Prediction model of *Populus simonii* seedlings based on growth characteristics in China

Huishuo Ge¹, Xiaoyu Zhang^{1,a}

¹ College of Science, Beijing Forestry University, No. 35, Qinghua East Road, Beijing 100083, P. R. China

Abstract. In this paper, we originally apply the BP neural network to predict the plant height of *Populus simonii* seedlings. Firstly, we explore correlation among the section length variables of *Populus simonii* seedlings in four growth periods by using principal component analysis and hierarchical clustering method, which obtain 5 principal components. In addition, we utilize Fuzzy C-Means Clustering (FCM) to classify the *Populus simonii* seedlings, and are obviously classified into two subpopulations. Furthermore, we utilize BP neural network to establish seedlings height growth model and aboveground biomass prediction model, respectively. Through numerical experiments, prediction accuracy of the seedling height growth models in four periods reaches about 84.89%. Meanwhile, the prediction accuracies of stem and leaf fresh weight and stem and leaf dry weight are 91.15% and 83.79%, respectively. This paper provides an effective method for studying phenotypic characteristics and predicting the height of *Populus simonii* seedlings, which supplies a reference for genome-wide association analysis.

1 Introduction

For the past few years, *P. simonii* has become the main species of protection forest in China for its adaptability, drought resistance and cold resistance. It has important values in the fields of wind-break and sand fixation, civil architecture and papermaking [1]. Therefore, the research of *P. simonii*'s growth characteristics is of practical significance in the selection of seedlings and genome-wide association analysis. On the other hand, biomass is not only the energy base and nutrient source of ecosystem operation, but also the important embodiment of ecosystem productivity [2]. It is also an important index to reflect the function of community or ecosystem [3].

At present, there are many researches on the growth of woody plants. In 2016, Peng et al. explored the growth rhythm of cuttage seedlings accorded with S-shaped curve [4]. On the other hand, Tang et al. took advantage of the modified Richards model, exponent function and the growth prediction functions to predict the tree height, volume and diameter at breast height (DBH) for *P. simonii* in the gully rolling region of Shaanxi Loess Plateau in 2000 [5]. And in 2001 Zhang et al. used logistic model to explore the growth law of *P. simonii* [6]. Furthermore, in 2014, Zhao et al. utilized logistic equation model and power function to estimate the biomass [7]. In 2017, Liu et al. established the regression models of aboveground biomass of *Populus euphratica* based on the

analysis of the different ages [8]. However, most of the current researches for predicting seedling height and aboveground biomass of *P. simonii* are based on time-dependent longitudinal models, but few studies reflect the relationship between dependent variables and corresponding phenotypic variables. Besides, Zeng et al. used BP neural network model to construct the aboveground biomass model of grassland in the source area of the Three Rivers in 2017, but the simulated and measured R² of the aboveground biomass of alpine meadow and alpine meadow only were 0.73 and 0.79, respectively [9].

In this paper, we originally apply the BP neural network to establish the prediction model of plant height with phenotypic variables (leaf number, section number, section length) and the aboveground biomass prediction model for 442 individuals as core populations that represented almost the entire geographic distribution of *P. simonii* in China. Firstly, we use principal component analysis to obtain 5 principal components of the section length variables of *P. simonii* seedlings in four growth periods. In addition, we apply FCM Clustering obviously to classify the *P. simonii* seedlings into two subpopulations, which shows significant differences in phenotypic traits (plant height, leaf number and section number). Furthermore, we let phenotypic variables be independent variables in order to propose the prediction model of plant height with phenotypic variables (leaf

Table 1. Principal component analysis results in four growth periods

Date	SS	TSLN	PC	SL	CV /%
April 29	78	16	5	10~12 13~15	≥80.0

^a Corresponding author: xyzhang@bjfu.edu.cn

				04~07 02~03 1,8,9,16	
				12~18 06~11	
May 20	78	19	5	02~05 19 1	≥80.0
				14~19 08~13	
June 04	79	23	5	20~23 01~04 05~07	≥80.0
				14~21 08~13	
June 17	77	23	5	01~04 05~07 22~23	≥80.0

*SS, TSLN, PC, SL and CV represent sample size, total section length number, principal component, section length and cumulative variance.

number, section number, section length) and the aboveground biomass prediction model. Through the numerical experiments, the prediction accuracy of the seedling height growth models in four periods would reach about 84.89% by utilizing BP neural network. Meanwhile, the prediction accuracies of stem and leaf fresh weight and stem and leaf dry weight would reach about 91.15% and 83.79%, respectively. Our research provides an effective method for studying phenotypic characteristics and predicting the height of *P. simonii* seedlings, which supplies a reference for genome-wide association analysis.

2 Plant material

In 2017, the 385 individuals with phenotypic and physiological traits sampled from 16 different localities in China were studied in this paper. *P. simonii* seedlings were growing under the conditions of 16 hours light and 8 hours dark in a greenhouse at Beijing Forestry University, Beijing, China (40°0'N, 116°20'E). The phenotypic traits of growth were recorded about every 20 days, including plant height, leaf number, section number and length. Especially, we recorded the data of 57 phenotypic and physiological traits individuals. The steps of our experiment are as follows: 1) cut the corresponding tissues of plants, then quickly put them into the known weight aluminum boxes and weigh the fresh weight (Wf); 2) put the plant tissue together with the aluminum box into the oven which has been heated to 105°C, fixing 30min, and then bake to the constant weight at 80°C, so got the dry weight (Wd). The relative water content of samples can be obtained by Wf and Wd.

3 Results

3.1 Classification of *P. simonii* seedlings based on principal component analysis and fuzzy c-means clustering

3.1.1 Principal component analysis

In order to compress the section length variables, we use principal component analysis method to reduce the variable dimension. The principal component analysis transforms the original variables into new axes, or principal components (PCs), which are orthogonal, so that the data are uncorrelated with each other. As we all know, the goal of PCA is to replace a large number of correlated variable with a smaller number of uncorrelated variables while capturing as much information in the original variables as possible [10]. In 2000, Destefanis et al. analyzed the first three principal components of beef characteristics, which explained about 63% of total variability [11]. In this paper, we obtain 5 principal components of the section length variable of *P. simonii* seedlings in four growth periods, and the contribution rate would reach about 80%.

Table 1 shows that the section length variables of four stages of *P. simonii* seedlings are all integrated into five principal components. We combine the section variables of larger loads into a principal component and reduce the original variable to 5 dimensions. Thus, the section length variables of *P. simonii* seedlings are roughly divided into five parts from bottom to top, among which the last two principal components on April 29 and May 20 are more dispersed and less representative, while the other principal components are more evenly distributed.

3.1.2 Correlation analysis of phenotypic variables of *P. simonii* seedlings

Furthermore, we explore the relationships among the phenotypic characteristics of *P. simonii* and the adjacent section length variables by hierarchical clustering method (the result can refer to Figure 1).

Figure 1. Hierarchical cluster diagram of phenotypic variables of *P. simonii* seedlings

From Fig. 1, the phenotypic variables of *P. simonii* seedlings including section length(m1~m16), plant height(y), leaf number(x1) and section number(x2), which are divided into three categories. Apparently, the continuous section length variables are assigned to the same category, and they have significant correlation. The results are similar to the principal component analysis results, and reveal the internal growth law of *P. simonii*.

3.1.3 Fuzzy c-means clustering

Among the 385 *P. simonii* samples collected in China, we apply FCM clustering algorithm and NbClust package (version3.0(<https://cran.rproject.org/web/packages/NbClust/index.html>)) in the R software to cluster *P. simonii* samples based on phenotypic characteristics in order to analyze the difference of *P. simonii* seedlings in different regions. In detail, the steps of the FCM clustering algorithm are as follows:

1. Determine the optimal clustering number c , denote the iteration stop threshold be ε . And let the iteration counter be $t=0$.

2. Let X be a matrix:

$$X_{n \times m} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where m represents the number of characteristic variables, n represents the number of samples and $x_j (j=1,2,\dots,n)$ represents the j^{th} sample. We firstly determine the optimal clustering number $c (2 \leq c \leq n)$. Thus we initialize the fuzzy membership matrix U :

$$U_{c \times n} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{c1} & u_{c2} & \cdots & u_{cn} \end{pmatrix} = (u_{ij})_{c \times n}$$

where u_{ij} represents the membership relation between

sample x_j , and satisfies $0 \leq u_{ij} \leq 1, \sum_{i=1}^c u_{ij} = 1$. Meanwhile,

we obtain the matrix of random-initial clustering centers (c is the number of clustering center):

$$P_{c \times n} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{c1} & p_{c2} & \cdots & p_{cn} \end{pmatrix} = (p_{ij})_{c \times n}$$

3. In our work, during the clustering analysis, objective function is defined:

$$J = \sum_{i=1}^c \sum_{j=1}^n (u_{ij})^h \|x_j - p_i\|^2$$

satisfying the additional conditions:

$$\begin{cases} \min J = \min \left\{ \sum_{i=1}^c \sum_{j=1}^n (u_{ij})^h \|x_j - p_i\|^2 \right\} \\ = \sum_{j=1}^n \min \left\{ \sum_{i=1}^c (u_{ij})^h \|x_j - p_i\|^2 \right\} \\ \sum_{i=1}^c u_{ij} = 1 \end{cases} \quad (1)$$

Solving the equation (1), we would have:

$$u_{ij} = \frac{\|x_j - p_i\|^{-\frac{2}{h-1}}}{\sum_{i=1}^c \|x_j - p_i\|^{-\frac{2}{h-1}}}, \quad p_i = \frac{\sum_{j=1}^n (u_{ij})^h x_j}{\sum_{j=1}^n (u_{ij})^h} \quad (2)$$

4. Update fuzzy cluster matrix $U^{(t)}$ and cluster center matrix $P^{(t)}$ according to the equation (2)

5. If $|P^{(t)} - P^{(t-1)}| < \varepsilon$, stop calculation, output fuzzy membership matrix U and cluster center matrix P . Otherwise, let $t = t + 1$, and return to step 2.

We consider optimal number by the application of the NbClust package of R language software. The specific results are shown in Figure 2. There are seven indicators of the 26 clustering evaluation indexes suggested to be clustered into two categories ($N_c = 2$). The bar graph is visualized using NbClust package of R language software.

Next, we take advantage of Euclidean distance to perform FCM clustering, and let the maximum number of iterations be $t=100$, the iteration stop threshold $\varepsilon = 10^{-5}$. Through multiple simulations and calculations, the seedlings of *P. simonii* are divided into two distinct subpopulations (refer to Figure 3 and Table 2).

Further analysis, we find that the samples of the first subpopulation are mainly distributed in Shaanxi Province, Hebei Province, Liaoning Province, and Qinghai Province. These provinces and cities are mainly located in the middle and lower reaches of the Yellow River. The samples of second subpopulation are mainly from Shaanxi, Qinghai, and Henan provinces, all located in the upper and middle reaches of the Yellow River.

Figure 2. The number of recommended clustering

In four growth periods, the first subpopulation samples are characterized by lower plant height (with an average of only 13.37cm, 20.38cm, 21.49cm and 21.87cm, respectively), less leaf numbers and less section numbers. For the second subpopulation, samples have higher plant height (with an average of 18.31cm, 30.03cm, 31.02cm and 32.09cm, respectively). Concretely, on April 29, there is a 5cm difference in plant height between the two subpopulations of samples. At June 17, this difference unexpectedly reaches 10.22cm. The difference between leaf numbers and section numbers are all fluctuated between 2~4 in four growth periods. Therefore, the phenotypic characteristics of *P. simonii* seedlings in different geographical locations are significantly different.

Figure 3. Clustering results of *P. simonii* seedlings in four periods

Table 2. Clustering results of FCM clustering algorithm in *P. simonii* seedlings

Region	April 29th		May 20th		June 4th		June 17th	
	1st sub	2nd sub	1st sub	2nd sub	1st sub	2nd sub	1st sub	2nd sub
Chicheng	3	4	2	4	2	4	2	4
Zhangjiakou	2	1	2	1	2	1	2	1
Fuxian	7	6	6	6	5	7	6	6
Gaoling	6	5	7	4	6	5	5	6
Langao	1	0	1	0	0	1	0	1
Linyou	1	1	1	1	1	1	1	1
Zhongning	2	4	2	4	2	4	2	4
Xinghai	3	1	2	2	3	1	3	1
Dulan	3	0	2	1	2	0	2	1
Menyuan	2	0	2	0	2	0	2	0
Neimeng	1	1	2	0	2	0	2	0
Henan	0	1	0	1	0	1	0	1
Yichuan	8	6	9	5	8	7	7	7
Songxian	2	0	2	1	2	1	1	1
Taor-ting	0	2	1	1	1	1	1	1
Gaizhou	2	0	1	0	1	0	1	0
Jinzhou	1	0	1	0	1	0	1	0
Guangdong	1	0	1	0	1	0	1	0
Fumin	0	1	0	1	0	1	0	1
Yushu	0	1	1	0	0	1	1	0

*1st sub are 2nd sub are represent First subpopulation and Second subpopulation.

3.2 Prediction model of *P. simonii* seedlings based on BP neural network

3.2.1 Prediction model based on BP neural network

In this part, we establish a prediction model based on BP neural network to research the relationship between the plant height and *P. simonii*'s phenotypic characteristics (leaf number, section number, section length).

The steps of BP neural network algorithm are as follows:

1. Network initialization. Give each connection weight a random interval of one interval $(-1, 1)$. Set the error function is e , the given calculation accuracy is ε and the maximum number of learning times is M .

2. Select the k input sample and the corresponding expected output randomly:

$$x(k) = (x_1(k), x_2(k), \dots, x_n(k))$$

$$d_o(k) = (d_1(k), d_2(k), \dots, d_q(k))$$

3. Calculating the input and output of each neuron in the hidden layer:

$$h_{ih}(k) = \sum_{i=1}^n w_{ih} x_i(k) - b_{ih} \quad h = 1, 2, \dots, p$$

$$h_{oh}(k) = f(h_{ih}(k)) \quad h = 1, 2, \dots, p$$

$$y_{io}(k) = \sum_{h=1}^p w_{ho} h_{oh}(k) - b_o \quad o = 1, 2, \dots, q$$

$$y_{oo}(k) = f(y_{io}(k)) \quad o = 1, 2, \dots, q$$

4. Using the expected output and actual output of the network, the partial derivative $\delta_o(k)$ of the error function to each neuron in the output layer can be obtained:

$$\begin{aligned} \frac{\partial e}{\partial w_{ho}} &= \frac{\partial e}{\partial y_{io}} \frac{\partial y_{io}}{\partial w_{ho}} \\ \frac{\partial y_{io}(k)}{\partial w_{ho}} &= \frac{\partial(\sum_h w_{ho} h_{oh}(k) - b_o)}{\partial w_{ho}} = h_{oh}(k) \\ \frac{\partial e}{\partial y_{io}} &= \frac{\partial(\frac{1}{2} \sum_{o=1}^q (d_o(k) - y_{oo}(k))^2)}{\partial y_{io}} \\ &= -(d_o(k) - y_{oo}(k)) y_{oo}'(k) \\ &= -(d_o(k) - y_{oo}(k)) f'(y_{io}(k)) = -\delta_o(k) \end{aligned}$$

5. Utilizing the connection weight between hidden layer and output layer. The calculated error function of the output $\delta_o(k)$ of the output layer and the output of the

hidden layer on the partial derivative $\delta_h(k)$ of each neuron of the hidden layer can be obtained:

$$\begin{aligned} \frac{\partial e}{\partial w_{ho}} &= \frac{\partial e}{\partial y_{io}} \frac{\partial y_{io}}{\partial w_{ho}} = -\delta_o(k) h_{oh}(k) \\ \frac{\partial e}{\partial w_{ih}} &= \frac{\partial e}{\partial h_{ih}(k)} \frac{\partial h_{ih}(k)}{\partial w_{ih}} \\ \frac{\partial h_{ih}(k)}{\partial w_{ih}} &= \frac{\partial(\sum_{i=1}^n w_{ih} x_i(k) - b_{ih})}{\partial w_{ih}} = x_i(k) \\ \frac{\partial e}{\partial h_{ih}(k)} &= \frac{\partial(\frac{1}{2} \sum_{o=1}^q (d_o(k) - y_{oo}(k))^2)}{\partial h_{ih}(k)} \frac{\partial h_{oh}(k)}{\partial h_{ih}(k)} \\ &= \frac{\partial(\frac{1}{2} \sum_{o=1}^q (d_o(k) - y_{io}(k))^2)}{\partial h_{oh}(k)} \frac{\partial h_{oh}(k)}{\partial h_{ih}(k)} \\ &= -(\sum_{o=1}^q \delta_o(k) w_{ho}) f'(h_{ih}(k)) = -\delta_h(k) \end{aligned}$$

6. Applying the output of each neuron $\delta_o(k)$ in the output layer and the output of each neuron in the hidden layer to modify the connection weight $w_{ho}(k)$, that is:

$$\begin{aligned} \Delta w_{ho}(k) &= -\mu \frac{\partial e}{\partial w_{ho}} = \mu \delta_o(k) h_{oh}(k) \\ w_{ho}^{N+1} &= w_{ho}^N + \eta \delta_o(k) h_{oh}(k) \end{aligned}$$

7. Using the input parameter of each neuron $\delta_h(k)$ in the hidden layer to modify the connection weight:

$$\begin{aligned} \Delta w_{ih}(k) &= -\mu \frac{\partial e}{\partial w_{ih}} \\ &= -\mu \frac{\partial e}{\partial h_{ih}(k)} \frac{\partial h_{ih}(k)}{\partial w_{ih}} = \delta_h(k) x_i(k) \\ w_{ih}^{N+1} &= w_{ih}^N + \eta \delta_h(k) x_i(k) \end{aligned}$$

8. Calculating global error:

$$E = \frac{1}{2m} \sum_{k=1}^m \sum_{o=1}^q (d_o(k) - y_{oo}(k))^2$$

9. To determine whether the network error reaches the preset error value, the algorithm would ends when the error reaches the preset accuracy or the number of learning is greater than the maximum number set. Otherwise, select the next learning sample and corresponding expected output, and return to step 3 for the next round of learning. Through the above BP neural network algorithm, we obtain ideal prediction

Figure 4. Prediction results of *P. simonii* seedling height in four periods

results of the seedling height of *P. simonii* (refer to Figure 4), which show that the phenotypic variables of *P. simonii* seedlings are closely related to their own growth, and the prediction accuracy of the four stages is 84.89%. Numerical experiment point out that BP neural network has properties of not only higher accuracy but also good stability.

3.2.2 Aboveground biomass prediction model of *Populus simonii* seedlings

The physiological characteristics of *P. simonii* seedlings can be obtained by using the observed phenotype characteristics. We utilize BP neural network to obtain the predicted results of fresh and dry weight of stem and leaf (Figure 5 and 6), and the numerical results are shown in Table 3. We utilize three commonly numerical prediction and evaluation criteria as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2$$

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

where $\bar{y}_i$ is the mean of test set, and $\hat{y}_i$ is the corresponding predicted value.

Figure 5. Prediction of fresh weight of stem and leaf

Figure 6. Prediction of dry weight of stem and leaf

Table 2. Physiological characteristics prediction of 57 *Populus simonii* seedling samples

Stem and leaf fresh weight				Stem and leaf dry weight			
Prediction accuracy	e	MAE	MSE	Prediction accuracy	e	MAE	MSE
91.15%	0.0201	1.4336	0.2618	83.79%	0.0048	0.7949	0.0621

We obtain that the predicted results of the aboveground biomass of poplar seedlings, the prediction accuracies of stem and leaf fresh weight and stem and leaf dry weight are 93.18% and 83.36%, respectively. The average absolute errors are 1.6224 and 0.8116, and the relative errors are 0.0258 and 0.0091, respectively, which can save time and cost for scientific experiment of *P. simonii*.

4 Discussion

In this paper, we first utilize the principal component analysis to combine section length variables into five principal components and analyze the correlation among phenotypic variables, then use FCM clustering to divide the *P. simonii* into two subpopulations. In addition, For *P. simonii* seedlings in four periods, non-linear regression models are established to predict plant height, which prediction accuracy is 84.89%. Finally, we also established a model of aboveground biomass of *P. simonii* seedlings. Through our paper, we can accurately predict *P. simonii*'s plant height in four periods just based on the phenotypic data (leaf number, section length and section number), and aboveground biomass, which can effectively reveal the growth regularity of *P. simonii* seedlings and save time for scientific experiments, it also provides reference for selective breeding of *P. simonii* seedlings.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities No. 2017ZY30, No.2015ZCQ-LY-01 and the National Natural Science Foundation of China (Grant No. 11501032).

References

1. C.X. Zhang, S. Meng, Y.M. Li, L. Su, Z. Zhao. Nitrogen uptake and allocation in *Populus simonii* in different seasons supplied with isotopically labeled ammonium or nitrate. *Trees*. **30**, 6(2016):2011-2018.
2. S. Brown, AE Lugo. Biomass of tropical forests: a new estimate based on forest volumes. *Science*. **223**, 4642(1984):1290-1293.
3. T. Zhao, Z. Ouyang, L. Jia, H. Zheng, Ecosystem services and their valuation of China grassland, *Acta Ecol. Sin.* **24**, 6 (2004):1101-1110.
4. Y.H. Peng, Z.L. Huang, H.K. Hao et al., The Growth rhythm of Young Plantation of *Excentrodendron hsienmu* Cutting Seedling. *Guangxi Forestry Science*. 2016, **45**, 3(2016):249-252.
5. D. Tang, Y. Zhang. The growing regularity of *Populus simonii* in gully rolling region of Shaanxi Loess Plateau. *Journal of Northwest Forestry College*. **15**(2000).
6. Z.L. Xiang, M.X. Yan, J.Z. Rong. Study on growth law of *Populus simonii*. *Protection Forest Science & Technology*. **2**(2001):10-12.
7. M. Cao, S.K. Niu, L.Q. Li et al. Aboveground biomass models of five shrubs in Songshan Nature Reserve. *Journal of Fujian Agriculture and Forestry University*. **47**, 4(2018):436-438.
8. H. Zhou, S.W. Meng, Q.J. LIU et al., Allometric equations for estimating aboveground biomass of broad-leaved forests saplings and shrubs in subtropical China. *Journal of Nanjing Forestry University (Natural Sciences Edition)*. **41**, 6(2017):80-85.
9. N. Zeng, X.L. Ren, H.L. He et al. Estimation of aboveground biomass of grassland in the source area

of the Three Rivers based on neural network.
Research of Environmental Sciences. **30**, 1(2017):59-66.

10. G.L. Smith. Principal component analysis: An introduction. *Analytical Proceedings*. **28**(1991):150-151.
11. G. Destefanis, M.T. Barge. The use of principal component analysis (PCA) to characterize beef. *medsci*. **56** (2000):255-259.