https://doi.org/10.26525/jtfs2024.36.2.139
ISSN: 0128-1283, eISSN: 2521-9847

MULTIVARIATE ANALYSIS ON RESIN PRODUCTION FROM *PINUS ROXBURGHII* SARGENT: A CASE STUDY OF HIMACHAL PRADESH

Kumar P[1, *], Gupta RK[1], Dutt B[2], Chandel A[1], Sharma S[3] & Nand S[4]

[1]*Department of Basic Sciences, Dr. Y. S. Parmar University of Horticulture and Forestry Nauni Solan, Himachal Pradesh-173230, India*
[2]*Department of Forest Products and Utilization, Dr. Y. S. Parmar University of Horticulture and Forestry Nauni Solan, Himachal Pradesh-173230, India*
[3]*Department of Social Sciences, Dr. Y. S. Parmar University of Horticulture and Forestry Nauni Solan, Himachal Pradesh-173230, India*
[4]*Divisional Forest Officer, Himachal Pradesh Forest Department, Himachal Pradesh-173230, India*

**pawansingta@gmail.com*

Submitted January 2023; accepted July 2023

Multivariate technique is extremely helpful in assisting researchers in making sense of large, complicated and complex datasets that contain many variables measured on a large number of experimental units. Data for the present study was collected for different morphological characters, namely tree diameter, tree height, bole height, number of branches and age of tree from three major resin-producing districts of Himachal Pradesh, viz. Bilaspur, Hamirpur and Sirmour. The primary data on 360 trees of *Pinus roxburghii* were collected through well planned survey from six ranges randomly selected through a multistage random sampling technique. Two forest ranges were considered from each districts, viz. Swarghat and Bhradi forest ranges from Bilaspur district, Aghar and Hamirpur forest ranges from Hamirpur district, and Rajgarh and Narag forest ranges from Sirmour districts. In the present study, two independent methodologies, viz. discriminant analysis and principal component analysis were discussed. Although resin production may depend on a variety of characteristics, researchers are always interested in identifying those sets of attributes that most significantly affect resin production of the tree. Results from the principal component analysis revealed that based on the average of all forest ranges selected for the sampling procedure, tree diameter played a main role in resin production with an eigenvalue of 2.752, explaining 55.38% of the variation. Results from the discriminant analysis revealed that Rajgarh, Hamirpur and Swarghat forest ranges were found to be high yielders whereas Bhradi, Narag and Aghar forest ranges were found to be low yielders.

Keywords: Principal component analysis, discriminant analysis, resin, *Pinus roxburghii* Sargent

INTRODUCTION

Advanced methods for simultaneously analysing many variables' relationships are included in multivariate analysis. Multivariate analyses produce more insightful results by taking into account the interconnectedness and relative relevance of the numerous features involved (Morrison 1976). As principal components are derived in decreasing order of importance, the first few principal components that account for the majority of variation in the original data are picked and this reduces the dimensionality of the original data (Hotelling 1933). Simultaneous examination of numerous measurements of the individuals under investigation is possible with multivariate approaches (Hair et al. 1987). Advances in computer programming has led to the increased use of multivariate techniques such as discriminant analysis and principal component analysis in biological data analysis. By using an appropriate linear transformation, multivariate analysis aims to reduce the number of variables. It then selects a very small number of the resulting linear combinations in an optimal way, ignoring the remaining linear combinations in the hope that they do not contain important information. As a result, the dimensionality of the problem is decreased. These techniques are now employed in a wide variety of domains, particularly where formal probabilistic models cannot be used and the variables pertaining to the research projects are intended to be correlated with one another.

One of the oldest renewable forest products is resin. Chir pine produces high-quality oleo resin that when steam-distilled, yields the industrially significant byproducts of rosin and turpentine oil. Turpentine oil has a wide range of industrial applications, including the perfumery industry, production of synthetic pine oil, medicinal preparations, denaturants and disinfectants. Keeping in mind the economic significance of resin industry, the present study was conducted using principal component analysis and discriminant analysis to evaluate the relative contribution of morphological features on resin production in *Pinus roxburghii* Sargent.

MATERIALS AND METHODS

Multistage random sampling technique was employed for the selection of Pinus forests. In the first stage, three districts (Bilaspur, Hamirpur and Sirmour districts) were selected randomly. From each selected districts, two forest ranges were selected randomly, i.e. Swarghat and Bhradi forest ranges from Bilaspur district, Aghar and Hamirpur forest ranges from Hamirpur district, and Rajgarh and Narag forest ranges from Sirmour district (Figures 1 and 2). Resin-producing vegetation in each forest range was divided into different diameter classes i.e. 30–40 cm, 40–50 cm, 50–60 cm, 60–70 cm, 70–80 cm and 80–90 cm by using the statistical technique. In each forest range, three quadrats of 0.1 hectare were made in the Pinus forest and ten trees were selected randomly from each diameter class in those quadrats. A total of 60 trees were selected from each forest range and likewise the sampling procedure was repeated in rest of the forest ranges. Therefore, the overall sample size was 360 trees from six forest ranges (Hamirpur and Aghar forest ranges from Hamirpur district, Swarghat and Bhradi forest ranges from Bilaspur district, Rajgarh and Narag forest ranges from Sirmour district).

Data collected on different morphological characters to assess the contribution of each character towards the production of resin were: tree height, tree diameter, bole height, number of branches, and age of trees. Multivariate techniques, viz. discriminant analysis and principal component analysis were used for the assessment of relative contribution of morphological characters contributing significantly towards resin yield.

Discriminant analysis

Fisher (1936) introduced the discriminant function analysis. With this technique, a number of measurements were used to provide a discriminant function (linear) that was better than any other linear functions for the observation. By selecting a linear component of the original variables and constructing statistics appropriate for the univariate case, the problem is then reduced to that of a single variable. By making a suitable choice of coefficients, the maximised value of the statistics obtained is used as the test condition. In these situations, discriminant analysis frequently yields results that are more gratifying than those of regression or correlation analysis.

Let d_1, d_2,...,d_p are the 'p' normal variate with same dispersion matrix (α_{ij}) but distributed independently of W_{ij}, where W_{ij} (i j =1,2,........,p) is the matrix giving the estimates on n degrees of freedom of the elements in the dispersion matrix (α_{ij}) of p-normally correlated variables. Considering only the first r variables, d_1, d_2, ...,d_r the statistic T_r is defined by:

$$nT_r = \sum_{i=1}^{p} \sum_{j=1}^{p} W_r^{ij}\, d_i d_j$$

where W_r^{ij} is the matrix reciprocal to W_{ij} (i, j =1,2,...,r) such that, $\frac{n-r+1}{r} T_r \sim F\,(r, n\text{-}r+1)$

It can be easily shown that if d_{r+1},...,d_p are distributed independently of d_1, d_2,...d_r and $E(d_{r+i})$ = ... $E(d_p)$ =0, $E(d_i)$ being not necessarily zero when i = 1,2...r, the statistic is distributed as:

$$\frac{n-p+1}{p-r} U_{p\text{-}r,\, r} \sim F\,(p\text{-}r, n\text{-}p+1)$$

The average value of W for classes I and II having number of cases a and b, respectively, was given by:

$$\bar{W}_I^a = \lambda_1 \bar{X}_1^a + \lambda_2 \bar{X}_2^a + \ldots + \lambda_p \bar{X}_p^a$$

$$\vdots \qquad \qquad \qquad \qquad \vdots$$

$$\bar{W}_{II}^a = \lambda_1 \bar{X}_1^b + \lambda_2 \bar{X}_2^b + \ldots + \lambda_p \bar{X}_p^b$$

The cut-off point choosing between classes I and II lies between $\bar{W}_I^a$ and $\bar{W}_{II}^a$. Its exact value could be dependent on the relative cost of miss-classifying the units, but frequently it is taken as the midpoint between W_I and W_{II}. A test of the hypothesis that the discriminant function has no discrimination provided by the F-test, is constructed as follows:

$$F = \left[\frac{\frac{n_I n_{II}/(n_I+n_{II})D^2}{p}}{\frac{D}{n-p-1}}\right] \text{with p and n-p-1 df}$$

where $D = \bar{W}_I^a$ and $\bar{W}_{II}^a = \sum \lambda_j d_j$ and $n = n_I + n_{II}$

The real adequacy of the discriminant function, however, must be determined by how well it discriminates between classes I and II on a fresh sample of data. The six forest ranges i.e. Hamirpur and Aghar forest ranges from Hamirpur district, Swarghat and Bhradi forest ranges from Bilaspur district, and Rajgarh and Narag forest ranges from Sirmour district were divided into high and low yielder groups in terms of resin production.

Principal component analysis

Data containing a large number of associated variables is reduced using this multivariate statistical technique to a much smaller set of new variables. Through a few linear combinations of these variables, the principal component analysis attempts to describe the variance-covariance structure of a set of variables. Its general objectives are data reduction and interpretation.

The presumption of multivariate normality is not necessary. The analysis deals with the internal organisation of relevant variables. In order to maintain the fewest number of variables possible, it aims to sacrifice part of the information found in the original variables but the amount of information lost is kept as minimal as possible. Biology, economics, meteorology, anthropology, social sciences and agriculture are some of the areas where the technique of principal component analysis is widely used. It has been used in conjunction with discriminant analysis for improving the stability of the coefficients, multiple regression analysis to tackle the problem of multicollinearity.

Finding principal components—new variables with the majority of information included in the original variables—is the goal of principal component analysis. In most cases, the correlation matrix (R) or sample variance-covariance matrix (S) are used to estimate the principal components. Scale impact can alter the makeup of the derived components when the variables are measured in different units. It is preferable to standardise the variable in order to address this issue. Correlation matrix should also be applied.

When the variation described by subsequent principal components is relatively small and the first few principal components typically account for majority of variation in the original variables, it is often useful to keep only those initial principal components and discard rest of the components from the analysis. The reason for this is that the variable they convey is completely random and of no use in the analysis. For the number of primary components of the correlation matrix with eigen roots less than one, a number of rule-of-thumb have been provided. Less information is contained in the primary components with variances below one.

Various steps involved in working out the principal components can be summarised as below:

i) Firstly, the Kaiser-Meyer-Olkin (KMO) measure for sampling adequacy is computed. If the KMO value is more than 0.5, only then we should go for principal component analysis.

ii) After that, determine the eigen value of variance-covariance matrix or correlation matrix.

iii) Arrange eigen values in decreasing order. Let these values in decreasing order be $\lambda_1, \lambda_2, \ldots, \lambda_p$ and corresponding variability be $V_1, V_2 \ldots V_p$, where V_p is variability for $\lambda_{p.}$

iv) Starting from the first principal component, go on adding the variance of first few principal components whose value is more than unity. The variability described by them is of greater use. Discard the remaining principal components.

v) From the eigen vectors of chosen

principal components, variables which load the respective principal components are determined.

The output desired for interpretation and grouping should include:

i) Eigen value and percentage of total variation explained by each principal component.

ii) The eigen vector for each principal component.

iii) The principal component scores.

iv) The correlation between original standardised variable and the corresponding principal component scores (occasionally called loading).

Principal component analysis technique was used to identify the important characteristics contributing towards resin production in *P. roxburghii*.

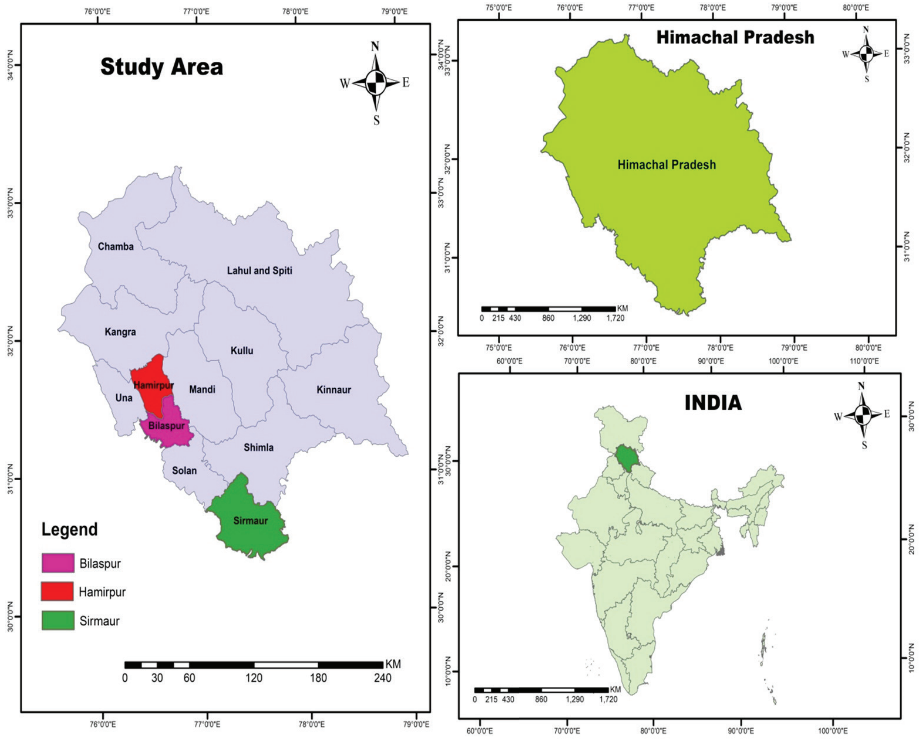

Figure 1 Study area (QGIS Software)

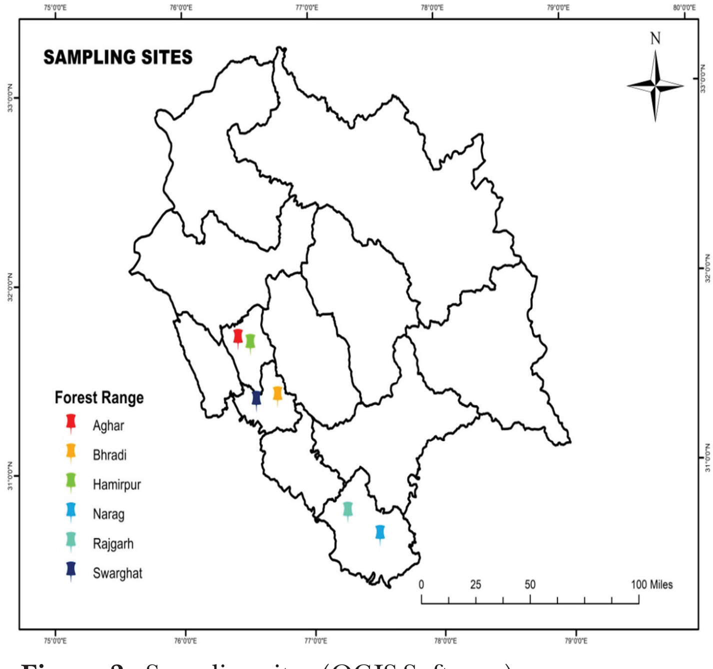

Figure 2 Sampling sites (QGIS Software)

RESULTS AND DISCUSSION

Discriminant analysis

The approach of categorising high and low yield resin on the basis of randomly selected characteristics is statistically weak. The discriminant analysis is a systematic and statistically valid procedure for this purpose. In the present study, the observation of different morphological characters of 360 trees were divided into two groups namely 'high yielder' and 'low yielder' on the basis of average value of resin yield and discriminant function was fitted. The discriminant function was found to be:

$$D = -6.571 + 0.109\ X_1$$

where D stands for resin yield, X_1 for tree diameter.

Thus, this equation reveals that tree diameter is the most important character that discriminate the two groups. The value of Wilk's lambda (λ) was obtained as 0.297 and which in turn, gave the computed value of chi square (χ^2) as 434.39. The forest ranges were assigned to group 1 (high yielder) if $D \geq m$ and if otherwise they were assigned to group 2 (low yielder), where $m = -0.068$ is the average of group centroids. Groups formed on the basis of the allocation rule are given in Table 1. The results of discriminant analysis revealed that Rajgarh, Hamirpur and Swarghat forest ranges were found to be high yielders, whereas Bhradi, Narag and Aghar forest ranges were found to be low yielders as shown in Table 1.

Table 1 High and low yielder forest ranges

High yielder forest ranges	Low yielder forest ranges
Rajgarh	Bhradi
Hamirpur	Narag
Swarghat	Aghar

Table 2 Eigen values and principal components for Rajgarh forest range (bold values indicate variables in linear combination)

Variables	PC_1	PC_2
Tree diameter	**0.694**	-0.016
Tree height	**0.673**	0.102
Bole height	-0.047	**0.981**
Number of branches	**0.741**	-0.012
Age of tree	**0.697**	-0.015
Eigen value	3.280	1.005
Cum. % of variance	65.591	85.698

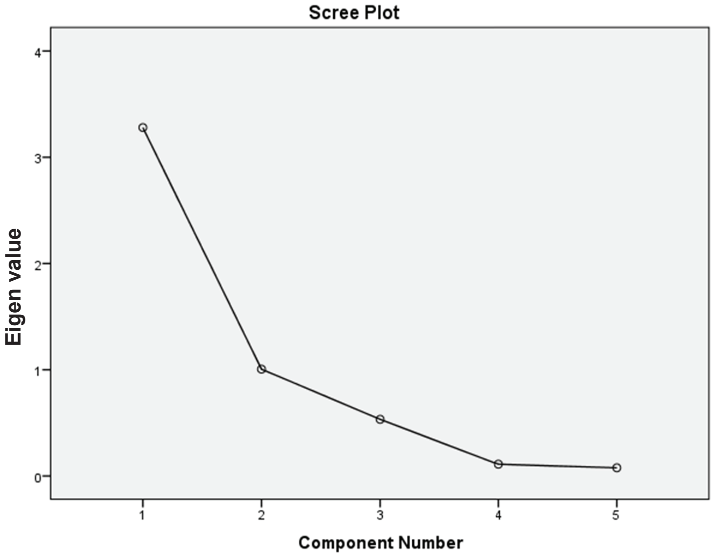

Figure 3 Scree plot for Rajgarh forest range

Table 3 Eigen values and principal components for Narag forest range (bold value indicate variable in linear combination)

Variables	PC_1
Tree diameter	**0.819**
Tree height	0.248
Bole height	0.205
Number of branches	0.180
Age of tree	0.324
Eigen value	2.915
Cum. % of variance	58.301

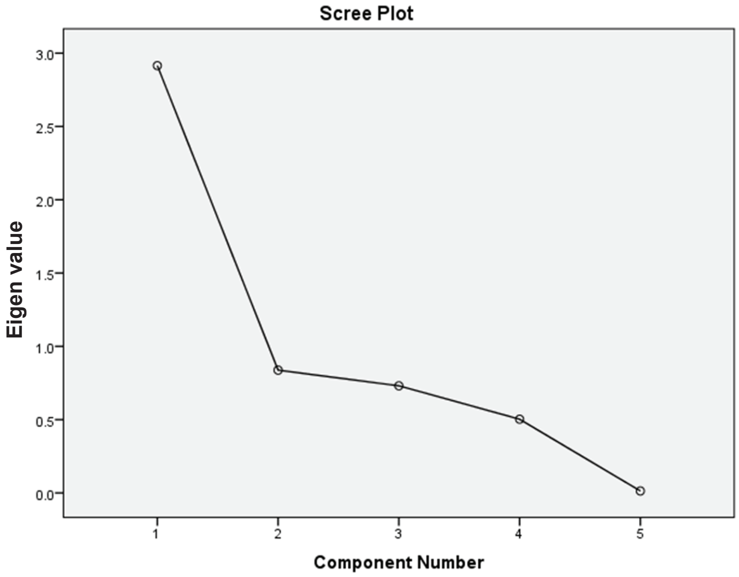

Figure 4 Scree plot for Narag forest range

Principal component analysis

To interpret the data in a more meaningful form, it is necessary to reduce the number of variables to a few interpretable linear combinations of variables. Thus, principal component analysis was employed to reduce the observed variables into a number of principal components that will account for most of the variation in observed variables.

Table 2 and Figure 3 revealed that two of five principal components (PCs) for the Rajgarh forest range had eigen values greater than unity and therefore these principal components play main role in the analysis. The first two principal components had been retained in the analysis, which explained 85.698% of total variation. The first principal component showed eigen value of 3.280 and second principal component had eigen value of 1.005. The first principal component was linear combination of tree diameter, tree height, number of branches and age of tree. The second principal component comprised of bole height only.

Table 3 and Figure 4 revealed that one of five principal components (PCs) for the Narag forest range had eigen values greater than unity and therefore this principal component plays a main role in the analysis. The first principal component had been retained in the analysis, which explained 58.301% of variation. The first principal component which showed eigen value of 2.915, was linear combination of tree diameter only.

Table 4 and Figure 5 revealed that two of five principal components (PCs) for the Swarghat forest range had eigen values greater than unity and therefore these principal components

Table 4 Eigen values and principal components for Swarghat forest range (bold values indicate variables in linear combination)

Variables	PC_1	PC_2
Tree diameter	**0.950**	0.101
Tree height	**0.832**	0.083
Bole height	**0.698**	-0.323
Number of branches	0.056	**0.963**
Age of tree	**0.938**	0.090
Eigen value	2.970	1.054
Cum.% of variance	59.40	80.48

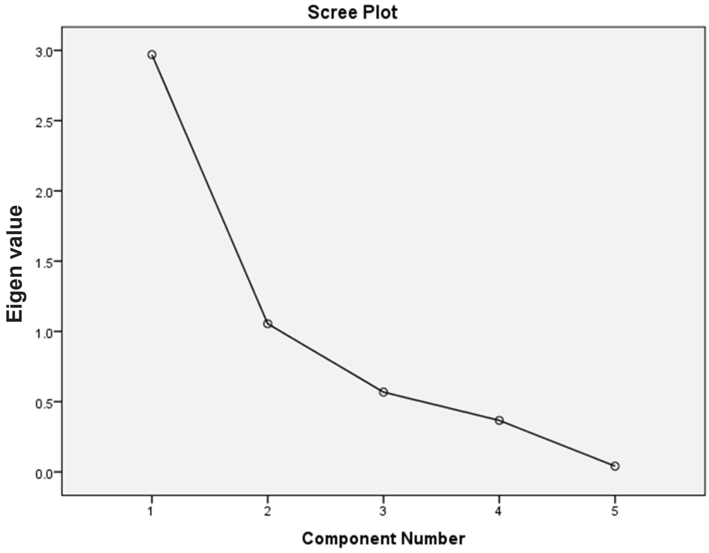

Figure 5 Scree plot for Swarghat forest range

play main role in the analysis. The first two principal components had been retained in the analysis, which explained 80.48% of total variation. The first principal component had eigen value of 2.970 and second principal component had eigen value of 1.054. The first principal component was linear combination of tree diameter, tree height, bole height and age of tree. While the second principal component comprised of number of branches only.

Table 5 and Figure 6 revealed that one of five principal components (PCs) for the Bhradi forest range had eigen values greater than unity and therefore this principal component plays a main role in the analysis. The first principal component had been retained in the analysis, which explained 72.89% of variation. The first principal component had eigen value of 3.644 and was linear combination of the tree diameter only.

Table 6 and Figure 7 revealed that two of five principal components (PCs) for the Hamirpur forest range had eigen values greater than unity and therefore these principal components play main role in the analysis. The first two principal components had been retained in the analysis, which explained 69.223% of total variation. The first principal component had eigen value of 3.757 and second principal component had eigen value of 1.089. Results showed that the first principal component was linear combination of tree diameter, tree height, bole height and age of tree. Whereas the second principal component comprised of number of branches only.

Table 7 and Figure 8 revealed that one of five principal components (PCs) for the Aghar forest range had eigen values greater than unity and therefore this principal component plays a main role in the analysis. The first principal

Table 5 Eigen values and principal components for Bhradi range (bold value indicate variable in linear combination)

Variables	PC_1
Tree diameter	**0.861**
Tree height	0.232
Bole height	0.193
Number of branches	0.218
Age of tree	0.261
Eigen value	3.644
Cum.% of variance	72.89

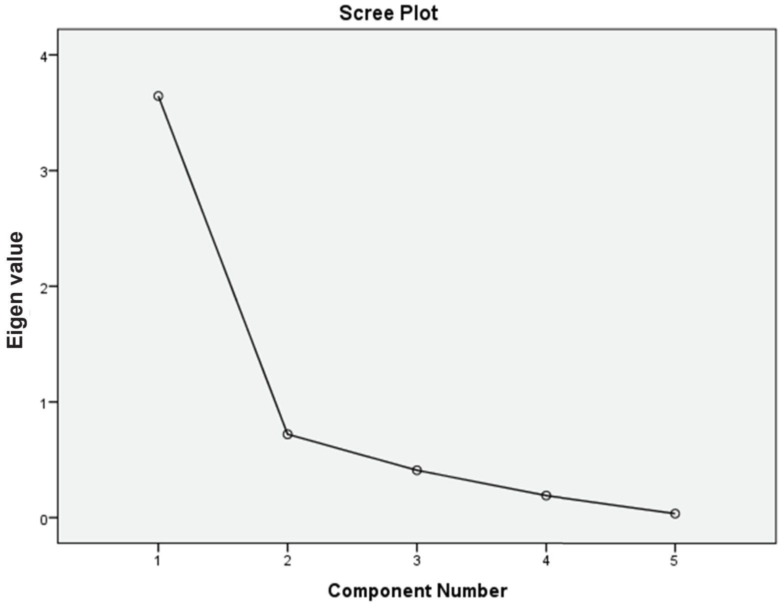

Figure 6 Scree plot for Bhradi forest range

component had been retained in the analysis, which explained 66.25% of variation. The first principal component had eigen value of 3.312 and was linear combination of tree diameter, tree height, bole height and age of tree.

Table 8 and Figure 9 which showed results for all the six forest ranges revealed that one of five principal components (PCs) had eigen values greater than unity, and therefore this principal component plays a main role in the analysis. The first principal component had been retained in the analysis, which explained 55.38% of variation. The first principal component had eigen value of 2.752 and was linear combination of tree diameter, tree height, bole height and age of tree.

The results of principal component analysis revealed that on an average among all the forest ranges selected for the sampling procedure, tree diameter plays an important role in resin yield which was in close affinity with previous findings on the usefulness of multivariate analysis in forestry research which used *P. roxburghii* as a case study (Sood et al. 2018).

CONCLUSION

The results of principal component analysis revealed that one of seven principal components (PCs) had eigen values greater than unity and therefore this principal component plays an important role in the analysis. The first principal component had been retained in the analysis, which explained 55.38% of variation and showed eigen value of 2.752. Thus, the principal component analysis has brought

Table 6 Eigen values and principal components for Hamirpur range (bold values indicate variables in linear combination)

Variables	PC_1	PC_2
Tree diameter	**0.958**	0.014
Tree height	**0.775**	0.043
Bole height	**0.671**	-0.274
Number of branches	0.062	**0.972**
Age of tree	**0.907**	-0.127
Eigen value	3.757	1.089
Cum.% of variance	53.671	69.223

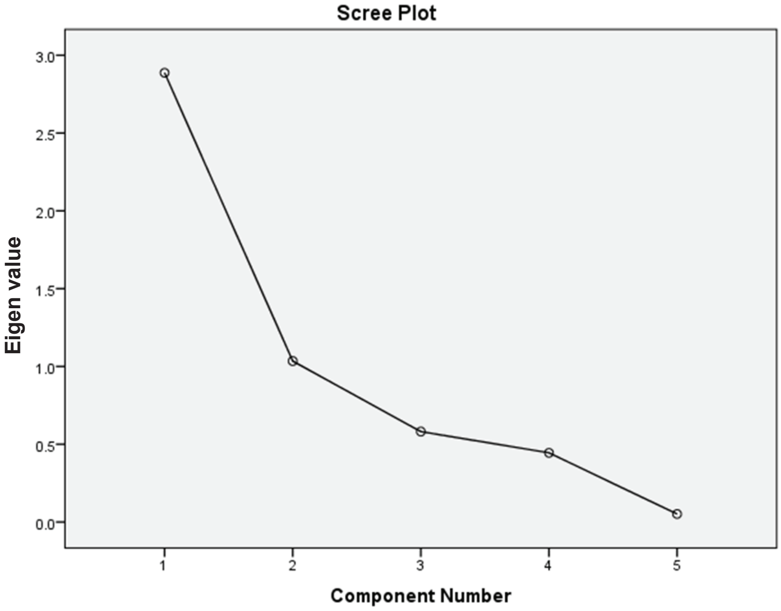

Figure 7 Scree plot for Hamirpur forest range

Table 7 Eigen values and principal components for Aghar range (bold value indicate variable in linear combination)

Variables	PC_1
Tree diameter	**0.949**
Tree height	**0.862**
Bole height	**0.747**
Number of branches	0.142
Age of tree	**0.942**
Eigen value	3.312
Cum.% of variance	66.25

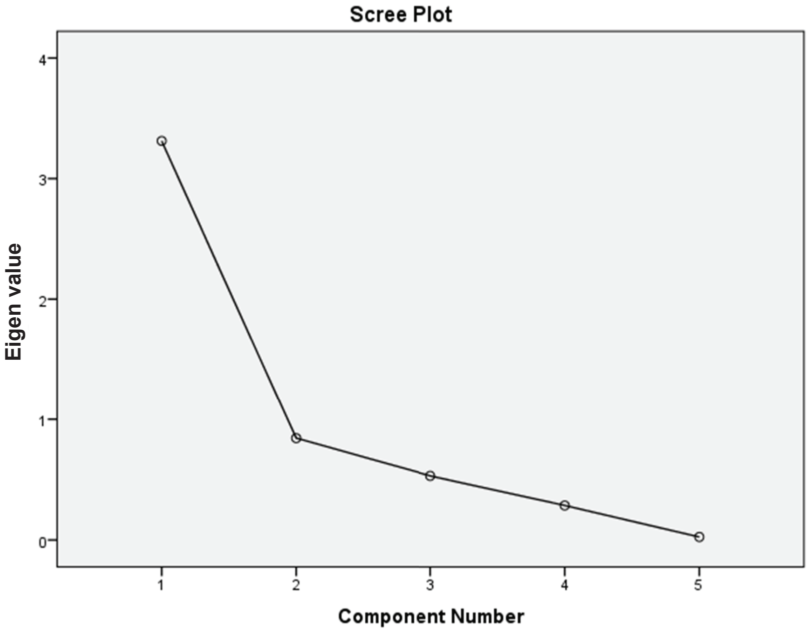

Figure 8 Scree plot for Aghar forest range

Table 8 Eigen values and principal components for all the six forest ranges together (bold values indicate variables in linear combination)

Variables	PC_1
Tree diameter	**0.929**
Tree height	**0.749**
Bole height	**0.624**
Number of branches	0.135
Age of tree	**0.894**
Eigen value	2.752
Cum.% of variance	55.38

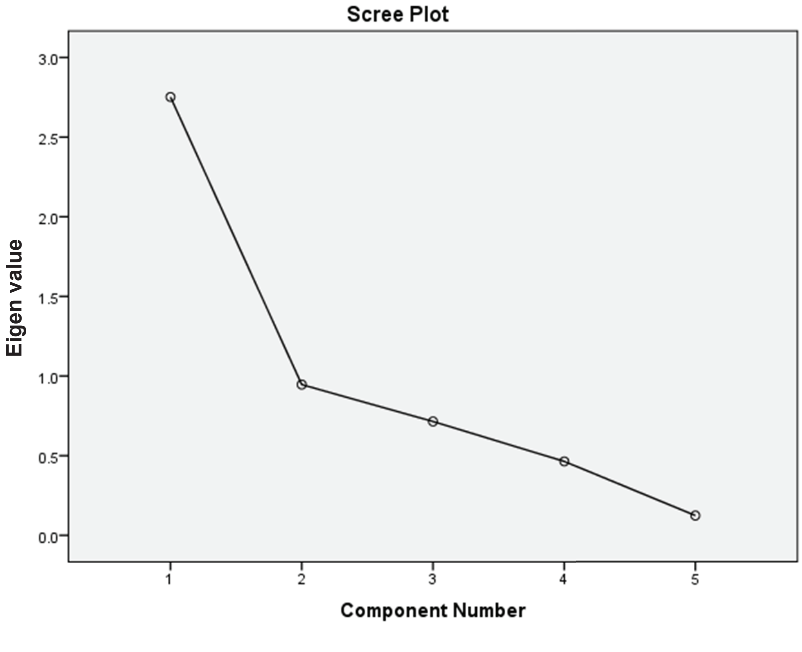

Figure 9 Scree plot for all the six forest ranges together

out some of the basic components associated with morphological characters of *P. roxburghii* and could be considered as important tool in explanatory work for optimising resin productivity. The results of discriminant analysis revealed that Rajgarh, Hamirpur and Swarghat forest ranges were found to be high yielders, whereas Bhradi, Narag and Aghar forest ranges were found to be low yielders of resin.

REFERENCES

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–188.

Hair JR, Anderson RE & Tathan RL. 1987. *Multivariate Data Analysis with Readings*. McMillan, New York.

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(6): 417–418.

Morrison DF. 1976. *Multivariate Statistical Methods*. McGraw-Hill Company, New York.

Sood Y, Mahajan PK, Bharti & Sharma KR. 2018. Usefulness of multivariate analysis in forestry research: a case study of *Pinus roxburghii*. *Journal of Pharmacognosy and Phytochemistry* 7(2): 3508–3509.