Classification and Risk Identification of Forest Ecology Based on Statistical Learning

Xuan Sun, Ting Lei *
Department of mathematics, College of Science, Beijing Forestry University, Beijing, China

*Corresponding author e-mail: Park_0662@126.com

Abstract. In this paper, the forest ecosystem classification and risk identification are studied based on statistical learning methods such as cluster analysis and discriminant analysis, taking Daxinganling Forest in Inner Mongolia, Beijing Forest, Jilin Changbai Mountain Forest, Hubei Huitong Forest, Guangdong Heshan Forest, and Linzhi Forest as the research objects. The results show that: (1) The classification of forest ecosystem activity is mainly determined by the forest species. At the same time, the composition and growth of forest species is an important basis for risk classification. (2) Climate factors and soil factors affect forest ecosystem stability and vitality, which is an important basis for testing models after risk identification. Therefore, according to the results obtained in this article, the key to study forest ecology is to plant the most suitable forest species and tree species on various site types and implement corresponding afforestation measures so that the entire region can achieve "suitable trees" and "rational management ", so that the potential of land production can be brought into full play and the goal of " use the land as much as possible " is achieved.

1. Introduction
Forest is the main body of terrestrial ecosystem, which is an important natural resource for human survival. The health status of forests will be directly related to global ecological security and the sustainable development of human society, so it is of great social significance to carry out forest health analysis and evaluation.

There is sufficient information about forest health, such as formulas, models. The specific ones include the domestic "health distance ", foreign "VOR model", "forest ecosystem health index formula", etc. The model uses BP neural network model, regression analysis, health index and other methods to carry out health evaluation and early warning of the forest through various indicators in the forest by using different weights of evaluation indicators. However, in the current research results, the evaluation system is incomplete. Due to the different academic directions of forest research, there is no Unified standard for the scale of forest evaluation [1].

In this paper, we study the forest based on statistical learning methods such as cluster analysis and principal component analysis, this paper studies the classification and risk identification of forest ecological environment.
2. Object and Method

2.1. Content Range
The five forests selected in this article are distributed in different latitudes and longitudes in China which have distinctive climatic characteristics. At the same time, the forests studied have different forest Category.

The geographical location of Huitong China Fir Ecological Station is 109 ° 45 ′ east longitude and 26 ° 50 ′ north latitude. At the same latitude in the world, except for the distribution of forest vegetation in China, the rest are basically deserts and semi-deserts, which is of great research significance.

2.2. Data Collection Method and Data Analysis
The data are from China National Ecosystem Research Station. Take Huitong fir forest as an example, attribute indexes were collected from the forest ecosystem detection indicators of Huitong fir forest from 2005 to 2010. The indicators include forest climate, temperature, soil moisture content, microbial content, and other 32 organic and inorganic factors.

3. Results and Analysis
Forest biomass represents the primary productivity of forests, and plays a great role in the research of forest ecosystems and other aspects. So, this paper chooses biomass carbon as the criteria for stability and vitality of forest ecosystems.[1]

3.1. Cluster Analysis to Classify Forest Ecosystems
According to the data collected from the five major forests, different types of forests are regarded as different types of sample. We use Q-type cluster analysis to classify five major forests which have twelve forest types in based on biomass carbon to obtain different levels of forest ecosystems.

In system cluster, the distance between samples (discrepancy coefficient) can be calculated by the same combined model (between groups linkage method), that is, the average distance of all samples in two small classes is the distance between the two small classes. Since the sample data is a continuous sample, the distance between samples uses euclidean distance squared:

\[ SEUCLID = \sum_{i=1}^{k} (x_i - y_i)^2 \]

According to the classification of biomass carbon(forest activity), we divided the twelve types forests into three categories, namely: the first type [Beijing Forest Pine Forest]; the second type: [Guangzhou Dinghu Mountain Monsoon Forest, Beijing Comprehensive Forest Observation Site, Zhejiang Tiantong Forest Schima superba, Beijing Forest Deciduous Forest, Zhejiang Tiantong Forest Sassafras Forest, Beijing Forest Deciduous Forest]; Category III: [Guangzhou Dingshan Forest Needle Broad II, Zhejiang Tiantong Forest Evergreen Irrigation Forest , Guangzhou Dinghushan horsetail pine forest, Zhejiang Tiantong Forest horsetail pine forest, Changbai Mountain forest red pine forest]

From the figure 1 we can draw the following conclusions:(1). Different types of forests in the same forest (same latitude and longitude) are at different forest levels. Basically, the same type of forest is in the same forest class. This is the same as "forest site evaluation " in forest ecology.

(2). By comparing the original data, it is found that the first type of forest biomass carbon has the highest carbon content, indicating high soil fertility and high forest vigor; the second type of forest biomass carbon content is moderate; the third type of forest biomass carbon content is the least. This is related to the biomass carbon content in forest ecology as evergreen broad – leaved forest > masson pine forest > bamboo forest > chinese fir plantation [2], which consistent with the actual situation.

Therefore, the analysis of forest activity from a mathematical perspective is consistent with forest ecology research and practice.
3.2. Find the Main Site Factor

The growth indexes of certain tree species on forest land are usually used to measure and evaluate the site quality of forests. Through the forest site evaluation, we can judge which tree species is dominant under different types of forest ecological conditions.

Because the factors that affect the activity and stability of forest ecosystems are very complex, we use the principal component analysis method to eliminate the relevant impacts between the evaluation indicators and reduce the workload of index.

First use communality to score the principal components. According to the figure, you can calculate the variance contribution of principal component to $X_{ij}$. It can be seen from the figure 2 and figure 3 that the variance contribution of the fifteen factors are all greater than 85%, which indicates that the extracted common factors are highly representative of the primitive variables, and the overall effect is good.

Because the dimensions of the variables in this example are relatively complicated, in order to keep the differences of the variables themselves, we choose to solve the principal components from the correlation matrix. After taking the four principal components, the cumulative variance contribution ratio reaches 97.2%, which contains most of the information.

Figure 4 gives the vector coefficients that can be used to calculate the linear expression of the principal components with respect to the original

| Communalties     | Initial | Extraction |
|------------------|---------|------------|
| Total solar radiation | 1.000   | .993       |
| Air pressure     | 1.000   | .997       |
| Air humidity     | 1.000   | .999       |
| Air temperature  | 1.000   | .971       |
| Cu               | 1.000   | .994       |
| Pb               | 1.000   | .982       |
| Zn               | 1.000   | .981       |
| Mn               | 1.000   | .992       |
| Co               | 1.000   | .910       |
| Cd               | 1.000   | .968       |
| Pb               | 1.000   | .988       |
| Ni               | 1.000   | .958       |
| pH               | 1.000   | .997       |
| Organic matter   | 1.000   | .999       |
| Biomass carbon   | 1.000   | .874       |

| Component | Total Variance Explained | Total Variance Explained | Total Variance Explained | Total Variance Explained |
|-----------|--------------------------|--------------------------|--------------------------|--------------------------|
| $\lambda_1$ | $\rho(Y_i, X_j) = \gamma_1 \sqrt{\lambda_1}$, where $\lambda_1$ is the first eigenvalue of the covariance matrix, $\gamma_1$ is the eigenvector. The column vector corresponding to the rescaling parameters is the factor load of the main component. So, the correlation coefficients $\rho(Y_i, X_j) =$
In order to write a linear expression of the first principal component with respect to the original variable, the column corresponding to the original data needs to be divided by $\sqrt{\lambda_1} = \sqrt{7.427}$ to obtain the transformation vector coefficient (eigenvector), and then, the expression of the first principal component is:

$$Y_1 = 0.223X_1 + 0.221X_2 + 0.078X_3 + 0.282X_4 + 0.331X_5 + 0.189X_6 + 0.356X_7 + 0.313X_8 + 0.284X_9 + 0.342X_{10} + 0.333X_{11} + 0.321X_{12} + 0.179X_{13} + 0.022X_{14} + 0.015X_{15}$$

Similarly, the four main components that affect the vitality and stability of forest ecosystems are:

1. $Y_1 = -0.223X_3 - 0.221X_2 - 0.078X_3 + 0.282X_4 - 0.331X_5 + 0.189X_6 + 0.356X_7 - 0.313X_8 - 0.284X_9 + 0.342X_{10} + 0.333X_{11} + 0.321X_{12} + 0.179X_{13} + 0.022X_{14} - 0.015X_{15}$
2. $Y_2 = 0.388X_1 - 0.398X_2 + 0.437X_3 - 0.335X_4 + 0.02X_5 + 0.23X_6 + 0.033X_7 - 0.126X_8 - 0.058X_9 - 0.074X_{10} + 0.002X_{11} - 0.086X_{12} + 0.428X_{13} + 0.135X_{14} + 0.315X_{15}$
3. $Y_3 = -0.0156X_1 - 0.089X_2 - 0.352X_3 + 0.078X_4 - 0.197X_5 + 0.449X_6 - 0.013X_7 - 0.014X_8 + 0.024X_9 - 0.081X_{10} - 0.205X_{11} - 0.248X_{12} + 0.196X_{13} + 0.595X_{14} - 0.308X_{15}$
4. $Y_4 = 0.239X_1 + 0.264X_2 + 0.171X_3 + 0.098X_4 - 0.228X_5 - 0.178X_6 + 0.150X_7 + 0.374X_8 - 0.441X_9 + 0.205X_{10} + 0.195X_{11} - 0.102X_{12} + 0.239X_{13} + 0.189X_{14} - 0.46X_{15}$

The size of each factor coefficient in the principal component expression can reflect the contribution of the factor to the principal component. In principal component analysis, it is generally considered that a load amount greater than 0.3 is significant. We can analyze what the four principal components are from the above four principal component expressions. The coefficient of $X_2$ (0.331), $X_3$ (0.356), $X_4$ (0.313), $X_5$ (0.342), $X_6$ (0.333), $X_7$ (0.321) in the first principal component is greater than the other factors. Those are the various constant elements in the soil, so we can call the first principal component the soil constant element factor; The coefficient of $X_1$ (0.388), $X_2$ (0.398), $X_3$ (0.437), $X_4$ (0.335), $X_5$ (0.428), $X_6$ (0.315) in the second principal component is greater than the other factors. those are the total solar radiation, air pressure, air temperature, and air humidity, so we can call the second principal component the climate factor; In the same way, we can call the third main component the soil organic factor , and the fourth principal component the soil trace element factor;

It can be seen from the principal component analysis that, in fact, the four principal components can be simply summarized into two main aspects, namely climate and soil.

1. The main factors that affect forest stability are the various elements in the soil. The effects of major elements, trace element, and organic compounds on plant growth are projected in different impact vectors, that is, the effects of the three components on plants are different. Each kind of forest has suitable soil conditions, which also corresponds to the fact that the parent rock under the soil determines the distribution of plant types. Therefore, it is very important to plant suitable forest species according to the soil content in different regions.

2. Climate is a combination of various climatic factors, including solar radiation, air temperature, air humidity, etc., which change from time to time and from day to day, resulting in complex phenomena. [3] Which affect the air temperature, humidity, solar radiation and precipitation, terrain, latitude and longitude and so on. At the same time, the growth of the forest requires years and months of accumulation. In the process of growing woody plants, they must adapt to the average temperature of a region and undergo weather changes from day to day, month to year. Combined with the clustering in the previous step, we can see that climate factors affect tree growth, and there is a general relationship between tree distribution and climate.

In Spar's <Forest Ecology>, which is also proposed that "the forest habitat commonly referred to when talking about trees is the sum of the atmospheric and soil conditions that surround the plants and can be used by the plants."

### 3.3. Fisher Discriminant Analysis

Based on the forest ecosystem classification obtained by cluster analysis and the four major site factors of principal component analysis as the criterion, the twelve forest species in the five major forests were classified to determine the correct rate.

First analyze the data set:
It can be seen from the figure 5 that the average values of the three indicators $x_1, x_2, x_3$ are different in the three types of pieces. This test shows that the discrimination is meaningful. The significance is less than 0.05, which is statistically significant.

Taking temperature and soil constant elements as examples, the correlation coefficient is -0.398, and the significance $p = 0.005 < 0.01$ is statistically significant. The correlation coefficient between major elements and trace elements in soil was -0.261, and the significance was $p = 0 < 0.01$.

The following fisher test is performed: three variables of 50 forests in the sample were analyzed by fisher discriminant analysis using SPSS. The results (figure 7) show that the correct rate of determining the forest level is 68%, indicating that the site factor of the principal component analysis can summarize the characteristics of the forest, and then classify the forest level.

Figure 8 shows the demarcation map of various types of forest samples. “1” stands for “most unstable forest ecosystem”; “2” stands for “more stable forest ecosystem”; “3” stands for “most stable ecosystem”, “+” represents the center of various types of projections. The “32” dividing line indicates the boundaries between the three types and the second type after projection. The remaining two meanings are similar.
3.4. Discriminant Analysis for Further Verification

Model accuracy can be tested from multiple correlation and partial correlation coefficient. If the t value of the partial correlation coefficient of a certain site factor is greater than $t_{0.01}$, it indicates that the factor has a certain effect. Greater than $t_{0.05}$ can be regarded as an important factor, and greater than $t_{0.01}$ can be regarded as the dominant factor.

According to figure 9 and figure 10, because the biomass carbon content determines the forest level, partial correlation analysis is performed with biomass carbon as the control variable and main impact factors (soil major elements, air temperature) as variables. Through analysis, the associated probability of each pair of variables is less than 1, which has strong significance. That is closely related to the stability of forest ecosystems. The model is established.

The coefficient matrix of the bayesian function is expressed as a mathematical expression:

$$
g_1(x) = 0.029x_1 - 3.225x_2$$
$$g_2(x) = 0.002x_1 - 1.106x_2$$
$$g_3(x) = 0.025x_1 - 2.754x_2$$

When making a judgment, you only need to bring them into these three classification functions and classify them into the class corresponding to the function with the largest value.

### Prior Probabilities for Groups

| Forest level | Prior | Cases Used in Analysis |
|--------------|-------|------------------------|
|              |       | Unweighted | Weighted |
| 1            | .333  | 5          | 5.000    |
| 2            | .333  | 26         | 26.000   |
| 3            | .333  | 18         | 18.000   |
| Total        | 1.000 | 49         | 48.000   |

### Classification Function Coefficients

|                      | Forest level | 1 | 2 | 3 | 4 |
|----------------------|--------------|---|---|---|---|
| air temperature      | .212         | .137| .158|
| Soil macroelements   | .001         | .013| .041|
| Exchangeable calciums| -.012        | -.003| -.008|
| Soil trace elements  | -.601        | -.234| -.140|

Fisher’s linear discriminant functions

**Figure 7.** Classification results.

**Figure 8.** Pedigree chart.

**Figure 9.** Prior probability

**Figure 10.** Classification function coefficient
4. The Advantages and Disadvantages of the Model

4.1. Comparison of existing forest discrimination with the model in this paper

Existing forest assessment models take quantitative theory as the theoretical premise. Variables that cannot be expressed numerically are mathematically transformed into quantities that can be expressed numerically. Qualitative variables are called items, and each of these different "values" are called categories.

After 0 ~ 1 digital treatment, the problem is quantified.

Although the results in this paper are good, compared with this model, some quantitative factors (gradient) are not considered. Quantitative factors can be added later for further research.

4.2. Realization of the model in this paper

Forest health is divided into five levels, which are high quality, healthy, sub-health, unhealthy and disease. The corresponding police levels are no warning, light warning, moderate warning, severe warning and giant warning. Among them, high quality corresponds to no warning, disease corresponds to giant police.

It can be seen from this that, although mentioned the first part of the article, the division of forests is mainly based on different forest species, and the classification of different forest species in the same forest is different. However, if we take the characteristic forest species in the forest (that is, the main planted forests in this region), we can see that due to abundant water and suitable climate in southern China, the forest ecosystems in the southern region are basically "high quality". Northern China is not suitable for plant growth due to the cold climate. Although the soil conditions are excellent, it cannot provide sufficient water and energy. Therefore, the northern forests are basically "healthy" and "sub-health".

Table 1. Forest health alert classification

| Health level | Forest Health Brief | Alertness | Signal |
|--------------|---------------------|-----------|--------|
| high quality | Very high species diversity, larger vegetation coverage, stronger resistance to pests and diseases, low forest fire risk level, no alarm | no warning | Green alert |
| healthy      | High species diversity, large vegetation coverage, strong resistance to pests and diseases, low forest fire risk level, no alarm | light warning | Blue alert |
| sub-health   | Species diversity is low, vegetation cover is small, ability to resist pests and diseases is weak, forest fire danger level is low, alert situation is in the gestation stage | moderate warning | Yellow alert |
| unhealthy    | Low species diversity, small vegetation coverage, weak resistance to pests and diseases, high forest fire risk level, and alert situation is in the development stage | severe warning | Orange alert |
| disease      | Species diversity is low, vegetation cover is small, resistance to pests and diseases is very low, forest fire danger level is high, and alert situation is in the outbreak stage | giant warning | Red alert |

Among the characteristic forests in China, two forests are quite special, namely "Zhejiang Tiantong Forest" and "Northeast Changbai Mountain Forest".

Although "Zhejiang Tiantong Forest" is a southern forest, the forest ecosystem is not stable. This is because of the deforestation of Zhejiang Tiantong Forest since 2003 due to human deforestation. The forest of Zhejiang Tiantong has been degraded severely, and different forest species have different degrees of damage [5]. Although Zhejiang Tiantong Forest is located in the south, it belongs to "disease".

The forest resources starring in "Northeast Changbai Mountain Forest" are divided into virgin forest, overcut forest and secondary forest. In the past 120 years, the vast virgin forest in Changbai Mountain has almost disappeared, but it has been over 70 years after the national key protection, the ecological
environment recovered well [6]. Therefore, although the forest of Changbai Mountain is in the north, it belongs to "healthy".

The effect of human factors on the forest ecological environment is very significant.

**Figure 11. Map**

5. Conclusion

1). The activity and stability of forest ecosystems are mainly determined by the forest species. Therefore, when planting planted forests, it is necessary to ensure the diversity of forest stands, avoid simple forest structures, and improve the ability to resist natural disasters. Through thinning, the stand density is adjusted to provide sufficient nutrient space for the preserved trees and improve the growth environment. Under-forest cutting and irrigation can promote the growth of other shrubs, increase species diversity, emit poor-quality trees, and improve the quality of forest trees.

2). According to the national soil activity level [4], classify the existing fourteen forest species and plant the appropriate forest species on suitable land. Under different depths of soil, the contents of organic matter and elements are different. Among them, the part that affects the growth potential of the forest is mainly the part of the soil that the root system occupies, and the availability of nutrients and water in this part of the soil. Therefore, it is necessary to constantly observe the soil composition at different depths, study the stability of the forest ecosystem, and judge the health status of the forest.

3). Because the factors affecting climate are very complicated, the method of averaging is not appropriate. It is better to study climate change, or to study multiple fixed-point changes in one area at the same time, so as to draw more accurate conclusions.

Acknowledgments

This paper is supported by Beijing University student scientific research training project of Beijing Forestry University (No. S201910022095), is also supported by the Fundamental Research Funds for the Central Universities (NO. 2015ZCQ-LY-01).

References

[1] Yu Xinxiao. Study on Monitoring and Assessment of Forest Health, Journal of Guangxi Agricultural Science Vol.40 No.8
[2] Li Rongbin. Review of Research Progress on Soil Microbial Biomass Carbon> Guangdong Forestry Science and Technology. Volume 24, Number 6 2008
[3] Ting HUA. Temporal and Spatial Variations in the Climate Controls of Vegetation Dynamics on the Tibetan Plateau during 1982–2011. Advances in Atmospheric Sciences November 2018
[4] Zhang Weili. Review of Soil Classification Studies and Revision of Chinese Soil Classification System, Chinese Agricultural Sciences August 2014.
[5] Guo Ming. Effects of Forest Degradation and Damage on Soil Respiration in Tiantong, Zhejiang. Journal of East China Normal University (Natural Science Edition), July 2011
[6] Yu Dapao. History and Research History of Broad-leaved Korean Pine Forest in Changbai Mountain, Chinese Journal of Applied Ecology, 2019-05.