificial neural network (ANN) to predict diameter distribution using data collected from 188 oak forest plots. The results demonstrated that the Weibull distribution performed well in fitting the diameter distribution. Compared with the stepwise regression model, the PPM model with stand type as a dummy variable reduced the predictional errors in estimating the parameters b and c of the Weibull distribution, but the prediction accuracy of the diameter distribution showed no significant improvement. Compared with the two PPM models, the ANN model with diameter class (C), average diameter (D) and stand type (T) as input variables decreased the RRMSE by 2.9% and 4.33% in estimating diameter distribution, respectively. The satisfactory prediction accuracy and simple model structure indicated that an ANN worked well for the prediction of the diameter distribution with few requirements and high practicality.

Addresses: [1]Faculty of Forestry, Central South University of Forestry and Technology, Changsha, Hunan, China| [2]Department of Geography and Environmental Resources, Southern Illinois University, USA.

✉ **Corresponding Author:** Guangxing Wang (gxwang@siu.edu).

Introduction

Oak trees are one of the dominant species found in the broad-leaved forests of subtropical, tropical and temperate regions (Nixon 1993, Perea et al. 2017). The 8th Chinese National Forest Inventory indicates that the total area of oak trees in China has reached 16.72 million hectares. The area and volume of oak trees account for 10.15% and 12.94%, respectively, of China's forests (Li et al. 2001, Wang et al. 2019). However, oak forests in China face quality problems related to excessive stand density and low growth rates. At present, the average volume of oak forests in China is approximately 77.39 m³·ha⁻¹, whereas in Germany it is 305 m³·ha⁻¹ (Hou et al. 2017). Therefore, understanding the stand structure, growth prediction and management of oak forests is particularly significant and urgent.

The diameter distribution refers to the number of trees within a certain diameter range at breast height (1.3 m height) (Hafley & Schreuder 1977). This measure is not only a basis for stand height, basal area and volume but is also a precondition for afforestation design (Meyer 1952, Leak 1964), the forest succession process (Goelz & Leduc 2002) and growing and harvesting models (Clutter et al. 1984, Borders & Pattreson 1990). At present, diameter distribution is mainly studied in reference to even-aged forests with few studies focusing on uneven-aged forests because uneven-aged forests are complex and diverse.

The study of the diameter distribution is mainly based on the PDF and on theoretical models. Related research has progressed from a fouce on static to dynamic methods. Static simulation methods simulate the pattern of stand distribution with PDFs such as normal, negative exponential, Johnson's SB (Johnson 1949), beta (Clutter & Bennett 1965), gamma (Nelson 1964), and Weibull distribution (Weibull 1951, Bailey & Dell 1973, Cao 2004). For natural forests, static simulations have been widely used to study the diameter

distribution. Many studies have shown that the Weibull function has the advantages of a flexible curve and simple mathematical operation among these PDFs (Lima et al. 2015, Quiñonez-Barraza et al. 2015). Bailey and Dell (1973) discussed the application of the Weibull function to diameter distribution studies, and the results from comparing available diameter distribution models show that no model could reveal the characteristics or the Weibull model. At present, the Weibull model has been widely used to describe the diameter distribution of *Pinus taeda* (Smalley & Bailey 1974, Lee & Coble 2006), *Eucalyptus robusta* (Miranda et al. 2018, Souza Retslaff et al. 2012, Binoti et al. 2010), *Pinus palustris* (Jiang & Brooks 2009), oak mixed forests (Sun et al. 2019, Carretero & Torres Alvarez 2013) and other mixed tree species (Lima et al. 2015, Coomes & Allen 2007). However, some researchers have reached different conclusions. Podlaski & Zasada (2008) found that the diameter distribution of a single-layer stand could be best simulated by the normal and logistic distributions, and the Weibull distribution and gamma distribution models were most suitable for the straight-diameter distribution of multilayer stands in fir and birch mixed forests. Kayes et al. (2012) found that both the lognormal distribution and Weibull distribution show the same applicability in describing the diameter distribution of *Acacia auriculiformis*. Mønness (1982) also found that the SB distribution is more suitable than the Weibull distribution in describing the diameter distribution of *Pinus sylvestris*. Thus, the applicability of the diameter distribution function may vary depending on tree species or stand structure.

The dynamic prediction of the stand diameter distribution mainly refers to the estimation of model parameters by the PPM (Hyink & Moser 1983, Siipilehto 1999), PRM (Bowling et al. 1989, Lindsay et al. 1996, Schütz & Rosset 2020) and percentile approach (Brooks et al. 1992, Pérez-López et al. 2019). At present, the

PPM is commonly used (Little 1983, Kilkki et al. 1989). The parameters of the diameter distribution model for a new stand can be estimated using the PPM by constructing a functional relationship between the parameters of the distribution model and the stand variables (average diameter at breast height (DBH), average tree height, cross-sectional area, etc.) (Siipilehto 1999). The PPM can greatly reduce the costs required for forestry measurements, but it suffers from low levels of predictive accuracy (Álvarez-González et al. 2002). Fortunately, a model with dummy variables can improve prediction accuracy by adding virtual stand-related variables (Pogoda et al. 2019). Dummy variable models have been widely used in regression analysis and modelling due to their high modelling accuracy (McCrohan & Harvey et al. 1989, Wang et al. 2008, Chen et al. 2017). Therefore, the application of a dummy variable model for the prediction of parameters for the diameter distribution model may be feasible.

The diameter distribution may not be characterized accurately in terms of the PDFs due to its complex nonlinear relationship (Huang 2000). For nonlinear problems, an ANN can approximate any type of arbitrary nonlinear function due to its intrinsic properties. Thus, an ANN may offer a new avenue for diameter distribution modelling. Compared to other pattern recognition tools, ANNs have some advantages, such as robustness and tolerance of noise (Samarasinghe 2006, Gurney 1997). ANNs have recently attracted considerable attention as a promising alternative for predicting the behaviour of complex, nonlinear systems (Dande & Samant 2018). At present, ANNs have been successfully applied in forestry for the estimation of tree height (Özçelik et al. 2013), the prediction of volume (Diamantopoulou 2005, Ashraf et al. 2013) and the prediction of other forest attributes (Corne et al. 2004). ANNs have been used for diameter distribution modelling, but such studies have focused on using ANNs to predict

the parameters of the PDF (Abbasi et al. 2008, Cai et al. 2010). However, few studies are available on the use of ANNs to directly fit and predict the stand diameter distribution.

Thus, the main purpose of this study was to use an ANN to predict the stand diameter distribution. This study 1) selected the most suitable distribution function from six models to fit the diameter distribution of oak secondary forest; 2) compared the fitting effects of the conventional stepwise regression method and dummy variable method in parameter prediction; 3) constructed a diameter distribution prediction model using ANN based on stand variables; and 4) selected the most suitable diameter distribution prediction model by evaluating the predictive performance of the stepwise regression model, dummy variable model and ANN model.

Materials and Methods

Study site

The study site is located in Hunan Province of southern China at longitudes and latitudes ranging from 108°47′ ~ 114°15′ E and 24°38′ ~ 30°08′ N, respectively (Figure 1). The study area has an elevation range of 24 ~ 2122 m and contains complex landforms, including hills, flatlands, and mountains. The mean annual temperature is 15 ~ 18°C, and the mean annual precipitation reaches 1200 ~ 1700 mm. The soil in the study area mainly includes red soil (Ultisols, US soil taxonomy) and yellow soil (Acrustox, US soil taxonomy) with a small amount of calcareous soil (Alfisols US soil taxonomy). Oak is a dominant species in the broad-leaved forests of Hunan Province. The oak forest can be partitioned into three stand types: pure oak forest, oak-coniferous mixed forest and oak-broad-leaved mixed forest. The main conifer species are *Pinus massoniana* and *Cunninghamia lanceolata*. The main broad-leaved species are *Cinnamomum camphora*, *Liquidambar formosana*, and *Carpinus turczaninowii*.
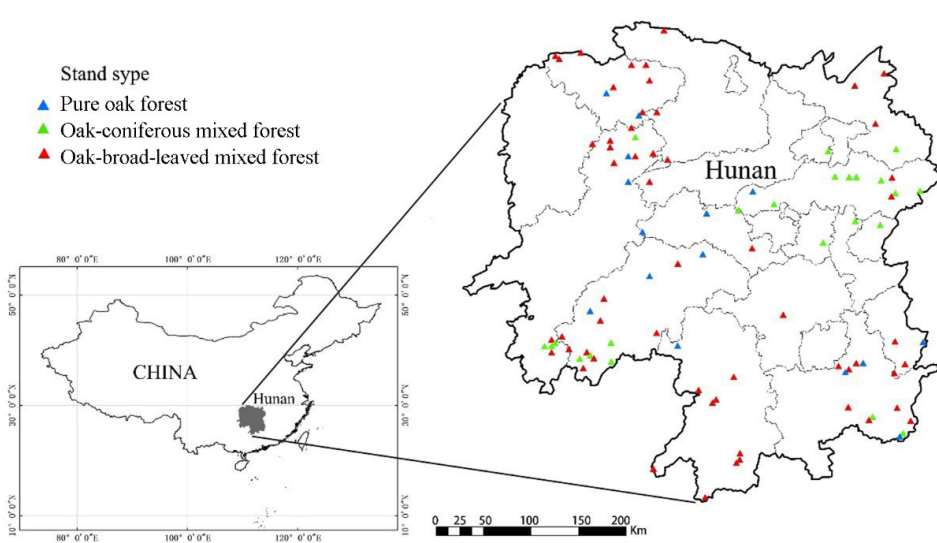
Figure 1 The location of the study area – Hunan – with the spatial distribution of sample plots.

Data

After statistical analysis and data collation, 22323 observations from 188 sample plots were selected. The predominant tree species of the plots were oak, conifer species and broad-leaved species. Sample plots were surveyed in 2004 and 2014, and their size was 25.82 x 25.82 m. The number of trees per hectare for each plot exceeded 750, and the oak species accounted for more than 30%. Within each plot, the DBH was measured, the basal area and trees per hectare were computed, and tree species, elevation, slope, aspect, slope position and soil were recorded. DBH was measured using a diameter tape. Plot elevation and slope position were measured using a global positioning system. The dataset was randomly split into two parts: 70% (133 plots) for modelling and 30% (55 plots) for validation. The modelling and verification sample plots are shown in Table 1.

Table 1 Descriptive statistics for modelling and validation data.

Data	Variables	Mean	Min	Max	SD
Modelling data (133 plots)	Mean DBH (cm)	10.26	6.67	22.52	2.69
	Basal area (m²/667m²)	1.245	0.222	4.408	0.656
	Trees per hectare (trees·ha⁻¹)	1833	795	4275	603
Validation data (55 plots)	Mean DBH (cm)	10.96	5.83	22.55	3.05
	Basal area (m²/667m²)	1.301	0.139	3.733	0.704
	Trees per hectare (trees·ha⁻¹)	1655	765	3000	484

Note: Max: maximum; Min: minimum; SD: standard deviation

Parameter prediction model

The parameter prediction model constructs the functional relationship between stand variables and parameters to predict the diameter distribution parameter of a new stand. Six common diameter distribution functions were developed and compared to fit the diameter distribution of oak forest. The specific expressions of the six functions are shown in Table 2.

The percentile procedure was used to estimate the parameters of the logistic distribution function (Zankis 1979). Moment estimation (García 1981) was used to estimate the parameters of the other five distribution functions. The Kolmogorov-Smirnov (KS) test was used to test the distribution type of the sample (Little 1983).

Table 2 Expression of the diameter distribution functions.

Name	Expression
Normal distribution	$f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Logarithmic normal distribution	$f(x) = \begin{cases} 0 & , \ when \ x \le 0 \\ \dfrac{1}{xc\sqrt{2\pi}} e^{-\frac{(lnx-b)^2}{2c^2}} & , when \ x > 0 \end{cases}$
Weibull distribution	$f(x) = \begin{cases} 0 & , \ when \ x \le a \\ \dfrac{c}{b}(\dfrac{x-a}{b})^{c-1} e^{-(\frac{x-a}{b})^c} & , when \ x > 0 \end{cases}$
Gamma distribution	$f(x) = \begin{cases} 0 & , \ when \ x \le a \\ \dfrac{b^{-c}}{\Gamma(c)}(x-a)^{c-1} e^{-\frac{x-a}{b}} & , when \ x > 0 \end{cases}$
Beta distribution	$f(x) = \begin{cases} 0 & , \ when \ x \notin [a_1, a_2] \\ (x-a_1)^{b-1}(a_2-x)^{c-1}(a_2-a_1)^{-(b+c-1)}/B(b,c) \end{cases} , when \ x > 0$
Logistic distribution	$f(x) = 1/(1 + e^{-\frac{x-a}{b}})$

Description f(x) is the frequency of diameter class, x is the diameter class, μ is the mathematical expectation, σ is the standard deviation, a, b and c are the model coefficients.

Stepwise regression model

At present, the stepwise regression method is commonly used to construct the parameter prediction model. In this study, the parameter of the distribution function was fitted as the dependent variable, and the alternative stand variables included average diameter (D), diameter squared (D^2), basal area (BA) and number of trees per hectare (N) as the independent variables. The stepwise regression model was constructed as follows:

$$y = \varphi_0 + \varphi_1 D + \varphi_2 D^2 + \varphi_3 BA + \varphi_4 N \tag{1}$$

where y is the dependent variable, φ_0 is the intercept, and φ_1-φ_4 are model coefficients.

The independent variables were selected based on stepwise regression. We allowed variables with variance inflation factors less than 10 and significant influence to enter the model to avoid serious collinearity problems between independent variables.

Dummy variable model

The growth of trees may vary among different forest types (Oettel et al. 2020). After data collation, we partitioned the oak forest stands into three types: pure oak forest (T1), oak-coniferous mixed forest (T2) and oak-broad-leaved mixed forest (T2). The average diameter and trees per hectare of the three stand types were analysed by one-way completely randomized analysis of variance using IBM SPSS Statistic 24 software. The average diameters of the stand types were ranked as T1>T2>T3 (Figure 2a), and significant differences were found among the average diameters of the three stand types (p<0.05). This result indicates that the stand type had a significant effect on DBH. The numbers of trees per hectare in the stand types were ranked as T1>T2>T3 (Figure 2b), and there were significant differences in the number of trees per hectare among the three stand types (p<0.05). This result indicates that the stand type had a significant effect on the stand density. Therefore, in this paper, stand type was introduced into the stepwise regression model as a dummy variable.

A dummy model is a model that includes dummy variables to address qualitative factors or classified variables (Lee 1974, Dorsett & Webster 1983). When the dummy variable was added to model (1), the model expression was as follows:

$$y_i = \varphi_0 + \varphi_1 D + \varphi_2 D^2 + \varphi_3 BA + \varphi_4 N + \sum \varphi_j * r_j \qquad (2)$$

where φ_j is a specific parameter of the dummy variable and r_j is a dummy variable to distinguish the three stand types.

## ANN model

Construction of the ANN model

An ANN is a popular type of deep learning algorithm with three or more hidden layers, and each layer consists of a number of neuron nodes. An ANN is usually composed of an input layer, a hidden layer, and an output layer. The number of input layer and output layer nodes is equal to the input and output vector dimensions, respectively. Each node between the two layers has
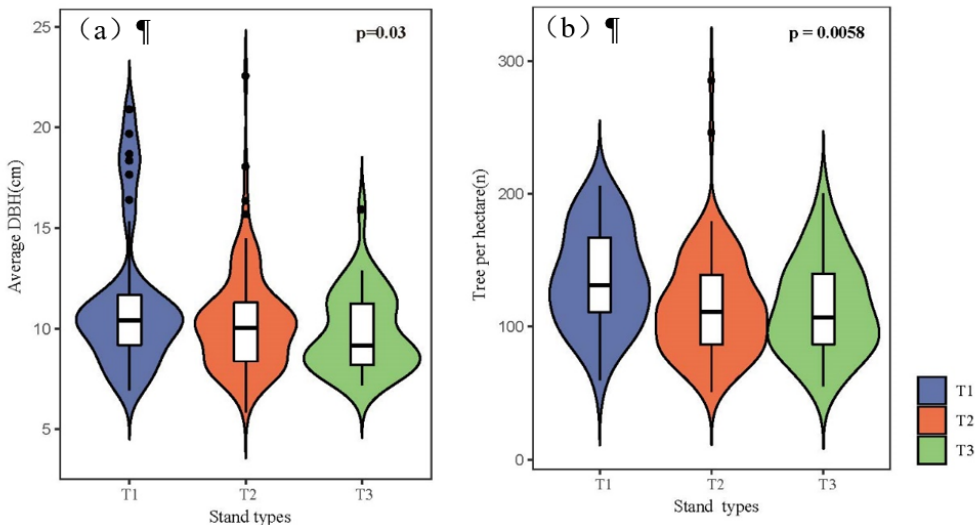
Figure 2 Analysis of variance of different stand types for the average DBH and number of trees per hectare (T1 denotes pure oak forest, T2 denotes oak-coniferous mixed forest, T3 denotes oak-broad-leaved mixed forest).

corresponding weight and threshold values. The optimal number ofneurons in the hidden layer is determined using the trial-and-error method. Some studies have shown that using too many hidden layers may lead to overtraining and increase the required number of training samples (Wekesa et al. 2019, Zhang et al. 2019). Therefore, a single hidden layer network was constructed in this study. That is, a three-layer back propagation (BP) network structure with one input layer, one hidden layer and one output layer was constructed. For example, the ANN structure is constructed with D, BA and N as input variables and using the frequency of the diameter class as the output variable (Figure 3).
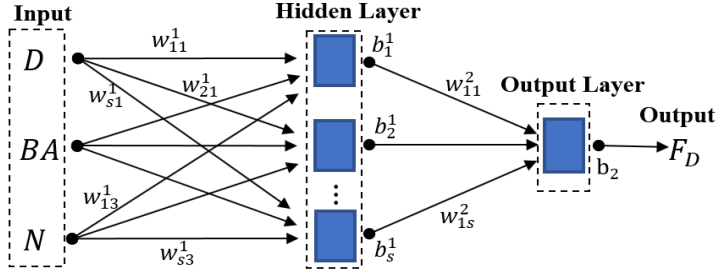
Figure 3 The topology of the artificial neural network (D, BA and N are the input variables, F_D is the output variable, w is the weight value, and b is the threshold value).

All neural networks were trained and tested using MATLAB 2016b software. The corresponding expression of the ANN is as follows:

$$F_D = \sum_{(i=1)}^{s} (w_{1i}^2 * tansig\ (w_{i1}^1 * r_D + w_{i2}^1 * r_{BA} + w_{i3}^1 * r_N + b_i^1)) + b_2 \tag{3}$$

$$tansig(z) = \frac{2}{1+e^{-2z}} - 1 \tag{4}$$

where F_D is the normalized value of the frequency of the diameter class, w_{i1}^1 is the weight from the first node in the input layer to neuron i in the hidden layer, w_{1i}^2 is the weight from neuron i in the hidden layer to neurons in the output layer, b_i^1 is the threshold value of neuron i in the hidden layer, b_2 is the threshold value of the output layer, i is the number of neurons (i =1, 2…s), tansig() is the transfer function, and r_D, r_{BA} and r_N are the normalized values of D, BA and N, respectively.

The input and output variables are normalized to a range of (−1,1) in the Neural Network Toolbox of MATLAB 2016b. The normalized expression is as follows:

$$\hat{x} = (\hat{x}_{max} - \hat{x}_{min}) * (x - x_{min}) / (x_{max} - x_{min}) + \hat{x}_{min} \tag{5}$$

The inverse normalized expression is as follows:

$$x = (\hat{x} - \hat{x}_{min}) * (x_{max} - x_{min}) / (\hat{x}_{max} - \hat{x}_{min}) + x_{min} \tag{6}$$

where $\hat{x}$ is the data after normalization, x is the data before normalization, *max* is the maximum value, *min* is the minimum value, $\hat{x}_{max}$=1, and $\hat{x}_{min}$=-1.

Model training

The purpose of the training data was to determine a suitable network structure, weight matrix and threshold matrix. In the following, we describe the training method using D, BA and N as input variables and F_D as the output variable. The datasets were randomly divided into training (70% of the data), validation (15%), and test (15%) sets. To determine a suitable network structure, the number of neurons should be optimized with a different number of hidden neurons. In the training process, the number of neurons was tentatively set to 4, 6, 8, 10, 12 and 14. Then, five good training results of each neuron were selected

after repeated training. The optimal network structure was determined by evaluating the training results of different neurons.

Evaluation indicator

To evaluate the predictive effect of the model, four statistics were computed: the coefficient of determination (R^2), residual sum of squared (RSS), mean squared error (MSE) and relative root mean square error (RRMSE). The larger the R^2 value and the smaller the RSS, MSE and RRMSE values are, the better the predictive performance is.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

$$MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n \tag{9}$$

$$RRMSE = 100 * \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-k}} / \bar{y} \tag{10}$$

where n is the number of observations, y_i is the observed value, $\hat{y}_i$ is the estimated value, $\bar{y}$ is the mean of the observed value, and k denotes the number of parameters.

Results

Comparison of the six distribution functions

The goodness of fit of six PDFs was tested using one-sample Kolmogorov-Smirnov (KS) tests. Table 3 shows the percentages of hypothesis acceptance for the 188 sample plots. The logistic distribution only gave a satisfactory fit for 34 out of 133 sample plots (33.3%). For the Weibull distribution, only 18 out of 133 sample plots (13.5%) failed to pass the KS test, showing that the diameter distributions followed the Weibull distribution.

Table 3 Results of the test of no difference between observed and predicted distributions within the 133 sample plots.

PDFs	Hypothesis rejected	Hypothesis accepted	Acceptance rate (%)
Normal	93	40	30.1
Logarithmic normal	24	109	82.0
Weibull	11	122	91.7
Gamma	18	115	86.5
Beta	32	101	75.9
Logistic	99	34	25.6

Note: criterion = one-sample Kolmogorov-Smirnov test; $\alpha = 0.05$; PDFs is the probability distribution functions

Evaluation and comparison of the two parameter prediction models

Stepwise regression model

The significant variables of each Weibull parameter model were identified using the stepwise regression method (Table 4). BA was selected as the independent variable for the prediction of parameter a (M1), D was selected for the prediction of parameter b (M2), and D^2 and N were selected for the prediction of parameter c (M3). All selected independent variables were found to be statistically significant, and the variance inflation factors (VIFs) were all less than 10, indicating no problematic collinearity. The R^2 values of M1 and M2 were only 0.134 and 0.258, respectively, but the overall multivariate regression models were statistically significant.

Dummy variable model

Stand type T was introduced into the regression models as a dummy variable to improve predictive performance (Table 5). Compared with model M2, the R^2 of the dummy variable model M5 increased by 0.005, while the RRMSE value decreased by 0.83%. Compared

Table 4 Stepwise regression for the prediction of parameters a, b and c.

Model#	Parameter	Independent variables					Evaluation index		
		Int	D	D^2	BA	N	R^2	RSS	RRMSE(%)
M1	a	4.877*			0.067*		0.134	2.684	2.89
M2	b	-5.507*	1.073*				0.983	18.528	6.86
M3	c	0.861*		0.001*		0.9×10^{-4}*	0.258	6.915	19.69

Note: Int is the intercept; D is the average DBH; D^2 is the square of DBH; BA is the basal area; N denotes the number of trees per hectare; R^2 is the coefficient of determination; RSS is the residual sum of squared; RRMSE is the relative root mean square error; symbol * indicates that the model parameter is significantly different from zero at the significance level of 0.05; M#: model numbers.

Table 5 The fitting results of the dummy variable model using stand type T as a dummy variable.

Model#	Parameter	Fixed variable				Dummy variable		Evaluation index		
		Int	D	D^2	BA	T1	T2	R^2	RSS	RRMSE
M4	a	4.877*			0.067*			0.134	2.684	2.89
M5	b	-5.353*	1.079*			-0.363*		0.987	14.305	6.03
M6	c	1.264*		0.001		-0.328*	-0.163*	0.392	4.858	16.51

Note: Int is the intercept; D is the average DBH; D^2 is the DBH square; BA is the basal area; R^2 is the coefficient of determination; RSS is the residual sum of squared; RRMSE is the relative root mean square error; symbol * indicates that the model parameter is significantly different from zero at the significance level of 0.05; M#: model numbers; T1 is the pure oak forest; T2 is the oak-coniferous mixed forest.

with model M3, the R^2 of the dummy variable model M6 increased by 0.134, while the RRMSE value decreased by 3.18%. The model accuracy did not change when the dummy variable was introduced into model M1. These results indicated that introducing dummy variables can increase the accuracy of the models with parameters b and c but the model with parameter a.

To compare the estimation accuracies from the multivariate regression models M2 and M3 and the dummy variable models M5 and M6, the significant differences among their average absolute residuals were tested, as shown in Figure 4. The average absolute residuals of model M5 were significantly lower than those in model M2 at the 0.05 level (P =0.0021) (Figure 4a). The average absolute residuals of model M6 were significantly lower than those in model M3 at the 0.1 level (P =0.088) (Figure 4b). These results indicated that introducing dummy variables

can significantly increase the accuracy of the models with parameters b and c.

Training results of the ANN

D was selected as one of the input variables in the neural network due to the collinearity and similar effects on the diameter distribution between D and D^2. The purpose of predicting the diameter distribution was to obtain the number or frequency of trees in each diameter class, so the diameter class (C) was considered a nonnegligible input variable. Therefore, D and C were used as the basic input variables for the neural network, and the stand variables BA, N and T were used as the input variables to be screened to determine an optimal network structure. Therefore, the oak forest ANN model was developed with C as the output variable and four variable combinations (D+C, D+C+BA, D+C+N, D+C+T) as the input variables.
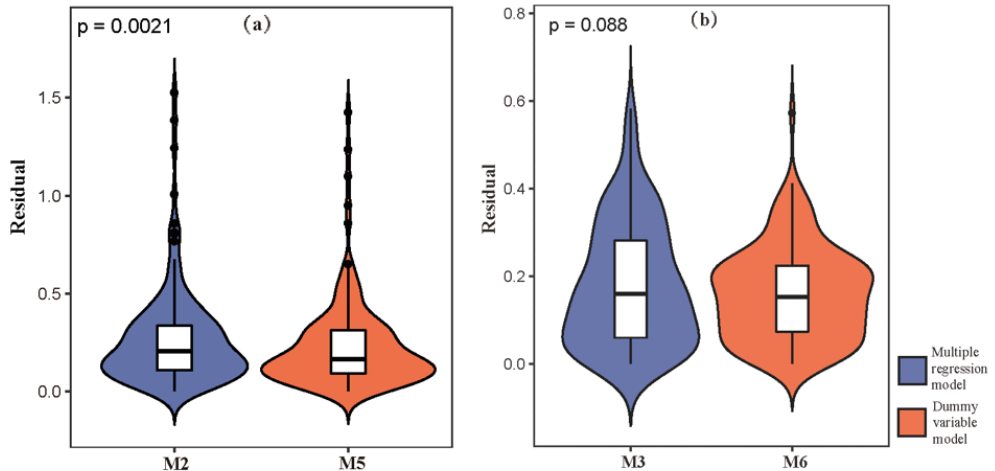
Figure 4 Analysis of variance for the average absolute residuals of M2 versus M5 and of M3 versus M6. M2 is the prediction model of parameter b, M3 is the prediction model of parameter c, M5 is the dummy variable model of parameter b, and M6 is the dummy variable model of parameter c.

Comparison of different numbers of neurons

The accuracy of the training results was influenced by the number of neurons during the training of the ANN. The training accuracy level with four neurons (S4) was significantly lower than that achieved from networks with more neurons (Figure 5). When D+C were used as input variables, there was no significant difference in the training accuracy from S6 to S14 (Figure 5a). When D+C+BA was used as the input variable, the training accuracy of S6 was significantly lower than the levels for S8 to S14, and there was no significant difference from S8 to S14 (Figure 5b). When D+C+N was used as the input variable, the training accuracies of S6 to S8 were significantly lower than those of S10 to S14, and there was no significant difference from S10 to S14 (Figure 5c). When D+C+T was used as the input variable, the results were consistent with the results obtained when D+C+N was used as the input variable (Figure 5d).

Optimal ANN

An analysis of variance was used to compare the optimal training results of different combinations of input variables (Table 6). No significant differences were observed between models M7 and M8, indicating that the introduction of BA did not substantially improve the prediction accuracy of the ANN. The training errors of models M9 and M10 were significantly lower than that of M7, indicating that the introduction of N or T substantially improved the prediction accuracy of the ANN. The training errors of model M10 were significantly less than those of the other models, indicating that the ANN model with D+C+T as input variables and 10 neurons was the optimal predictive model for oak forests.

The accuracy of the best neural network model M10 was evaluated using the training, verification and test sets, and the correlation coefficients between observations and predicted values are shown in Figure 6. The correlation coefficients in the training set, verification set, test set and overall data set exceeded 0.962. Thus, model M10 performed well in the training process and can be applied for the diameter distribution prediction of oak forests.
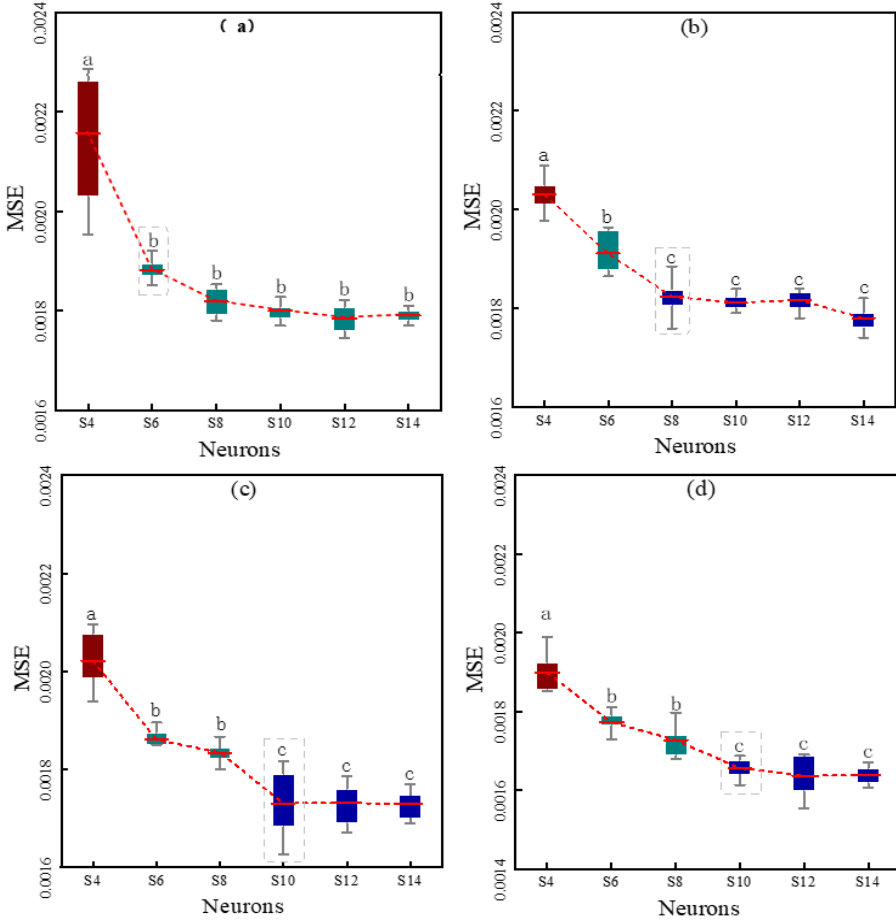
Figure 5 Analysis of variance comparing the MSE for different combinations of input variables and neurons for a), average DBH and diameter class as input variables, b) average DBH, diameter class and basal area as input variables, c) average DBH, diameter class and trees per hectare as input variables, d) average DBH, diameter class and stand type as input variables (MSE is the mean squared error, S# denotes the number of neurons, and different letters indicate that the MSE is different at a significance level of 0.05.

Table 6 Comparison of training errors for different combinations of input variables.

Model#	Input variable			Number of neurons	MSE
M7	C,	D		6	0.00187±0.00003a
M8	C,	D,	BA	8	0.00182±0.00005a
M9	C,	D,	N	10	0.00173±0.00008b
M10	C,	D,	T	10	0.00165±0.00003c

Note: C is the diameter class; D is the average DBH; BA is the basal area; N denotes the number of trees per hectare; T is the stand type; MSE is the mean squared error; M# denotes the model number; different letters indicate that the MSE is different at a significance level of 0.05.

13
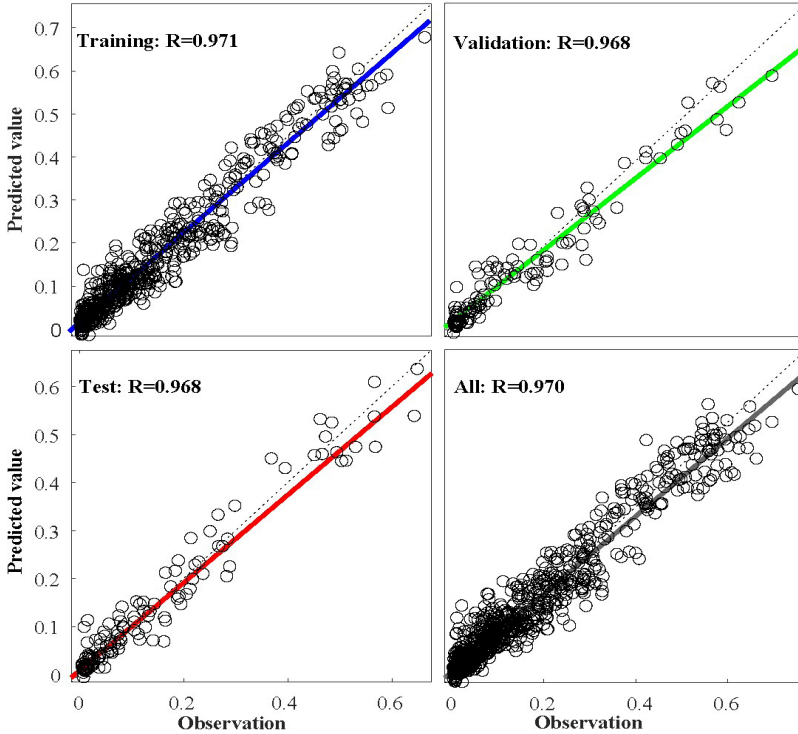
Figure 6 Correlation coefficient between observations and predicted value for the training set, verification set, test set and total set of the best neural network model (M10).

The weight and threshold matrix of the neural network model are as follows:

$$\{w_{1s}^2\} = \begin{bmatrix} -7.92 \\ 1.80 \\ -0.71 \\ 0.55 \\ 1.72 \\ -2.64 \\ -0.08 \\ 7.12 \\ 4.55 \\ 5.62 \end{bmatrix}, \{w_{s1}^1, w_{s2}^1, w_{s3}^1\} = \begin{bmatrix} -4.68 & -3.41 & -0.29 \\ -4.79 & 1.41 & 3.85 \\ -0.32 & 0.33 & -2.30 \\ -3.09 & 3.50 & 0.55 \\ -4.61 & 1.41 & 4.02 \\ -3.95 & 0.63 & 0.20 \\ 2.81 & -1.00 & 2.67 \\ 10.04 & -0.70 & 0.01 \\ 1.72 & 2.25 & 0.37 \\ 7.77 & -0.76 & -0.04 \end{bmatrix}, \{b_{1s}^1\} = \begin{bmatrix} 1.25 \\ 0.94 \\ -0.07 \\ 0.07 \\ -0.97 \\ 0.78 \\ -0.05 \\ -3.60 \\ 1.71 \\ 4.33 \end{bmatrix}, \{b^2\} = [-1.36]$$

Table 7 Significant difference test of the average absolute residuals from zero using the stepwise regression model (M11), the dummy variable model (M12) and the ANN model (M10) based on the modelling data.

Model#	R²	RRMSE (%)	RSS	Dummy variable		Evaluation index	
				T value	P value	T value	P value	T value	P value
M11	0.919	30.98	1.93	---		1.359	0.175	4.419	0.000
M12	0.926	29.55	1.75	-1.359	0.175	---		3.527	0.000
M10	0.941	26.65	1.43	-4.419	0.000	-3.527	0.000	---	

Note: R² is the coefficient of determination; RRMSE is the relative root mean square error; RSS is the residual sum of squared
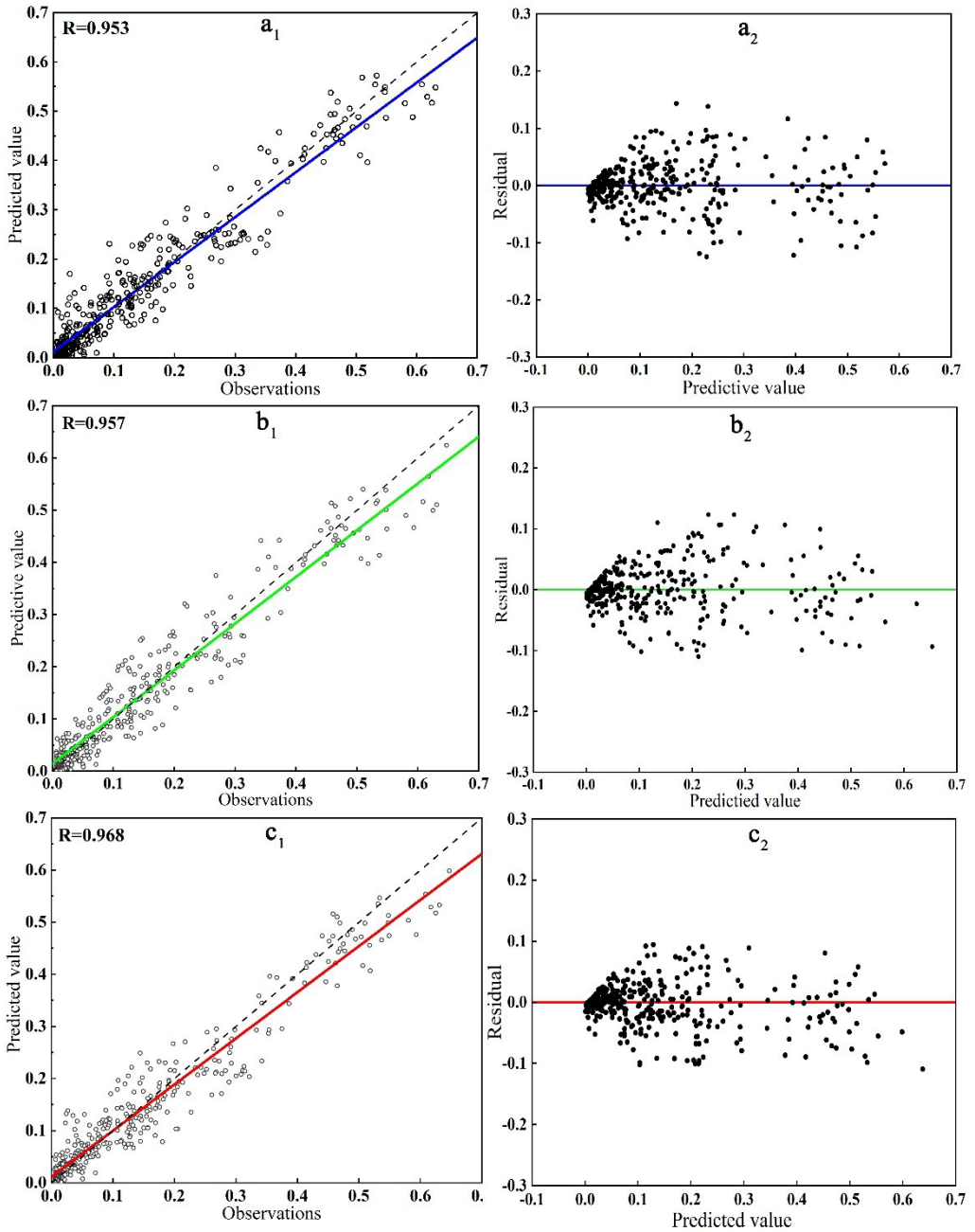
Figure 7 The correlation coefficient and residual distribution between observations and predicted values for a) the stepwise regression model, b) the dummy variable model, and c) the ANN model.

Comparison and test of models

The predictive performances of the diameter distribution of the stepwise regression, dummy variable model and ANN model were compared using a t-test. The significant differences among their average absolute residuals are shown in Table 7. Compared to the stepwise regression model (M11), the R^2 of the dummy variable model and ANN model increased by 0.007 and 0.022, while the RRMSE value decreased by 1.43% and 4.33%, respectively. The average absolute residual of the ANN model was significantly different from that of the stepwise regression model and dummy variable model. No significant differences were observed between the stepwise regression model and dummy variable model. These results show that the predictive performance of the diameter distribution based on ANN was more suitable than the stepwise regression model and dummy variable model.

The predictive performance of the three models was tested using validation samples (Figure 7). The correlation coefficients between the observations and predicted values were greater than 0.953. The residuals were evenly distributed with an absence of heteroskedasticity.

Discussion

The Weibull distribution is highly accurate in determining the diameter distribution of oak forests (Schreuder & Swank 1974). Carretero and Torres Alvarez (2013) argued that the Weibull distribution is a useful tool to predict the diameter distribution of oak forests. Our findings are consistent with those of other studies, and overall, the Weibull distribution showed very good applicability to oak forests. Although the Weibull distribution achieved the highest rate of accuracy (91.7%) of the six distribution functions, the gamma and lognormal distribution functions showed high rates of 86.5% and 82.0%, respectively.

Podlaski & Zasada (2008) found that the Weibull distribution and gamma distribution function were suitable for predicting the diameter distribution of multilayer stands in fir and birch mixed forests. Lima et al. (2017) also found that the lognormal distribution function exhibited good applicability in *Aspidosperma pyrifolium* and *Myracrodum urundeuva* forests after comparing the fitting results of several distribution functions. From the results of the present study, we conclude that the Weibull, gamma and lognormal distribution functions show good applicability to diameter distribution, and the prediction effects of distribution functions may differ by tree species or stand structure. Some studies found that the SB function also shows high levels of applicability and is even superior to the Weibull distribution (Mønness 1982). The prediction effect between Weibull and SB was not compared because comparing the applicability of distribution functions was not the main purpose of this study.

Some studies have verified that the dummy variable model exhibits a good fitting effect for qualitative data. Chen et al. (2017) found the dummy variable model to exhibit better performance than the general allometric equation. In this study, we introduced stand type as a dummy variable into the regression model to improve the predictive performance of the distribution function because related research has introduced dummy variables into the diameter distribution. We found the dummy variable to significantly increase the predictive performance of parameters b and c of the Weibull distribution (Table 5). Unfortunately, the prediction accuracy of the Weibull distribution in predicting the diameter distribution of oak secondary forests did not improve significantly (P=0.175<0.05). The possible reason for this result was that stand type may not fully reflect the differences in the diameter distribution due to the complex stand structure of oak secondary forests. Follow-up studies should prioritize other dummy

variables, such as forest layer and stand age.

The quantitative relationship between dependent variables and independent variables must be considered in advance using nonlinear regression to estimate model parameters, which makes it difficult to select mathematical functions and parameter estimation methods during modelling (Johnston et al. 2010). Compared to some definite mathematical models, the ANN offers substantial flexibility due to its ability to identify hidden relations in the data (Scrinzi et al. 2007). In this study, we found that the ANN model exhibited better predictive performance for the diameter distribution of oak forests than the traditional stepwise regression model or dummy variable model on the diameter distribution of oak forests. Compared with the stepwise regression model and the dummy variable model, the RRMSE of the ANN was significantly reduced by 1.33% and 2.9%, respectively. Meanwhile, the input variables of the ANN model included only three variables (C, D, T), and the independent variables of the multiple regression model and dummy variable model included five variables (D, D^2, BA, N, T), which indicated that the ANN model has higher practicability and simplicity. However, while a traditional regression model can be established using a small sample, a large dataset is needed for an ANN to avoid overfitting.

Unfortunately, factors such as stand age, forest layer and site index were not investigated in detail due to the complex stand structure and geomorphological structure of oak secondary forests, which led to few optional variables. Compared with the stepwise regression model and the dummy variable model, although the RRMSE of the ANN model in predicting the diameter distribution of oak secondary forest decreased significantly, the prediction accuracy may still have room for improvement. Follow-up studies could prioritize the use of other variables, such as site index and stand age.

Conclusion

In this study, we attempted to find a new method to predict the stand diameter distribution of oak secondary forests. Compared with traditional regression methods, an ANN can significantly improve the prediction accuracy of the diameter distribution, and ANN models have higher practicability and simplicity than traditional regression methods. The ANN model that utilized the average diameter, diameter class and stand type as input variables established in this study can be used to predict the diameter distribution of oak secondary forests.

Acknowledgment

References

Abbasi B., Rabelo L., Hosseinkouchack M., 2008. Estimating parameters of the three-parameter Weibull distribution using a neural network. European Journal of Industrial Engineering 2(4): 428-445. https://doi.org/10.1504/ejie.2008.018438

Álvarez-González J.G., Schröder J., Rodríguez Soalleiro R., Ruíz González A.D., 2002. Modelling the effect of thinnings on the diameter distribution of even-aged Maritime pine stands. Forest Ecology and Management 165(1-3): 57-65. https://doi.org/10.1016/S0378-1127(01)00650-8

Ashraf M.I., Zhao Z.Y., Bourque C.P.A., MacLean D.A., Meng F.R., 2013. Integrating biophysical controls in forest growth and yield predictions with artificial intelligence technology. Canadian Journal of Forest Research 43(12): 1162-1171. https://doi.org/10.1139/cjfr-2013-0090

Bailey R.L., Dell T.R., 1973. Quantifying diameter distributions with the Weibull function. Forest Science 19(2): 97-104. https://doi.org/10.1093/forestscience/19.2.97

Binoti D.H.B., Leite H.G., Nogueira G.S., Silva M.L.M., Garcia S.L.R., Cruz J.P., 2010. Three-parameter Weibull distribution in a diametric distribution model for thinning Eucaliptus stands. Revista Árvore 34(1): 147-156. https://doi.org/10.1590/S0100-

67622010000100016

Borders B.E., Patterson W.D., 1990. Projecting stand tables: a comparison of the Weibull diameter distribution method, a percentile-based projection method and a basal area growth projection method. Forest Science 36(2): 413-424. https://doi.org/10.1093/forestscience/36.2.413

Bowling E.H., Burkhart H.E., Burk T.E., Beck D.E., 1989. A stand-level multispecies growth model for Appalachian hardwoods. Canadian Journal of Forest Research 19(4): 405-412. https://doi.org/10.1139/x89-064

Brooks J.R., Borders B.E., Bailey R.L., 1992. Predicting diameter distributions for site-prepared loblolly and slash pine plantations. Southern Journal of Applied Forestry 16(3): 130-133. https://doi.org/10.1093/sjaf/16.3.130

Cai S., Kang X., Zhang L.X., Gong Z., Qin L., Chen P., 2010. A model for tree diameter distribution in stands based on artificial neural network. International symposium on intelligence processing and trusted computing (IPCT). IEEE Computer Society 332e336. https://doi.org/10.1109/iptc.2010.44

Cao Q.V., 2004. Predicting parameters of a Weibull function for modeling diameter distribution. Forest Science 50(5): 682-685. https://doi.org/10.1093/forestscience/50.5.682

Carretero A.C., Torres Alvarez E., 2013. Modelling diameter distributions of *Quercus suber* L. stands in "Los Alcornocales" Natural Park (Cadiz-Malaga, Spain) by using the two parameter Weibull functions. Forest Systems 22(1): 15-24. https://doi.org/10.5424/fs/2013221-02142

Chen D.S., Huang X.Z., Zhang S.G., Sun X.M., 2017. Biomass modeling of Larch (*Larix* spp.) Plantations in China based on the mixed model, dummy variable model, and bayesian hierarchical model. Forests 8(8): 268. https://doi.org/10.3390/f8080268

Clutter J.L., Bennett F.A., 1965. Diameter distributions in old-field slash pine plantations (Report 13). Macon: Report of Georgia Forest Research Council.

Clutter J.L., Harms W.R., Brister G.H., et al., 1984. Stand structure and yields of site-prepared loblolly pine plantations in the lower coastal plain of the Carolinas, Georgia, and north Florida. USDA Forest Service, Southeastern Forest Experiment Station, Asheville, General Technical Report SE-27.

Coomes D A., Allen R.B., 2007. Mortality and tree-size distributions in natural mixed-age forests. Journal of Ecology 95(1): 27-40. https://doi.org/10.1111/j.1365-2745.2006.01179.x

Corne S.A., Carver S.J., Kunin W.E., Lennon J.J., Van Hees W.W.S., 2004. Predicting forest attributes in southeast Alaska using artificial neural networks. Forest Science 50(2): 259-276. https://doi.org/10.1093/forestscience/50.2.259

Dande P., Samant P., 2018. Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review. Tuberculosis 108: 1-9. https://doi.org/10.1016/j.tube.2017.09.006

Diamantopoulou M.J., 2005. Artificial neural networks as an alternative tool in pine bark volume estimation. Computers and Electronics in Agriculture 48(3): 235-244. https://doi.org/10.1016/j.compag.2005.04.002

Dorsett D., Webster J.T., 1983. Guidelines for variable selection problems when dummy variables are used. The American Statistician 37(4a): 337-339. https://doi.org/10.1080/00031305.1983.10483138

García O., 1981. Simplified method of moments estimation for the Weibull distribution. New Zealand Journal of Forestry Science 11(3): 304-306.

Goelz J.C.G., Leduc D.J., 2002. A model describing growth and development of longleaf pine plantations: consequences of observed stand structures of structure of the model. Gen. Tech. Rep. SRS–48. Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station. pp. 438-442.

Gurney K., 1997. An introduction to neural networks. London: UCL Press. 234 pp.

Hafley W.L., Schreuder H.T., 1977. Statistical distributions for fitting diameter and height data in even-aged stands. Canadian Journal of Forest Research 7(3): 481-487. https://doi.org/10.1139/x77-062

Hou Y.S., Chen X.L., Sun G.J., 2017. Oaks Management. China Forestry Press: Beijing, China. 9 p

Huang J.R., 2000. Studies on the suitability of Weibull distribution in Masson pine plantations. Guizhou Forestry Science and Technology 1: 7-13.

Hyink D.M., Moser J.W., 1983. A generalized framework for projecting forest yield and stand structure using diameter distributions. Forest Science 29: 85-95. https://doi.org/10.1093/forestscience/29.1.85

Jiang L.C., Brooks J.R., 2009. Predicting diameter distributions for young longleaf pine plantations in Southwest Georgia. Southern Journal of Applied Forestry 33(1): 25-28. https://doi.org/10.1093/sjaf/33.1.25

Johnson N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrica 36(1-2): 149-176. https://doi.org/10.2307/2332539

Johnston M.H., Williamson T.B., Munson A.D., Ogden A.E., Moroni M.T., Parsons R, Price D.T., Stadt J.T., 2010. Climate change and forest management in Canada: impacts, adaptive capacity and adaptation options. A State of Knowledge report. Sustainable Forest Management Network, Edmonton. 54 p.

Kayes I, Deb J.C., Comeau P., Das S., 2012. Comparing normal, lognormal and Weibull distributions for fitting diameter data from Akashmoni plantations in the north-eastern region of Bangladesh. Southern Forests 74(3): 175-181. https://doi.org/10.2989/20702620.2012.717409

Kilkki P., Maltamo M., Mykkanen R., et al., 1989. Use of the Weibull function in estimating the basal area DBH-distribution. Silva Fennica 23(4): 311-318. https://doi.org/10.14214/sf.a15550

Leak W.B., 1964. An expression of diameter distribution

for unbalanced, uneven-aged stands and forests. Forest Science 10(1): 39-50. https://doi.org/10.1093/forestscience/10.1.39

Lee F.Y., 1974. On the dummy variable technique and covariance analysis in testing equality among sets of coefficients in linear regressions: An expository note. The Journal of Financial and Quantitative Analysis 9(3): 491-495. https://doi.org/10.2307/2329876

Lee Y.J., Coble D.W., 2006. A new diameter distribution model for unmanaged loblolly pine plantations in East Texas. Southern Journal of Applied Forestry 30(1): 13-20. https://doi.org/10.1007/s10035-008-0092-4

Li W.Y., Wang B., Li G.C., 2001. Ecological benefits and economic values of Oaks species and countermeasures for their resource protection. For. Sci. Techol 8: 13-15. https://doi.org/10.13456/j.cnki.lykt.2001.08.004

Lima R.A.F. de, Batista J.L.F., Prado P.I., 2015. Modeling Tree Diameter Distributions in Natural Forests: An Evaluation of 10 Statistical Models. Forest Science 61(2): 320-327. https://doi.org/10.5849/forsci.14-070

Lima R.B. de, Bufalino L., Alves F.T., da Silva J.A.A., Ferreira R.L.C., 2017. Diameter distribution in a Brazilian tropical dry forest domain: predictions for the stand and species. Anais Da Academai Brasileira De Ciencias 89(2): 1189-1203. https://doi.org/10.1590/0001-3765201720160331

Lindsay S.R., Wood G.R., Woollons R.C., 1996. Stand table modeling through the Weibull distribution and usage of skewness information. For. Ecol. Manage 81: 19-23. https://doi.org/10.1016/0378-1127(95)03669-5

Little S.N., 1983. Weibull diameter distributions for mixed stands of western conifers. Canadian Journal of Forest Research 13 (1): 85-88. https://doi.org/10.1139/x83-012

McCrohan K.F., Harvey J.W., 1989. A comparison of dummy variable versus traditional multiple discriminant function analysis. Health Marketing Quarterly 6(4): 147-157. https://doi.org/10.1300/J026v06n04_11

Meyer H.A., 1952. Structure, growth, and drain in balanced uneven-aged forests. Journal of Forestry 50(2): 85-92. https://doi.org/10.1093/jof/50.2.85

Miranda R., Fiorentin L., Péllico Netto S., Juvanhol R., Corte A.D., 2018. Prediction system for diameter distribution and wood production of *Eucalyptus*. Floresta e Ambiente 25(3): e20160548. https://doi.org/10.1590/2179-8087.054816

Mønness E.N., 1982. Diameter distributions and height curves in even-aged stands of *Pinus sylvestris* L. Reports of the Norwegian Forest Research Institute 36(15): 43 p.

Nelson T.C., 1964. Diameter distribution and growth of loblolly pine. Forest Science 10(1): 105-114. https://doi.org/10.1093/forestscience/10.1.105

Nixon K.C., 1993. Infrageneric classification of *Quercus* (Fagaceae) and typification of sectional names. Annals of Forest Science 36: 25-34. https://doi.org/10.1051/forest:19930701

Oettel J., Lapin K., Kindermann G., Steiner H., Schweinzer K.M., Frank G., Essl F., 2020. Patterns and drivers of deadwood volume and composition in different forest types of the Austrian natural forest reserves. Forest Ecology and Management 463: 118016. https://doi.org/10.1016/j.foreco.2020.118016

Özçelik R., Diamantopoulou M.J., Crecente-Campo F., Eler U., 2013. Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. Forest Ecology and Management 306: 52-60. https://doi.org/10.1016/j.foreco.2013.06.009

Perea R., López-Sánchez A., Dirzo R., 2017. Differential tree recruitment in California oak savannas: Are evergreen oaks replacing deciduous oaks? Forest Ecology and Management 399: 1-8. https://doi.org/10.1016/j.foreco.2017.05.018

Pérez-López E., Santiago-García W., Quiñonez-Barraza G., Rodríguez-Ortiz G., Santiago-García E., Ruiz-Aquino F. 2019. Estimation of diameter distributions for *Pinus patula* with the Weibull function. Madera y Bosques 25(3): e2531626. http://doi.org/10.21829/myb.2019.2531626

Podlaski R., Zasada M., 2008. Comparison of selected statistical distributions for modelling the diameter distributions in near-natural Abies-Fagus forests in the Świętokrzyski National Park (Poland). European Journal of Forest Research 127(6): 455-463. https://doi.org/10.1007/s10342-008-0229-3

Pogoda P., Ochał W., Orzeł S., 2019. Modeling diameter distribution of black alder (*Alnus glutinosa* (L.) Gaertn.) stands in Poland. Forests 10(5): 412. https://doi.org/10.3390/f10050412

Quiñonez-Barraza G., De los Santos Posadas H.M., Cruz-Cobos F., Martinez A.V., Perez G.Á., Valverde G.R., 2015. Dynamic modeling for diameter distribution on *Pinus* mixed stands in Durango, Mexico. Madera y Bosques 21(2): 59-71.

Souza Retslaff F.A. de, Figueiredo Filho A., Dias A.N., Bernett L.G., Figura M.A., 2012. Growth and yield prognosis in diameter classes for thinning *Eucalyptus grandis* stands in Brazil south. Revista Árvore 36(4): 719-732. https://doi.org/10.1590/S0100-67622012000400013

Samarasinghe S., 2006. Neural networks for applied sciences and engineering. Florida, USA: Taylor and Francis Inc. 608 p.

Schreuder H.T., Swank W.T., 1974. Coniferous stands characterized by the Weibull distribution. Canadian Journal of Forest Research 4: 518-523. https://doi.org/10.1139/x74-075

Schütz J.P., Rosset C., 2020. Performances of different methods of estimating the diameter distribution based on simple stand structure variables in monospecific regular temperate European forests. Annals of Forest Science 77(47): 1-11. https://doi.org/10.1007/s13595-020-00951-3

Scrinzi G., Marzullo L., Galvagni D., 2007. Development of a neural network model to update forest distribution data for managed alpine stands. Ecological Modelling 206(3-4): 331-346. https://doi.org/10.1016/j.ecolmodel.2007.04.001

Siipilehto J., 1999. Improving the accuracy of predicted basal-area diameter distribution in advanced stands by determining stem number. Silva Fennica 33(2): 281-301. https://doi.org/10.14214/sf.650

Smalley G.W., Bailey R.L., 1974. Yield tables and stand structure for loblolly pine plantations in Tennessee, Alabama, and Georgia highlands. USDA For. Serv. Res: Paper SO-96. 81 p.

Sun S.C., Cao Q.V., Cao T.J., 2019. Characterizing diameter distributions for uneven-aged pine-oak mixed forests in the Qinling mountains of China. Forests 10(7): 596. https://doi.org/10.3390/f10070596

Wang M., Borders B.E., Zhao D., 2008. An empirical comparison of two subject-specific approaches to dominant heights modeling-the dummy variable method and the mixed model method. Forest Ecology and Management 255(7): 2659-2669. https://doi.org/10.1016/j.foreco.2008.01.030

Wang W.W., Chen X.Y., Zeng W.S., Wang J.J., Meng J.H., 2019. Development of a mixed-effects individual-tree basal area increment model for Oaks (*Quercus* spp.) considering forest structural diversity. Forests 10(6): 474. https://doi.org/10.3390/f10060474

Weibull W., 1951. A statistical distribution function of wide applicability. J. Appl. Mech 18: 293-297. https://doi.org/10.1093/qjmam/6.4.453

Wekesa J.S., Luan Y, Chen M, Meng J., 2019. A hybrid prediction method for plant lncRNA-protein interaction. Cells 8(6): 521. https://doi.org/10.3390/cells8060521

Zankis S.H., 1979. A simulation study of some simple estimations for the 3-parameter Weibull distribution. Journal of Statistical Computation and Simulation 9: 101-116. https://doi.org/10.1080/00949657908810302

Zhang L., Ye Y., Zeng W., Chiaradia A., 2019. A systematic measurement of street quality through multi-sourced urban data: a human-oriented analysis. Int J Environ Res Public Health 16(10): 1782. https://doi.org/10.3390/ijerph16101782