ISSN: 2688-8653

Determine the Ecological Reserves by Random Forest and K-means Method

Hongyi Yang[1], Yueheng Wu[2], Yajie He[2]

[1]Nanjing Tech University, Nanjing, 211816, China

[2]Jiangxi University of Finance and Economics, Jiangxi, 330013, China

Abstract

This paper extracted seven indicators through the analysis of the relevant reasons of whether to establish the ecological area and constructed a 0-1 classified data set through a data query. A two-level classifier model combining Random Forest, K-means clustering, and distance classification were established. Firstly, the Random Forest classifier was trained with classification data set, and the parameters were adjusted by the grid search method. Then, the accuracy of the model was verified by 86%, which was used as the first-level classifier to decide whether to establish ecological regions. Secondly, K-means clustering was used for the straight label data in the dataset, and the mean value was taken for each category. The second level classifier was established based on the principle of nearest distance to determine the number of ecological regions.

Keywords

Random Forest; Multilevel Classifier; K-means.

1. Introduction

As China holds the largest population and carbon emissions, reaching the peak carbon dioxide emissions before 2030 is a goal of the Chinese government. To achieve this, building ecological reserves is one of the most effective ways.

It is a "natural oxygen bar", which can greatly alleviate the excessive carbon emissions caused by human activities; it can preserve many species and various types of ecosystems and provide a place for human beings to study natural ecosystems. It is a "natural laboratory" for ecological research and convenient for continuous and systematic long-term observation.

2. Determine the Ecological Reserves

2.1. Index Selection and Data Collection

2.1.1. Index Selection

To measure whether one should set up nature reserves, the indices can be divided into natural factors and human factors.

Natural factors include the area's natural landscape proportion, this is the most fundamental condition of establishing nature reserves. The second to consider is the plentiful of scarce animals & plants. Establishing reserves in regions with large numbers of rare species is critical to protecting species diversity.

Human factors include local population, air pollution degree, increasing rate of industrial output value, local GDP, and GDP of the province where it is located, which will all affect the capital investment capacity of establishing nature reserves, along with the transformation and development of the region.

ISSN: 2688-8653

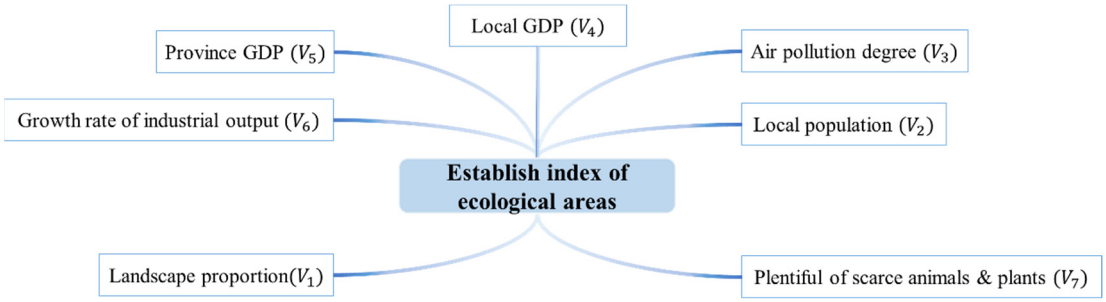

Figure 1. Diagram of Index

2.1.2. Data Collection

Firstly, we select some areas with established nature reserves around 2000 and collect the data of the years near the establishment year, as well as the data 20-25 years before the establishment year. For the year having reserves, we label it as 1; for the year without reserves, we label it as 0.

Table 1. Example: Data with Reserves

Year	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Label
2003	11.33	21	0.21	16.8	0.04	0.02	1	1
2004	12.05	19	0.11	16.3	0.03	0.017	1	1
2005	11.77	16	0.12	15.7	0.026	0.015	1	1
2006	11.45	15	0.15	15.2	0.025	0.012	1	1
2007	11.32	13	0.104	14.7	0.027	0.011	1	1
1985	10.92	6.86	0.01	9.8	0.017	0.012	1	0
1984	10.87	5.33	0.011	9.5	0.019	0.013	1	0
1981	10.23	5.21	0.013	8.3	0.013	0.012	1	0
1979	10.44	4.75	0.009	8.1	0.015	0.015	1	0
1978	10.76	4.21	0.007	7.9	0.012	0.011	1	0

To ensure coverage rate, we also include data from areas where nature reserves had never been established. Then the situation and basis of establishing nature reserves in various areas of our country are summarized.

Table 2. Data without Reserves

V_1	V_2	V_3	V_4	V_5	V_6	V_7	label
0.32	118	0.19	113	0.41	0.411	0	0
0.36	150	0.18	76	0.31	0.312	0	0
0.29	144	0.13	87	0.28	0.311	0	0
0.35	178	0.11	112	0.29	0.289	0	0
0.31	203	0.14	31	0.27	0.287	0	0

2.2. Random Forest and K-means Clustering Bilayer Classifier

2.2.1. Theory of the Random Forest Model

The Random Forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each tree to try to create an uncorrelated forest of trees whose prediction by the committee is more accurate than that of any individual tree.

First of all, a training set is generated by the bootstrap method. After constructing a decision tree for each training set, the nodes to find characteristics were falling apart and can not find

ISSN: 2688-8653

the characteristic of all indicators (such as information gain) of the largest. But in the random part of characteristics, in the middle of the chosen characteristics to find the optimum solution, and applied it to the nodes to split. The Random Forest approach avoids over-fitting because of bagging, the idea of integration, which is equivalent to sampling both samples and features (if the training data is treated as a matrix, as is common in practice, it is a process of sampling both rows and columns).

Information entropy is a common index to measure the purity of a sample set. Let p_k (k=1,2,...,m) represent the proportion of k-th sample in the input sample set D. Assuming there are m types in total, then the information entropy of set D is:

$$\mathrm{Ent}(D) = -\sum_{k=1}^{m} p_k \log_2 p_k \tag{1}$$

The smaller the $\mathrm{Ent}(D)$ is, the higher the purity of set D.

Assuming that the attribute a has V types of values, then the input data set D can be divided into V branches according to the attribute a. And the information gain obtained according to the attribute is:

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} |D_v||D|\mathrm{Ent}(D_v) \tag{2}$$

Among this, D_v represents the gather of branches obtained from one of the value a_v based on attribute a.

The larger the information gain is, the greater the purity will be improved by using attribute a for partitioning. Therefore, attributes can be selected according to the information gained.

The reason for this wonderful effect is that the trees protect each other from their errors (as long as they don't constantly all error in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees can move in the correct direction. So the prerequisites for Random Forest to perform well are:

(i) There needs to be some actual signal in our features so that models built using those features do better than random guessing.

(ii) The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

2.2.2. Establish the Model

We apply the model to a training set containing 130 pieces of data, which also holds a balance of the existence of reserves.

Firstly, we use K-cross validation to make the generalization error closer to the real model representation. Due to the small number of samples, we divide the original data into 7 groups and make a validation set for each subset, while the remaining 6 groups of subset data were used as training sets.

In this way, 6 models were obtained, and the accuracy reached 97.7% after training. The Area Under Curve (AUC) was 0.98, which was very close to 1. We reckon it is pretty good at recognizing things. The confusion matrix shows that 128 out of 130 pieces of data are correctly identified.

The next step is the tuning of parameters. The two main parameters are the number of decision trees and the maximum number of splits. We choose grid search, which is a tuning technique that attempts to compute the optimum values of hyperparameters. We set the range of decision

ISSN: 2688-8653

trees in a Random Forest from 10 to 200 and the maximum number of splits in a decision tree from 100 to 300.

Table 3. Range of Parameters

Parameter	Min	Max
Decision Trees Number	10	200
Maximun Number of Splits	100	300

We use the harmonic mean F_1 of precision rate and recall rate to measure the accuracy of the model. The formula is as follows:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

The optimal value comes out to 0.99, as shown in Figure 2.

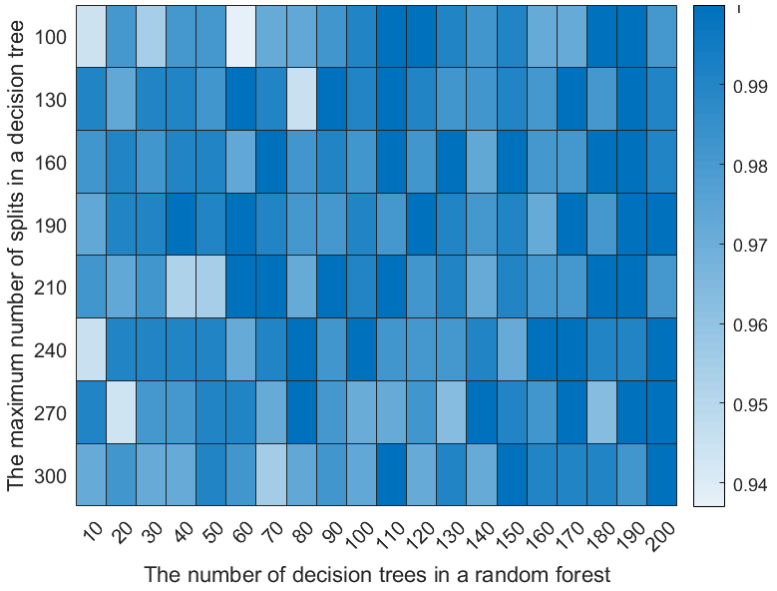

Figure 2. Grib Search Tuning Technique

We nailed down the number of decision trees as 40 and the maximum number of splits as 190. After tuning the parameters, the accuracy rises to 98.8%, and AUC changes to 0.99.

We also take other classifiers into account, such as the logistic regression, but it generally doesn't give discrete output so the accuracy it reaches is only 84.2%. In contrast, the forest classifier performs better with more categorical data than numeric. As for SVM, it is intrinsically two-class so the accuracy of this classifier is merely 83.2% whereas Random Forest is intrinsically suited for multiclass problems. When it comes to a Naive Bayes classifier, a key challenge is that if a categorical variable has a category that was not checked in the training data set, then the model will assign a 0 (zero) probability, which makes it unable to predict, so finally, the accuracy is 80.2%. The comparison of accuracy is displayed in Table 4.

Table 4. Accuracy Comparison

Model	Random Forest	Logistic Regression	Support Vector Machine	Naïve Bayes
Accuracy/%	98.80	84.20	83.20	80.20

ISSN: 2688-8653

2.2.3. Test the Model

First, we search for some data that is both crosswise and lengthways to test the model and compare the results with true value. We choose three cities to verify the model and search 5 years of data in every one of them. The first two cities had reservations around 2000 and the data was selected from 1980-2000 and 2000-2015 equally. The last city does not have a reservation until now. After applying this model, 13 out of 15 results are the same as the true value.

Then we aim to test the reasonability of the index, so we calculate the weight of every index based on the test set, as shown in Table 5.

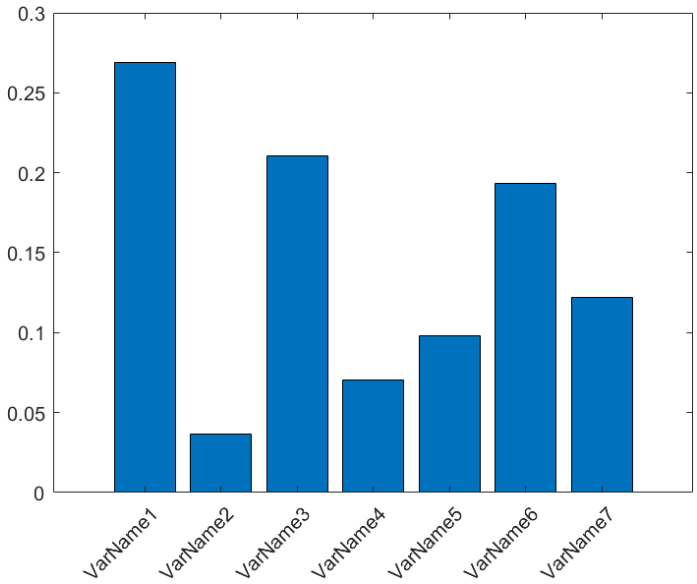

Figure 3. Column histogram of indices

From Figure 3, we can deduce that the proportion of landscape, local GDP, and the scarcity of animals & plants are the substantial indices we need to take consider in, which are identical to the information provided by research and documents. In contrast, the local population is not a primary condition that only needs to be larger than a particular threshold value.

We have enough evidence to conclude that the choice of indices is reasonable.

2.2.4. K-means Clustering to Determine the Scale

As for the scale of each reservation, we use K-means Clustering to solve this problem. Based on the preceding process, 78 areas that are labeled 1 in the training set were selected. We define the target number k as 2, then the dataset $X = \{x_1, x_2, \dots x_m\}$ is divided into two clusters (C_1, C_2). Their initial centroid vectors are $\{\mu_1, \mu_2\}$:

$$\mu_i = \frac{1}{|C_i|}\sum_{i \in C_i} x \tag{4}$$

After that, every data point is allocated to each of the clusters by reducing the in-cluster sum of squares, that is, minimizing the square error E.

$$E = \sum_{i=1}^{k}\sum_{i \in C_i} \|x - \mu_i\|_2^2 \tag{5}$$

ISSN: 2688-8653

Now define v_i as a 7×1 vector of indices of the place x_i we are going to evaluate, N as the number of ecological areas aimed to be built, and D as the distance between vectors. Those two clusters respectively represent small scale, that is, the number of reserves in the city is 1-2, and large scale, in which the number rises to 3-5 in one place, with the following expression:

$$N = \begin{cases} 1-2 & D(v_i, \mu_1) > D(v_i, \mu_2) \\ 3-5 & D(v_i, \mu_1) < D(v_i, \mu_2) \end{cases} \quad (6)$$

From the clustering results of all positive samples in the training data and the mean values of the two types of indicators, there is little difference in the proportion of natural landscape between the two categories in this indicator, which indicates that the natural landscape of all cities that should establish ecological zones must reach a certain range.

Secondly, in terms of population, GDP, and industrial change, the mean value of the first type is significantly higher than that of the second type. Moreover, we also found that most of the cities in the first category have established multiple ecological zones, which indicates that the number of ecological zones should be considered from the perspectives of population, economy, and industry. The population provides the necessity and labor for the construction of many ecological zones. The economy provides financial support for the construction of the eco-zone.

Compared with a large and wide ecological zone, multiple and smaller ecological zones are more flexible and can balance the contradiction between industrial land and ecological land.

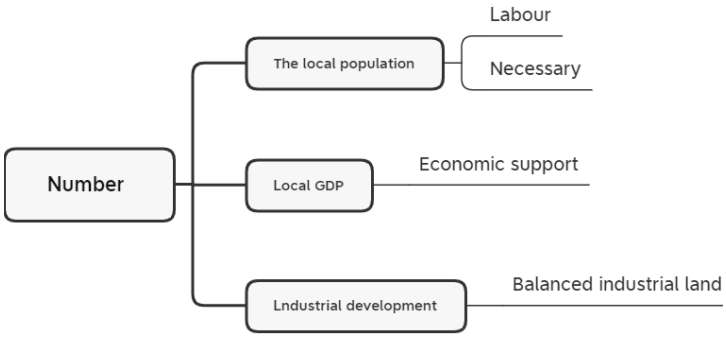

Figure 4. Influence Factors on the Scale

In summary, it comes out that if one place has a high proportion of natural landscape or many scarce animals and plants, it is more possible to build more than 3 reserves. On the other hand, if one place has a developed industry and large population, it tends to only establish one reserve. The result accords with the explanation of indices.

2.3. Result Analysis

2.3.1. The Necessity of the Establishment

The main goal of this model is to test whether a natural reservation should be built in a particular city. Based on the results, cities that need to build reservations can be classified into three types.

The first type is the place with a large proportion of landscape, small population, and low GDP. The reason for the establishment is the plentiful scarce animals & plants there and an urgent need to protect the diversity of creatures.

The second type is aimed at mitigating industrial hazards. For some places that used to be hostile but now are overwhelmed by pollution problems whereas the surge of GDP and economic situation, it is time to save the environment.

ISSN: 2688-8653

The third type is cities with a low level of population, GDP, and development of industry, but is poor the air quality.

The index value of the three types is demonstrated in Table 5.

Table 5. Example Data of Three Types

Type	V_1	V_2	V_3	V_4	V_5	V_6	V_7
First	34.78	4.3	0.098	5.1	0.071	0.018	2
Second	10.34	347.7	0.55	49	0.39	0.78	0
Third	5.43	44.7	0.77	48.5	0.11	0.43	0

2.3.2. Examples in Practice

We take some representative places across the country to decide if they should build reservations and if so, on what scale should they build. Given the rationality of index selection and weight distribution, we select cities based on their weight. The data of these cities are displayed in Table 6.

Table 6. The Data of Examples

City	V_1	V_2	V_3	V_4	V_5	V_6	V_7	Label
A	49.5	47	0.23	1.84	0.145	0.17	0	0
B	33.3	54	0.33	16.9	0.251	0.14	0	0
C	3.11	318	0.67	34.1	0.277	0.33	0	0
D	33	201.4	0.77	54.1	0.258	0.22	0	1
E	29.8	75.1	0.89	0.238	0.138	0.17	0	1
F	0.141	920	0.86	101.4	0.731	0.44	0	0
G	0.387	908	0.78	73.19	0.99	0.43	0	0

For those places that are not suitable for eco-zone, the reasons are given as follows. F and G are the same types of cities. They don't have the main condition: enough landscape, so they have to take other measures to reduce air pollution, such as replacing polluting energy with new energy sources.

The reasons why A and B have no need to build lying on large territory and sparsely populated so that their ecology has not suffered much damage. Moreover, their industry and economy are still developing. Likewise, C is a developing city with a certain natural landscape and tourism, but the economy is not too developed and pollution is acceptable, so there is no need to build an ecological zone.

As for the reasons why D and E set up ecological zones, firstly is the large proportion of landscape. Moreover, the pollution caused by industrial development began to be serious in D, so with sufficient economic and labor support, it was necessary to set up ecological zones to relieve the ecological pressure. The quantity of the reserves is 3-5 because $D(v_i, \mu_1) = 102.33$ which is smaller than $D(v_i, \mu_2) = 137.57$.

As for E, its landscape is vast and sparsely populated, and the air pollution is due to the influence of the geographical environment. Now E's population has increased to a certain scale leading to the necessity to build. The quantity of the reserves is 1-2 because $D(v_j, \mu_1) = 238.93$ which is bigger than $D(v_j, \mu_2) = 39.6$. In line with the principle of "people-oriented", it is necessary to establish ecological protection areas and artificial afforestation.

These examples also fit the three main types of building ecozones that we summarized earlier.

ISSN: 2688-8653

2.3.3. Positive Effect on Carbon Neutral

The contribution of green vegetation to carbon emission reduction and carbon neutralization is profound. In terms of the ability of carbon absorption among different plants, evergreen plants, broad-leaved plants, and bunched bamboo are relatively stronger.

The land condition in D is suitable for planting sycamore trees, which can sequester carbon dioxide from the atmosphere, then use it for photosynthesis and store it as cellulose in their trunks, branches, and leaves. It is one of the best tree species for carbon storage and sequestration. D's industrial development level is not so high, but the total apparent CO_2 emissions in 2017 reached 7.22 million tons, so carbon emission reduction is very necessary for this area.

E is dry all year round and has unique conditions of light, heat, and water resources. It is suitable for planting shrubs or semi-shrubs. They are often born in the desert or at the edge of the desert. They can resist drought, wind, and sand, as well as withstanding light and moderate saline-alkali environments. With the goal of carbon neutrality, efforts to increase forest cover, wind prevention, and sand fixation in ecologically fragile areas have become the focus of policy.

3. Model Evaluation

The establishment of a multi-layer classifier to solve the establishment and number of ecological areas is more reliable and scientific from the perspective of actual data.

In the classification model, the weights of indicators are not taken into account. If a weighted classification model can be established, the model results and accuracy may be improved.

References

[1] Research on Saihanba Forest Farm (1962-2017) -- Master's Thesis of Hebei Normal University (CNki.com.cn).

[2] Wind and Weather regularity in Saihanba, Zhang Shushan 25 29 (CNki.net).

[3] Cause Analysis of "3·15" Strong Sandstorm in North China, Duan Bolun (CNki.net).

[4] Dust Storms in US History 35 (CNki.net).

[5] Current Nature Reserves and Their Distribution in China (baidu.com).

[6] A Study on temperature Variation in Saihanba - Docin.com.

[7] Eco-city Environmental Index System (baidu.com).