

Research On the Pricing Model of Second-Hand Sailboats Based on GDBT Model

Jieli Chen*, Jiewen Zhong, Zenghao Chen

School of Economics, Guangdong University of Technology, Guangzhou, China, 510000

*Corresponding author: jerry3556236546@163.com

Abstract. Sailboats have varying values with changes in market conditions and aging. This paper analyzed the characteristics and economic characteristics of monohull along with catamaran in their respective regions. The paper also carried out Spearman correlation analysis on their listing prices. The gradient boosting decision tree model was established to calculate the feature importance of each variable on the listing price, and the accuracy of the model reached over 86%. It indicated that the model can effectively price the second-hand sailboat market. This paper used one-way ANOVA to determine whether there was a significant difference between the listing prices of monohull and catamaran in the secondary markets. There were significant differences in the listing prices of monohulled sailboats and catamarans in different geographical regions. The listing prices of monohull and catamaran in different geographical regions were significantly different, and the regional effects were inconsistent.

Keywords: Second-hand Sailboat Pricing, Gradient Boosting Decision Tree Model, Regional Effect, Kruskal-Wallis Test.

1. Introduction

1.1 Background

Shipowners are more interested in second-hand ships than in the market for new ships due to considerations such as budget and delivery time. The deviation of the value judgment of the two parties to the ship trading leads to poor liquidity in second-hand sailboat trading, so the final sailboat trading price is usually determined after a sailboat broker's inspection. In terms of sailing brokers, they need to consider the value of second-hand sailing boats, investigate the models and materials of second-hand sailing boats in different regions. The sailing brokers evaluate the price of second-hand sailing boats based on various indicators that may affect their price, and make an accurate evaluation of their price to promote market trading.

1.2 Literature Review

Goulielmos [1] pointed out that the Greeks and a few maritime countries were aware of the importance of the second-hand shipping market and concluded that shipowners could predict the price of second-hand ships, during which time they could choose the highest or lowest price. Liang FANG et al.[2] studied the dynamic relationship between price and trading volume in the second-hand ship market based on the VAR-GARCH model, and concluded that there is a unilateral causal relationship between past trading volume and volatility in the return rate of all second-hand ship markets. Roar Adland [3] explored whether energy efficiency will affect the value of ships in the second-hand market, concluded that there was a negative correlation between energy efficiency and sales prices, with an elastic value of around 0.4, indicated that the actual operating energy efficiency depends on the speed of ships. Lim S S [4] used artificial neural network models outperform simple stepwise regression analysis and satisfied both statistical soundness and accuracy of results inscientific models. Floriano [5] established a model of the relationship between shipbuilder countries and second-hand ship prices and concluded that ships built in Japan and Europe often receive higher prices than other countries. Azhar A [6] builded a cost estimation model of a used ship to determine the price of a used ship andbuilds a cost estimation model of a used ship. The data was processed and analyzed by multiple methods including market price method, comparative ship and physical pricing method, and the

estimated price or appraised ship is obtained from the average of the three methods. Nam H S[7] argued that the prices of used and new ships were closely related and influenced by market dynamics. The results showed that the ship's main engine type and country of construction were statistically significant in most ship categories, while other ship specific and economic factors also explained the value of second-hand ships consistent with the literature.

We possessed a batch of second-hand sailboat data. We used useful predictive factors such as the characteristics of specific sailboats and regional annual economic data to explain the search price for each sailboat. The accuracy of price estimation for each type of sailboat was discussed. The impact of regions on pricing was explained based on the established model. Then we discussed the impact of regions on sailboat prices and determined whether the regional effects were consistent for all sailboat species.

1.3 Main Work

We conducted Spearman correlation analysis between the characteristics of the two types of sailboats and the listing price respectively to determine whether the variables were significantly correlated with the listing price. The gradient boosting decision tree model was established to calculate the characteristic importance of each variable to the listing price, and to evaluate the accuracy of the model results.

We used one-way ANOVA to explore whether there was a significant difference between the listing prices of the two types of sailboats in the secondary market in different regions. Sailboat characteristics and economic data from different regions were substituted into the GBDT model developed to discuss the effect of different regions on listing prices. Kruskal-Wallis tests were conducted separately for the listing prices of the two sailboat varieties in different regions to discuss the effect of region on sailboat prices and to determine whether the region effect was consistent for all sailboat varieties.

1.4 Arrangement

This paper was organized as follows. Section I introduced the research background, literature review, research problem and analysis of this paper. Section II presented the principles and assumptions of the three models used in this paper: GBDT, one-way ANOVA, and Kruskal Wallis test principle, and proposed the basic assumptions of this article. Section III applied the GBDT model to explain the listing price of each sailboat and discussed the accuracy of price estimation. Section IV explained the impact of different regions on the price based on the establishment of the GBDT model, and judged whether the regional effect is consistent for all sailboats. Section V summarized the research questions, methods, results and implications of this paper.

2. Preliminary

2.1 Gradient Boosting Decision Tree

The gradient boosting decision tree algorithm is one of the more advanced machine learning strategies available[8]. The GBDT algorithm has excellent performance in regression and classification problems.

The GBDT algorithm expression is:

$$F(x, P_{arm}) = \sum_{n=0}^N \alpha_n T_n(x, P_{armk}) = \sum_{n=0}^N f_n(x, P_{armk}) \quad (1)$$

n is the subregression decision tree ($k = 0, 1, \dots, K$)

2.2 One-way Analysis of Variance

One-factor ANOVA is an analysis of one category of independent variables on numerical variables. Also known as one-way ANOVA, analysis of whether one factor has a significant effect on the outcome [9].

2.3 Kruskal-Wallis Test

The Kruskal-Wallis test is essentially a generalization of the Mann-Whitney U test for two independent samples to multiple independent samples, and is used to test whether the distributions of multiple aggregates are significantly different[10].

2.4 Assumption

We made the following reasonable assumptions and conditional constraints based on the actual situation to construct a more accurate mathematical model.

Hypothesis 1: The pricing of used sailboats is related to the market environment in their region, excluding other unrelated factors. The price of a sailboat is influenced not only by the characteristics of the sailboat itself, but also by the supply and demand in the market.

Hypothesis 2: The selling price of a sailboat has a lot to do with its geographical location. Because market conditions and supply and demand vary geographically, this has an impact on the selling price.

Hypothesis 3: It is assumed that regional economic indicators can be reflected by GDP and GDP per capita.

Hypothesis 4: The data provided in the topic are true and reliable to a certain degree. Because the model we built is based on the data provided in the topic, Only the high validity of the data can guarantee the high reliability of the model.

2.5 Notations

The symbol description of the paper was shown in Table 1.

Table 1: Notations

Symbol	Description
x_{Year}	Year of production of the sailboat(<i>ft</i>)
x_{Length}	Length of the sailboat (<i>ft</i>)
x_{LWL}	Waterline length of the sailboat (<i>ft</i>)
x_{Beam}	Beam of the sailboat (<i>ft</i>)
x_{Draft}	Draft of the sailboat (<i>ft</i>)
x_{Dis}	Displacement of the sailboat (<i>lbs</i>)
x_{SA}	Sail Area of the sailboat (<i>sq ft</i>)
x_{GDP}	GDP of the region(<i>USD billion</i>)
x_{GPC}	GDP per capita of the region (<i>USD</i>)
y_{Price}	List price of the sailboat (<i>USD</i>)
P_{armk}	GBDT algorithm model parameters

3. Pricing Model for Second-hand Sailboats

3.1 Spearman Correlation Analysis

We conducted Spearman correlation analysis of year, characteristics, regional economic data and listing price for used monohulled sailboats and used catamarans, respectively. The obtained correlation coefficients were shown in Table 2. These variables were significant at the 1% level, indicating that these variables explained the listing price of each used sailboat.

Table 2: Spearman Correlation Analysis

	Year	Length	LWL	Displacement	Beam
monohulled sailboat	0.434 (0.000***)	0.542 (0.000***)	0.484 (0.000***)	0.583 (0.000***)	0.446 (0.000***)
	Sail Area	Draft	GDP	GDP per capita	
	0.5 (0.000***)	0.379 (0.000***)	0.199 (0.000***)	0.205 (0.000***)	
catamaran	Year	Length	LWL	Displacement	Beam
	0.586 (0.000***)	0.597 (0.000***)	0.537 (0.000***)	0.563 (0.000***)	0.613 (0.000***)
	Sail Area	Draft	GDP	GDP per capita	
	0.625 (0.000***)	0.521 (0.000***)	0.124 (0.000***)	0.034 (0.000***)	

3.2 Gradient Boosting Decision Tree Model

3.2.1 Modeling Steps

Step1: A gradient boosting decision tree (GBDT) regression model was established using training set data.

Step2: The feature importance was calculated using a gradient boosting decision tree established.

Step3: The established gradient boosting decision tree (GBDT) regression model was applied to training and testing data to obtain model evaluation results.

3.2.2 Modeling Process

Given second-hand sailboat listing price data training set:

$$X = \begin{bmatrix} [x_{Year}, x_{Length}, x_{LWL}, x_{Beam}, x_{Draft}, x_{Dis}, x_{SA}, x_{GDP}, x_{GPC}, y_{Price}]_1 \\ [x_{Year}, x_{Length}, x_{LWL}, x_{Beam}, x_{Draft}, x_{Dis}, x_{SA}, x_{GDP}, x_{GPC}, y_{Price}]_2 \\ \dots \\ [x_{Year}, x_{Length}, x_{LWL}, x_{Beam}, x_{Draft}, x_{Dis}, x_{SA}, x_{GDP}, x_{GPC}, y_{Price}]_{n-1} \\ [x_{Year}, x_{Length}, x_{LWL}, x_{Beam}, x_{Draft}, x_{Dis}, x_{SA}, x_{GDP}, x_{GPC}, y_{Price}]_n \end{bmatrix} \quad (2)$$

The strong learner expression of the model is

$$F_n(x) = F_0(x) + \sum_{n=1}^N \sum_{k=1}^{K_n} c_{n,k} I(x \in R_{n,k}) \quad (3)$$

3.2.3 Result of the Model

We used gradient boosting decision tree (GBDT) model to regress the data, and Figure 1 showed the characteristic importance ratios of the respective variables for monohulled sailboats and catamaran.

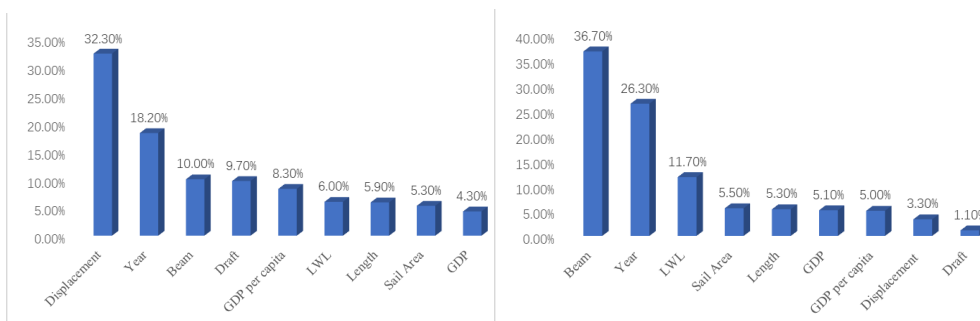


Figure 1: Ranking of Listing Price Characteristic Values of Monohulled Sailboat and Catamaran

Overall, displacement was the most important factor affecting the listing price of monohulled sailboats, with an importance of 32.30%, followed by year at 18.20%. The impact of other factors was less than 10%. This suggests that the price of monohulled sailboats is mainly influenced by displacement and year, with these two factors accounting for over 50% of the cumulative importance.

For catamarans, beam was the most important factor affecting the listing price, with an importance of 36.70%, followed by year at 26.30% and LWL at 11.70%. The impact of other factors was less than 10%. It suggested that the price of catamarans was mainly influenced by beam and year, with these two factors accounting for over 60% of the cumulative importance.

From this perspective, it was clear that the price formation of second-hand sailing boats in a given year is not negligible, as the price of sailing boats changes with their aging. For monohulled sailboats, however, displacement was the most important factor influencing price formation, while for catamarans it was beam.

3.2.4 Model Evaluation

After repeated iterative training of GBDT model, $MAPE$ and R^2 evaluation indexes of the model were calculated based on the five-fold cross validation test set.

Table 3: Model Evaluation Result

Type		MAPE	R^2
monohulled sailboat	Training set	6.473	0.977
	Cross verification set	15.329	0.788
	Test set	15.708	0.864
catamaran	Training set	3.046	0.979
	Cross verification set	10.071	0.792
	Test set	9.295	0.862

The model evaluation results were shown in Table 3, MAPE (mean absolute percentage error) is a percentage value. The smaller the value, the more accurate the model is. Compared with the predicted value when only the mean is used, the closer the R^2 value is to 1, the more accurate the model will be, indicating that the model has a good fitting effect. The R^2 value of the training set of monohulled sailboat and catamaran was as high as 97%, and the R^2 value of the test set was more than 86%, indicating that the model has a good fitting effect and a high accuracy in estimating the price of each sailboat.

Figure 2 showed the prediction of Gradient Boosting Decision Tree (GBDT) on the test data.



Figure 2: Test Data Prediction Diagram of Monohulled Sailboat and Catamaran

As can be seen in Figure 5, the curves of the true and predicted values were very close to each other and predict the test data very well. It further indicated that the gradient boosting decision tree model had high accuracy in estimating the prices of each sailboat.

4. The Effect of Region on Listing Prices

We conducted one-way ANOVAs on the listing prices of second-hand monohulled sailboats and catamarans separately for geographic regions. The results were shown in Table 4.

Table 4: Model Evaluation Results

Type	Variable Name	Region	Sample size	Average value	Standard deviation	F	P
Monohulled Sailboat	Listed price (USD)	USA	375	237652.635	113527.943	11.026	0.000***
		Europe	1745	213992.752	111619.274		
		Caribbean	178	193937.309	91291.222		
Catamaran	Listed price (USD)	USA	104	465940.519	190940.256	8.884	0.000***
		Europe	728	453907.81	171264.696		
		Caribbean	295	407738.637	155021.207		

Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively

It can be found that among used sailboats, the p-value of the results of the one-way ANOVA on the listing price of used sailboats in Europe, the United States and the Caribbean was less than 0.05. The results were statistically significant, indicating that there were significant differences in the listing price of single-hull sailboats in different geographic regions.

We entered the data for second-hand monohulled sailboats and second-hand catamarans from different regions (USA, Europe, Caribbean) into the mathematical model we developed and obtained the proportional importance of the characteristics of the respective variables for monohulls and catamarans from different regions.

The model evaluation results obtained using the GBDT model were shown in Table 5.

Table 5: Model Evaluation Results R^2 value

Type		USA	Europe	Caribbean
Monohulled Sailboat	Training set	0.997	0.976	0.996
	Cross-validation sets	0.806	0.812	0.801
	Test set	0.823	0.824	0.886
Catamaran	Training set	0.998	0.973	0.994
	Cross-validation sets	0.755	0.869	0.705
	Test set	0.824	0.828	0.875

The value of the training set R^2 and the value of the test set R^2 of the sailing vessels in the three regions were all over 0.97 and 0.82 respectively, indicating that the fitting effect of the model was good and the price estimation of each sailing species in the three regions was relatively accurate.

We used the GBDT model to regression the data to get the feature importance.

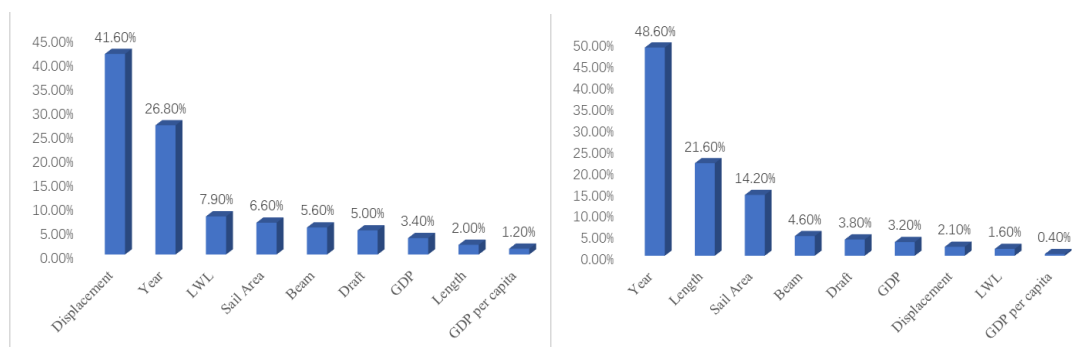


Figure 3 :Ranking of Listing Price Characteristic Values of USA Monohulled sailboats and Catamarans

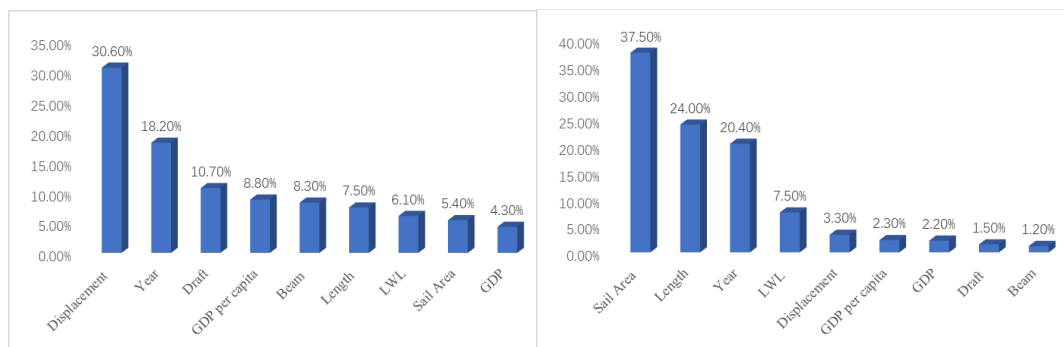


Figure 4 :Ranking of Listing Price Characteristic Values of Europe Monohulled sailboats and Catamarans

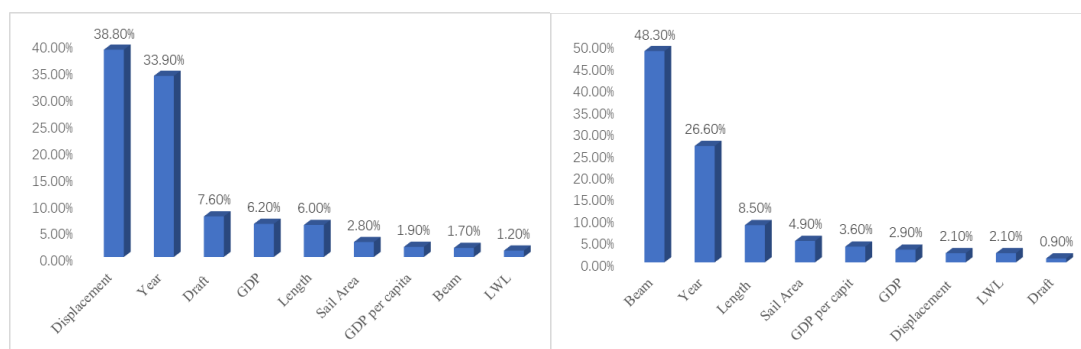


Figure 5 :Ranking of Listing Price Characteristic Values of Caribbean Monohulled sailboats and Catamarans

As shown in Figure 3, in the United States, the importance of Displacement for a monohulled sailboat increased from 32.30% to 41.60%. This was because displacement was one of the most important measures of sailboat size and carrying capacity. The displacement of a sailboat was directly proportional to its carrying capacity. Catamaran length and price were generally positively correlated. It was because longer hulls could provide more space, more facilities and better performance. In terms of both monohulled sailboats and catamarans, sail area became more important. It was because sailboats were powered by wind caught by their sails. Therefore, whether a sailboat could go faster or not played a decisive role. In cases where wind direction was correctly mastered, sail area could be used to obtain greater sail power and maintain smooth sailing.

As shown in Figure 4, the importance of the price characteristics of European second-hand sailboats was basically the same as that of the whole. The importance of Draft features had increased to a certain extent for monohulled sailboats, because draft refers to the depth of the hull in the water. Draft and hull weight were positively related, so the price will increase accordingly. Generally speaking, an increase in GDP per capita will increase consumption levels in the region, so it will also had an impact on the price of monohulled sailboats. For catamarans, reasons for the increase in importance of Sail Area and Length characteristics were similar to those in the United States.

As shown in Figure 5, the importance of year features had increased significantly for monohulled sailboat in Caribbean. Year was an important factor in the price of monohulled sailboats. The price of a new sailboat was higher than that of an old one. In general, the gross domestic product of a region may have affected the price of a single sailboat. If the GDP of a region was higher, then people in that region may have had more money to buy expensive monohulls. So the price of monohulls may have been higher. Beam and length features were increasingly important for catamarans. Because catamarans were wider, which could lead to higher mooring charges and taxes. Typically, docks charged 1.5 to 2 times the docking rate for catamarans. The length of a catamaran was also proportional to the price. It was because longer hulls could provide more space, more facilities and better performance.

The Kruskal-Wallis's test was performed on the listed prices of monohulled sailboats and catamarans in different regions, and the results of the test analysis were shown in Table 6.

Table 6: Kruskal-Wallis Test Analysis Results Table

Type	Region	P	Cohen's f Value
Monohulled Sailboat	USA	0.000***	0.821
	Europe	0.000***	0.63
	Caribbean	0.000***	0.817
Catamaran	USA	0.000***	0.992
	Europe	0.000***	0.399
	Caribbean	0.000***	0.541

The Kruskal-Wallis's test results showed that based on the variable listing price, the p-value of the test was 0.000 less than 0.05 in different regions and therefore statistically significant. This indicated that there was a significant difference in the listing price of different varieties in the same region and the regional effect was inconsistent across all sailing varieties.

Among the monohulled sailboats, the Cohen's f values for the magnitude of difference were 0.821, 0.63, and 0.817 for the United States, Europe, and the Caribbean, respectively, with a large degree of difference. Among the catamarans, the Cohen's f values for the United States and the Caribbean were 0.922 and 0.541, respectively, with a large degree of difference, while the Cohen's f values for Europe were 0.399, with a moderate degree of difference.

5. Conclusion

This paper studied the pricing of second-hand sailboats. We used Spearman correlation analysis to determine whether the characteristics of sailboats and regional economies were significantly related to the listing price. It was found that the two were significantly related. We established the gradient boosting decision tree model to price the second-hand sailboat, with an accuracy of over 86%. When determining the price of second-hand sailboats, the manufacturing year cannot be ignored. For Monohulled sailboats, displacement and width were the most important price forming factors. For catamarans, length and width were the most important factors. We used one-way ANOVA model to explore whether there was a significant difference in the listing prices of monohulled sailboats and catamarans in different geographical regions. There were significant differences in the listing prices of monohull and catamaran in different geographical regions, and the regional effects were inconsistent.

References

- [1] Alexander M. Goulielmos. Did the price of second-hand ships follow a predictable path over the period 1976-2008? [J]. International Journal of Shipping and Transport Logistics, 2010, 2(4): p.383-410.
- [2] Liang Fang, Jin Peng Liang, Jie Wang . Price and Volume Dynamics in Second-Hand Ship Market Based on VAR-GARCH Model [C]. International Conference on Instrumentation and Measurement, Computer, Communication and Control. 2015
- [3] Adland R, Cariou P, Wolff F C. Does energy efficiency affect ship values in the second-hand market? [J]. Transportation Research Part A: Policy and Practice, 2018, 111: 347-359.
- [4] Lim S S, Lee K H, Yang H J, et al. Panamax second-hand vessel valuation model [J]. Journal of Navigation and Port Research, 2019, 43(1): 72-78.
- [5] Floriano C. M. Pires Junior, Ricardo Brito Fiasca, Luiz Felipe Assis . Shipbuilder country and second-hand price [J]. International Journal of Shipping and Transport Logistics, 2020, 12(5): 426-444
- [6] Azhar A. Method for Estimating Price of Secondhand Ship with Multi Method [C] IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2021, 1052(1): 012011.

- [7] Nam H S, De Alwis N, D'agostini E. Determining factors affecting second-hand ship value: linkages and implications for the shipbuilding industry [J]. WMU Journal of Maritime Affairs, 2022, 21(4): 493-517.
- [8] CHEN Tiankai,WANG Guiyong,SHEN Lizhong. Performance prediction of diesel engine based on GBDT algorithm[J]. Automotive Engine,2022,No.262(05):51-58.
- [9] Alassaf M, Qamar A M. Improving sentiment analysis of Arabic tweets by One-Way ANOVA [J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(6): 2849-2859.
- [10] Pu Hu. Introduction and Application of Kruskal Wallis Test Principle [J]. Journal of Xingyi Ethnic Normal University, 2019 (04): 108-111