# Used Sailboat Price Prediction Based on GA-BP Neural Wetwork

Jinda Wang[1], Jinxin Xu[1,*], Dong Tang[1], Liru Xu[2]

[1]College of Civil Engineering, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, 050043

[2]College of Management, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, 050043

*Corresponding author: xujx187@163.com

**Abstract.** The paper presents a comprehensive study on sailboat pricing, considering various factors affecting listing price prediction. Through correlation analysis, eight key indicators influencing sailboat prices are identified. These indicators form the basis of a prediction model using the GA-BP neural network, resulting in a high accuracy rate of 0.87 in price estimation. The research also explores the impact of regions on sailboat listing prices by defining price level variables and selecting five region-representative indicators. The regional effect analysis model, based on ANOVA multiple regression, shows that the region significantly affects the markup, with factors like GDP and waterfront area playing a crucial role. The model's regression coefficients effectively explain the degree of impact. Comparing regional impact factor models using Hong Kong and Croatia data, the study finds that the Hong Kong sailboat price prediction model is accurate, providing valuable insights for the local market. Differences in coastal area and per capita GDP explain the regional price variations between Hong Kong and Croatia for both monohulled and catamaran sailboats. In conclusion, the research contributes to understanding sailboat pricing dynamics and emphasizes the significance of regional factors in price prediction. The findings offer valuable guidance for sailboat buyers, sellers, and market analysts.

**Keywords:** Used Sailboat, Price Evaluation, GA-BP Neural Network, Multiple Linear Regression.

## 1. Introduction

Used sailboats have different prices depending on their condition and market conditions. Because of the special property of "one boat, one condition" [1]and the lack of criteria for measuring the value of used sailing boats, it is important to study the factors affecting the price of used sailing boats to facilitate cooperation and exchange between countries. In recent years, several research studies have focused on developing machine learning-based models for second-hand car valuation. The first study by Liu Chang (2021) from Chongqing University explores a machine learning approach for second-hand car valuation[2]. The author presents a model that employs unspecified machine learning techniques to estimate car values, potentially contributing to the improvement of the existing valuation systems. In a similar vein, Li Xuelei (2020) from Chongqing University of Technology proposes a valuation model based on BP (Backpropagation) neural networks[3]. The study showcases the potential of neural networks in predicting second-hand car prices, which could offer enhanced accuracy compared to traditional methods. Furthermore, Xu Yuanlei (2021) from Hebei University of Economics and Trade proposes an optimized BP neural network model using the particle swarm algorithm for second-hand car valuation[4]. By leveraging the optimization capabilities of the algorithm, the proposed model aims to achieve better performance in estimating car values. Additionally, Yu Zitong (2022) from Dongbei University of Finance and Economics introduces a different application of BP neural networks optimized with genetic algorithms[5]. Their approaches, which involve BP neural networks and optimization algorithms, showcase the potential for improving the accuracy and efficiency of the valuation process in the dynamic used car market. Further research and refinements to these models may lead to more precise and reliable predictions, benefiting both sellers and buyers in making informed decisions.

# 2. The basic funamental of BP neural network

## 2.1. The structure of BP neural network

BP neural network is composed of input layer, hidden layer and output layer[6]. The structure of a three-layer BP neural network with M nodes in the input layer, N nodes in the single hidden layer and L nodes in the output layer is illustrated in Figure1.

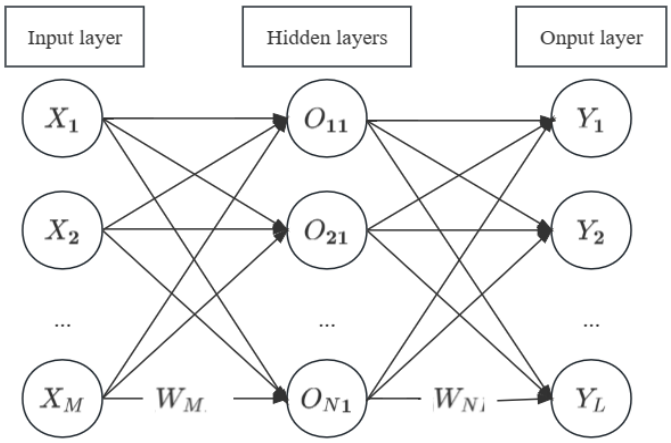


**Figure 1.** Neural network structure

Where the output value of the input layer is $a_i$ (i=1,2,...,M), the output value of the hidden layer is $a_j$ (j=1,2,...,N), and the output value of the output layer is $a_k$ (k=1,2,...,L).

BP neural networks can be divided into two parts:

**Positive dissemination of information:**

First, the paper set the activation function to be a Sigmoid function with the expression:

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

The input $net_j$ of the jth node of the hidden layer is:

$$net_j = \sum_{j=1}^{N} w_{ij} a_i + \theta_j \tag{2}$$

In the above equation,$w_{ij}$ and j denote the weight of the hidden layer and the threshold value of the jth node, respectively.

The output $a_j$ of the jth node of the hidden layer is:

$$a_j = f(net_j) = \frac{1}{1+e^{(1-net_j)}} \tag{3}$$

Input value of the kth node of the output layer $net_k$:

$$net_k = \sum_{k=1}^{L} w_{jk} a_j + \theta_k \tag{4}$$

In the above equation, $w_{jk}$ and k denote the weight of the output layer and the threshold value of the kth node, respectively.

The output ak of the kth node of the output layer is:

$$a_k = f(net_k) = \frac{1}{1+e^{(1-net_k)}} \tag{5}$$

**Back propagation of errors:**

First the paper define the step size as μ and the error calculation formula as:

$$E = \frac{1}{2}\sum_{k=1}^{L}\left(y_{pk} - a_{pk}\right)^2 \tag{6}$$

In the above equation, P and L are the sample size and the number of output values, respectively. The output layer weight correction formula is:

$$\Delta w_{jk} = -\eta \frac{\partial E_P}{\partial w_{jk}} = -\eta \frac{\partial E}{\partial net_k} a_j \tag{7}$$

The derivation of the above equation yields:

$$\delta_k = -\frac{\partial E}{\partial net_k} = a_k(1 - a_k)(y_{pk} - a_k) \tag{8}$$

So the output layer weights result in:

$$w_{jk}(k+1) = w_{jk}(k) + \eta \delta_k a_j \tag{9}$$

The correction formula for the hidden layer weights is:

$$\Delta w_{ij} = -\eta \frac{\partial E_P}{\partial w_{ij}} = -\eta \frac{\partial E}{\partial net_j} a_j \tag{10}$$

Deriving the above equation yields:

$$\delta_j = -\frac{\partial E}{\partial net_j} = a_j\left(1 - a_j\right)\sum_{k=1}^{L} \delta_k w_{jk} \tag{11}$$

So the result of finding the hidden layer weights is:

$$w_{ij}(k+1) = w_{ij}(k) + \eta \delta_j a_i \tag{12}$$

### 2.2. The determination of the number of network layers

The paper optimize the BP neural network by using a genetic algorithm[7]: first, for the hidden layer neuron nodes, and second, for the initial weights and thresholds of the neural network.

The paper use Matlab software to build the model. The initial weights and thresholds of the neural network are first found to be optimal by a genetic algorithm. Then the optimal solution is assigned to the neural network, and the BP neural network can make data prediction by the optimized weights and thresholds.

## 3. Results

There are nine possible influencing factors, including the average price sold by the existing manufacturer, length of the boat (feet), year of manufacture, GDP of the region, GDP per capita of the region, beam, draft, displacement, and sail area, and all of them are continuous variables. Since there is a great difference in the magnitude between different factors, which affects the similarity between the samples, resulting in not being able to filter out the reasonable variables correctly. Therefore, the paper standardized the data line. The data were normalized to the interval $[y_{min}, y_{max}]$. The large data prediction model for the user's electricity consumption is implemented in the Clementine software.

The paper performed correlation analysis on the constructed variables: the Pearson correlation coefficient between the variables was calculated. A positive result obtained indicates a positive correlation, while a negative value indicates a negative correlation. A higher absolute value indicates a higher correlation. Figure 2 shows the obtained results.

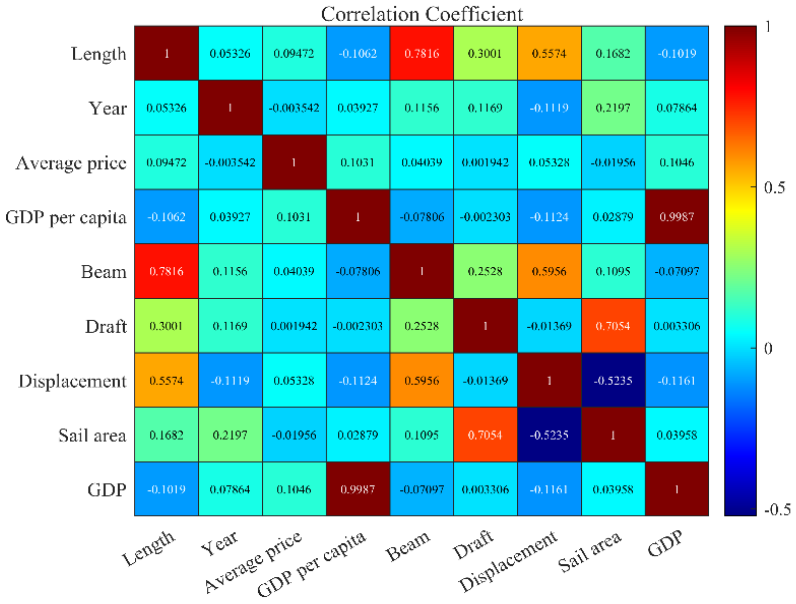


**Figure 2:** Correlation of each variable

From figure 2, the paper can see that the correlation between GDP and GDP per capita is as high as 0.9987. The absolute values of the correlation between the remaining two variables are below 0.8 and cannot be ignored. Therefore, the paper choose the remaining eight factors other than GDP as independent variables and the markup of sailboats as dependent variables to construct the following model.

### 3.1. Price Prediction Based on GA-BP Neural Network

The paper solve the model through programming, first the paper set the parameters in the model, the number of neurons in the hidden layer, the number of overall evolutionary iterations, population size, crossover probability, and mutation probability are 6, 20, 10, 0.4, 0.05, respectively. Then the paper run the code to obtain the optimal individual fitness curve as shown in Figure 3.

Ultimately, the fit of the GA-BP neural network model to a single sailboat is obtained as shown in Figure 4.

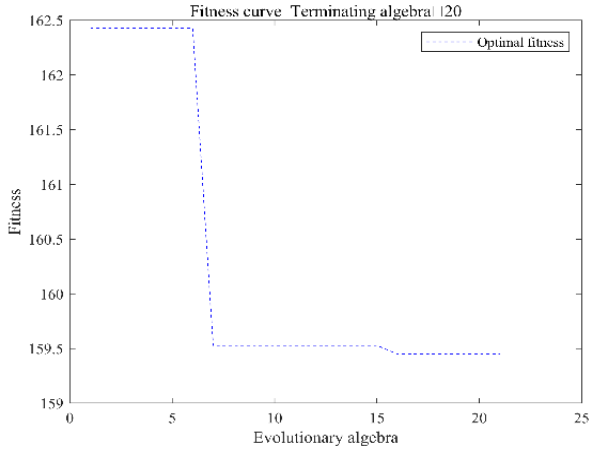


**Figure 3:** Optimal individual fitness curve

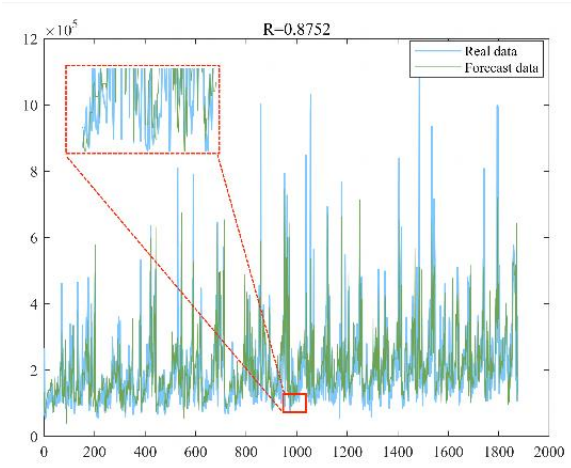


**Figure 4:** Fitting effect

In Figure 4, the true values are in blue and the model predicted values are in green. The predicted and true values show roughly the same curve trend in the figure, and it can be tentatively concluded that the model fits the monohull better. Using the same model to solve for catamarans, the paper can also get better fitting results for catamaran sailing price.

### 3.2. Analysis of the Accuracy of Sailboat Price Prediction

The paper collated the used monohull price estimates and the used monohull real values output by the GA-BP price prediction model into a table and calculated the prediction accuracy of the dataset separately, as shown in Table 1.

**Table 1:** Accuracy of sailboat price forecasting

| Make | Variant | Real Listing Price | Forecast Listing Price | Prediction accuracy |
|---|---|---|---|---|
| Alabama | 423 | 5.45 | 5.26 | 0.034 |
| Bahamas | Cruiser 50 | 7.32 | 7.78 | -0.063 |
| California | 41 DS | 2.64 | 2.89 | -0.098 |
| Dominican Republic | 40S Cruiser | 16.21 | 16.01 | 0.012 |
| France | 37 B&C | 3.49 | 3.31 | 0.052 |
| Georgia | First 44.7 | 4.6 | 4.31 | 0.063 |

The average prediction accuracy of the training set of the neural network optimized by the genetic algorithm is 89.62%, and the training result of the model is better. This shows that the accuracy of the price assessment of each monohull sailing vessel variant is high. Using the same method to calculate the accuracy of the price evaluation of catamaran sailboats, it can be similarly concluded that the accuracy of the price evaluation of each catamaran sailboat variant is also higher.

### 3.3. Forward Multiple Linear Regression Model

#### 3.3.1 Define Price Level

The listing price of sailboats varies from region to region and from year to year. To discuss only the effect of region on listing price, the paper define a new variable, price level (PL). Price level is defined by the equation 13:

$$PL = \frac{LP}{GDP_{per}} \tag{13}$$

Where LP stands for Listing Price(USD) and GDP per stands for Gross Domestic Product per capita.

GDP per capita varies from region to region and from year to year, so the paper use equation 13 to eliminate the effect of year on price.

#### 3.3.2 ANOVA Regional Effects on Price Levels

First, the paper apply the model of the first question to predict the predicted value of the listing price for all the given used sailboats in the Annex. The price level is found according to the equation 13. Figure 5 shows the predicted versus the true value.

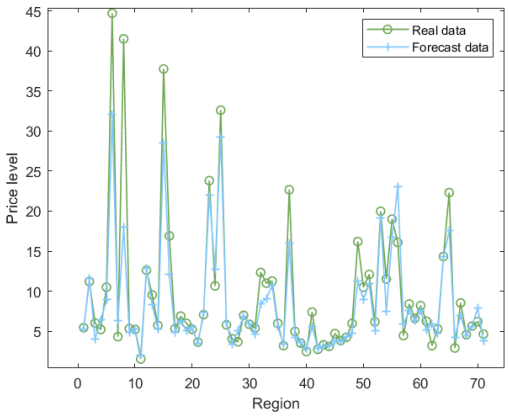


**Figure 5:** Comparison of predicted and true values

As can be seen from Figure 5, its predicted value is very close to the true value. To some extent, it can be used for the prediction of other cases.

Next, the paper performed ANOVA on 79 different regions and price levels. ANOVA is applied to analyze the relationship between categorical and continuous variables. In this problem, region is a categorical variable and price level is quantitative. Therefore, the paper chose to use ANOVA. Based on the type of data, the paper choose the F-test. This type was judged to be a one-way ANOVA. This leads to the ANOVA model with the formula 14:

$$F = \frac{SS_A/(m-1)}{SS_E/(N-m)} \tag{14}$$

Where F is the test statistic. m-1 is the between-group variance degrees of freedom, N-m is the within-group variance degrees of freedom. $SS_A$, $SS_E$ are the within-group sum of squared deviations and between-group sum of squared deviations, respectively, and Eq. 15, Eq. 16 are their calculation formulas.

$$SS_A = \sum_{i=1}^{m}(\bar{x}_i - \bar{x})^2 \tag{15}$$

$$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i) \tag{16}$$

In the above equation, $x_{ij}$ is the result of any one test, and $\bar{x}_i$ is the total mean of the test results.

First, ANOVA requires that the quantitative variables need to satisfy normality. Figure 6 shows the results of the analysis of normality for the price level. The graph basically shows a bell shape with a high middle and low ends, which is basically acceptable as a normal distribution.

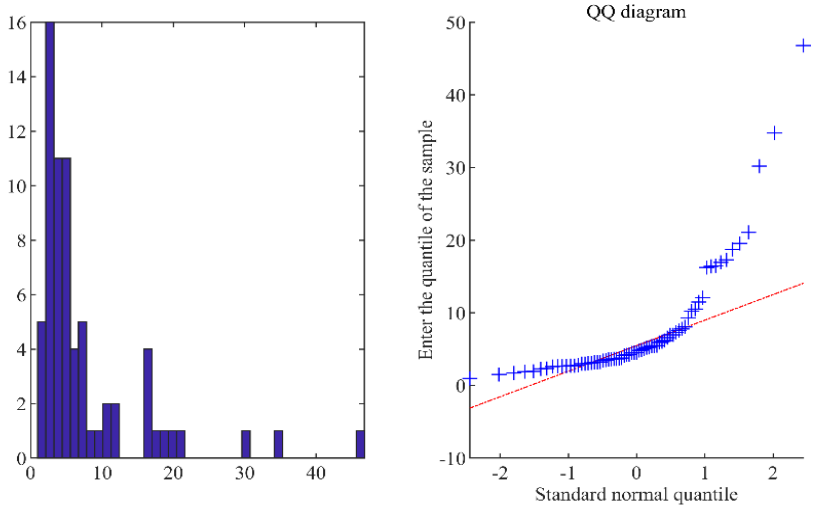


**Figure 6:** Normality test chart

In addition, variance chi-square is a prerequisite for ANOVA. The paper used SPSS to perform a chi-square test on the data. The results are shown in Table 2. Observe the P-value, P>0.05, indicating the chi-square.

**Table 2:** Chi-square test

| Region(standard deviation) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Oceanis 45 | Oceanis 50 | Cyclades 39.3 | Oceanis 40 | Oceanis 46 | Sun Odyssey 54 DS455 | 50 | Sun Odysey 39%i | Sun Odyssey 50DSo |
| 0.12 | 0.44 | 0.36 | 0.17 | 0.25 | 0.37 | 0.16 | 0.55 | 0.39 |

Next, SPSS was used to perform ANOVA on the two variables. Table 3 shows the obtained results. In Table 5, p<0.0.5, indicates significant and rejects the original hypothesis that there is a significant difference between the two data sets.

**Table 3:** Analysis of variance results

| Variable Value | Sample size | Average value | Standard deviation | F | p |
|---|---|---|---|---|---|
| Croatia | 437 | 12.617 | 5.726 | | |
| Italy | 312 | 6.984 | 4.495 | | |
| Sweden | 11 | 6.257 | 2.932 | | |
| Germany | 46 | 5.237 | 2.196 | | |
| British Virgin Islands | 59 | 4.322 | 2.402 | 25.91 | 0.000*** |
| Virginia | 21 | 4.571 | 1.752 | | |
| Antigua and Barbuda | 11 | 11.213 | 7.167 | | |
| Greece | 291 | 7.064 | 4.518 | | |
| Turkey | 71 | 22.312 | 11.599 | | |

Because there are too many data, only some of them are listed here.

***,**,*represent the significance levels of 1%,5%,10%, respectively.

Finally, the paper investigate the magnitude of the differences using effect size indicators. Table 4 shows the effect quantification analysis table. Table 6 shows that the value of $\eta^2$ is 0.444,indicating that 44.4% of the data are derived from the differences between groups. the Cohen's f value is 0.894, indicating that the degree of variation in the quantification of effects of the data is a large degree of variation. That is, it indicates that the region affects the markup.

**Table 4:** Quantitative Analysis of Effectiveness Table

| Analysis items | Difference between groups | Total deviation | Partial Eta side(n2) | Cohens f value |
|---|---|---|---|---|
| Real Price Level | 47833.681 | 107719.49 | 0.444 | 0.894 |

### 3.4. Analysis of regional effects for all sailing variants

The statistical analysis shows that there are 79 different regions and more than 400 different variants. To investigate whether the regional effects of different sailing variants are consistent, the paper drew the price level change curves for the different types with different regions as the horizontal axis and the price level as the vertical axis. The results are shown in Figure 7. Because there are many variants of sailboat b and many different regions, it is difficult to express them in one graph, so only some of the comparison results are drawn here. The paper selected regions with more common types of sailboats for d comparison. Table 5 shows the final selection of 5 different regions and 10 different variants of sailboats.

**Table 5:** Filtered sailboat areas and sailboat variants

| | Country/Region/State | Variant | Variant |
|---|---|---|---|
| 1 | Italy | Oceanis 45 | Sun Odyssey 54 DS |
| 2 | Croatia | Oceanis 50 | 455 |
| 3 | France | Cyclades 39.3 | 50 |
| 4 | Spain | Oceanis 40 | Sun Odyssey 39i |
| 5 | Greece | Oceanis 46 | Sun Odyssey 50 DSO |

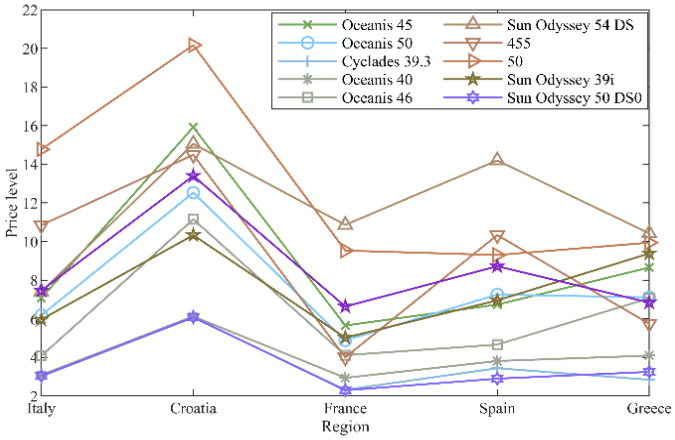


**Figure 7:** Comparison of the regional effects of different variants of sailboats

As shown in Figure7, it is very obvious: the regional effect is largely consistent for all sailboat variants.

### 3.5. Practical and statistical significance of regional effects based on multiple regression

#### 3.5.1 Data pre-processing

By counting the data of monohull and catamaran in the annex, the paper found that there are 79 different regions. Since the price levels of different variants of vessels in the same region are different, the paper choose the average price levels of different variants of vessels in the same region as a representative value of the price levels in that region.

The paper chose a total of six regional factors, namely GDP per capita, the longitude of the regional center, the latitude of the regional center, temperature, wind speed, and territorial sea area, as independent variables in the multiple regression, and price level as the dependent variable to establish a multiple regression equation models to study the practical and statistical significance of regional impacts.Since the units of the six independent variables are different, the paper first standardized the independent variables according to the formula17.

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{17}$$

#### 3.5.2 Validation of a Multiple Linear Regression-based Impact Factor Model

Before conducting the multiple regression analysis, it is necessary to investigate whether the correlation between the independent and dependent variables is significant. The paper applied Pearson correlation coefficient to analyze the correlation of these seven continuous variables. Figure 8 shows the results of correlation analysis.

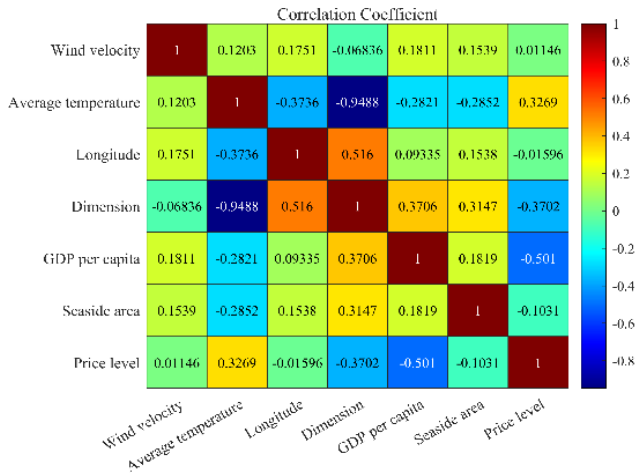


**Figure 8:** Variable correlation analysis heat map

From Figure 8, it can be seen that the relationship between the four independent variables, mean temperature, latitude, GDP per capita, and territorial sea area, and the dependent variable, price level, is linear and significant. Wind speed, mean temperature and price level are positively correlated. Longitude, latitude, GDP per capita, and territorial sea area are negatively correlated with price level. Since the correlation coefficient of latitude and mean temperature is 0.9488, which is highly correlated, it is sufficient to choose one of them. Finally, it is determined that the independent variables are wind speed, longitude, latitude, GDP per capita, and territorial sea area, and the dependent variable is price level for multiple regression analysis.

### 3.5.3 Regression analysis and solution of the dependent and independent variables

The paper used a forward stepwise regression approach for the analysis. Modeling the basic regression equation[8], equation 18.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \mu \tag{18}$$

In the formula 18, $x_1 x_2, x_3, x_4, x_5$ represent the independent variables wind speed (km/h), longitude, latitude, GDP per capita, and territorial sea area, respectively.$\hat{y}$represents the dependent variable. $\mu$ represents the uncertainty factor.[9] The data are brought in and a regression model is built, and the results are shown in Table 6.

**Table 6:** Multiple linear regression model results

| | Coefficient | F-test | Average relative error | VIF |
|---|---|---|---|---|
| $\beta_1$ | 19.841 | | | |
| $\beta_2$ | 2.568 | | | |
| $\beta_3$ | 20.006 | 265.636 | | |
| $\beta_4$ | -67.185 | | 0.1593 | 1.6 |
| $\beta_5$ | -39.327 | | | |

Next, residual analysis and multicollinearity tests were performed. QQ plots can be used to test whether the data obey a normal distribution. Figure 9 shows the Q-Q plot of the residuals.

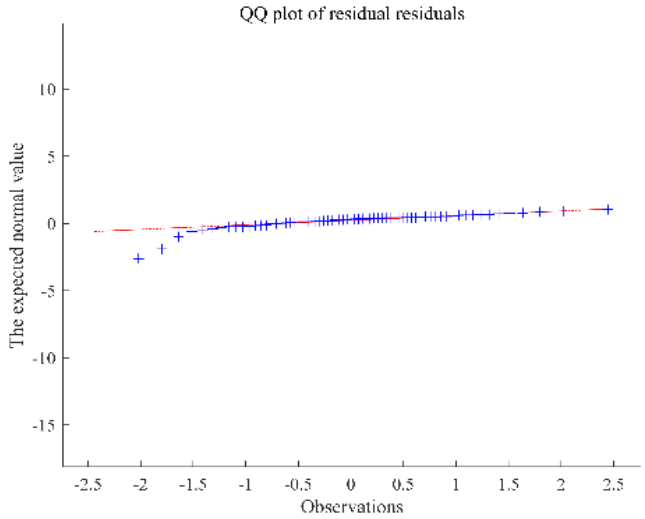


**Figure 9:** Residual Q-Q plot

Finally, regression analysis was performed by adding independent variables one by one using stepwise forward regression[10]. Table 7shows the results of the multiple regression.

**Table7:** Stepwise forward regression analysis of variance table

| Models | | Square and | Degree of freedom | Mean Square | F | Sig |
|---|---|---|---|---|---|---|
| 1 | Regression | 561.702 | 1 | 561.7 | 134.386 | .00b |
| | Residuals | 5434.205 | 363 | 39.66 | | |
| | Total | 5955.907 | 364 | | | |
| 2 | Regression | 1091.162 | 2 | 545.36 | 103.783 | .00c |
| | Residuals | 4864.745 | 362 | 36.88 | | |
| | Total | 5955.907 | 364 | | | |
| 3 | Regression | 1323.438 | 3 | 536.54 | 73.362 | 00d |
| | Residuals | 4632.469 | 361 | 35.62 | | |
| | Total | 5955.907 | 364 | | | |
| 4 | Regression | 1537.289 | 4 | 524.69 | 58.523 | 00e |
| | Residuals | 4418.618 | 360 | 33.57 | | |
| | Total | 5955.907 | 364 | 518.32 | | |
| 5 | Regression | 1836.423 | 5 | 32.36 | 48.779 | 00f |
| | Residuals | 4119.484 | 359 | 507.89 | | |
| | Total | 5955.907 | 364 | 30.57 | | |

From the ANOVA table in Table 7, the residual sum of squares decreases to 4119.484 when all variables are added. the significance test F-statistic is 48.779, corresponding to all P-values of 0, indicating that the original hypothesis is not valid and it is reasonable to consider the model as a linear model.

The model with the best fitting effect is selected. From Table 7, it can be seen that the model fits best when a total of five independent variables, wind speed (km/h), longitude, latitude, GDP per capita, and territorial sea area, are added. Figure 10 shows the comparison between the predicted and true values of the multiple regression model. As can be seen from the figure, $R^2$ is 0.93, and the fitting effect is very good.

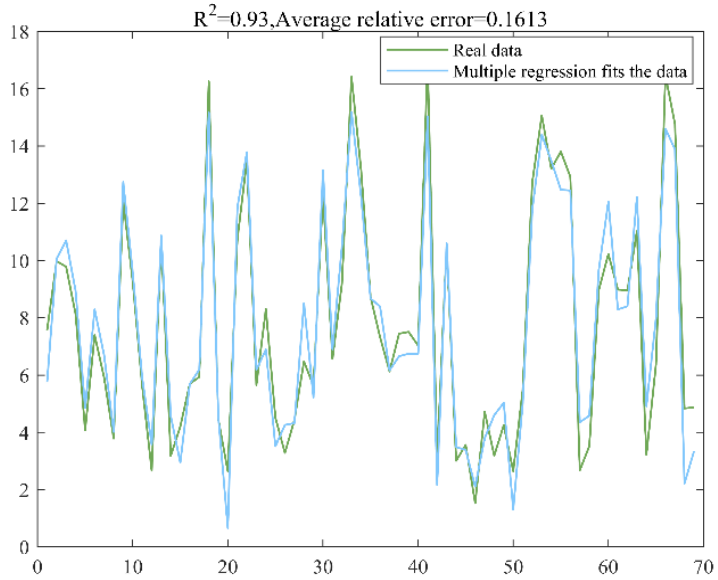


**Figure 10:** Regression model fitting effect

## 4. Conclusions

The paper investigates sailboat pricing, considering factors such as boat characteristics and regional effects. Using a GA-BP neural network model, the study achieves a goodness-of-fit of 0.87, enabling accurate price estimation for different sailboat varieties. It highlights the significant impact

of region on price markup through variance testing, with sailboats in Hong Kong generally selling at higher prices. The findings suggest potential for increased sales through appropriate price reductions to stimulate consumer demand. The paper proposes further model improvement, suggesting multiple linear regression repetitions to discover a better-fitted equation for sailboat price prediction.

With the above analysis, the paper offers advice to help brokers understand market trends and buyer preferences in the sailboat industry. Firstly, brokers are advised to pay attention to price differences between sailboat variants, as monohulls tend to fetch higher prices than catamarans. Secondly, focusing on major brands that dominate the market can lead to more sales opportunities. Finally, brokers should closely monitor market trends and buyer demand, considering that sailboat prices tend to rise with updated manufacturing years. In-depth market analysis, understanding customer needs, offering personalized services, and collaborating with other sailing industry companies can facilitate the growth of the sailboat market.

## References

[1] DU Ke, JIAO Fangfang. Influencing factors in the evaluation of second-hand ship price market method [J]. Water Transport Management, 2020, 42(11): 1-4+7.

[2] LIU Chang. Research on used car value evaluation based on machine learning [D]. Chongqing University, 2021.

[3] LI Xuelei. Construction and application of used car value evaluation model based on BP neural network [D]. Chongqing University of Technology, 2020.

[4] Xu Yuanlei. Research on used car value evaluation based on BP neural network model optimized by particle swarm algorithm [D]. Hebei University of Economics and Business, 2021.

[5] YU Zitong. BP neural network stock selection model based on genetic algorithm optimization [D]. Dongbei University of Finance and Economics, 2022.

[6] Ding, S., Su, C., & Yu, J. (2011). An optimizing BP neural network algorithm based on genetic algorithm. Artificial intelligence review, 36, 153-162.

[7] Wang, S., Zhang, N., Wu, L., & Wang, Y. (2016). Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method. Renewable Energy, 94, 629-636.

[8] Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences, 106, 234-240.

[9] Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics, 99-114.

[10] Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural computation, 16(12), 2639-2664.