

Used sailboats pricing – sail the profit up

Summary

As economic globalization continues to develop, the second-hand ship industry has broad market prospects. To explore the intrinsic factors affecting sailboat prices, we have established three models based on big data: **a static price prediction model, a regional impact model, and a dynamic fluctuation model.**

Firstly, we clean the original data to accurately quantify the factors affecting prices and define eight indicators with the most significant impact on price: *length, year, beam, draft, the proportion of water area, per capita GDP, tariff rate, and length of the coastline*. Then, we expand the data and classify the data types accordingly.

For Model I, we creatively combine principal component analysis and stepwise linear regression to build the regression equation. We find that some independent variables have strong coupling in the correlation analysis of the original independent variables. To avoid the impact on the fitting goodness of the regression equation, we extract **principal components** as new variables for regression. Meanwhile, we use **stepwise regression** to balance precision and significance tests. For both monohull and catamaran sailboats, we obtain regression equations for each original independent variable and price. Our research shows that for monohull sailboats, *length, beam, and draft* are positively correlated with price, with the *beam* having the greatest influence. *The proportion of water area* has no significant impact on price, while *per capita GDP* has the opposite effect and shows obvious fluctuations. For catamaran sailboats, the impact strength of the hull parameters is greater.

For Model II, we emphasize the strong significance of the regional impact on prices through intuitive data graphs and SPSS statistical results. We use **bivariate analysis** of variance to demonstrate that changes in regions have significant differences in the impact on different types of boats. We also apply the model to a relevant dataset in Hong Kong and find that the predicted prices are very close to the actual prices in the Hong Kong market. We also find that ship prices in Hong Kong are generally lower than in other regions, and the decline in prices for individual ships is more severe. We analyze the practical significance of the results and speculate that this may be caused by ship industry transfer and tariff preferential policies.

Model III is an extension of Model I. We find an **interesting phenomenon**: for sailboats in different initial price ranges, the relationship between *year* and price may be positively correlated or opposite. We also find that the *year* has a good explanatory effect on the fluctuation of *per capita GDP*. Therefore, we combine **autoregressive modeling** and Model I to establish a dynamic model of price over time and predict the optimal selling price in the Hong Kong region for the next five years. Based on this, we also creatively explore the mechanism of the COVID-19 epidemic on sailboat prices.

Finally, we conduct a sensitivity analysis on the core model, Model I. We find that the stability of catamarans is slightly inferior to that of monohulls, but the stability of both is at a high level, demonstrating the reliability of our model.

Keywords: principal component analysis; time series; stepwise regression

Contents

1 Introduction	3
1.1 Problem Background	3
1.2 Restatement of the Problem	3
1.3 Literature Review.....	3
1.4 Our Work.....	4
2 Assumptions and Justifications.....	5
3 Notations	5
4 Model I: PCSL regression model of price.....	6
4.1 Data Description	6
4.2 The Establishment of PCSL regression model.....	9
4.3 The Solution of PCSL regression model.....	11
5 Model II: Regional impact model.....	15
5.1 Regional impact analysis	15
5.2 Practical application of the model: Hong Kong.....	17
6 Model III: Sailboat Dynamic Pricing Model.....	19
6.1 The establishment of the model	20
6.2 Prediction results and analysis	20
7 Sensitivity Analysis.....	21
8 Model Evaluation	21
8.1 Strengths	21
8.2 Possible improvement.....	22
9 Conclusion.....	22
9.1 Basic conclusion	22
9.2 Further conclusion	23
References	23

1 Introduction

"Mathematics is the key to unlocking the secrets of big data."

- Dr. Hannah Fry

Associate Professor in the Mathematics of Cities at University College London






1.1 Problem Background

With the continuous development of economic globalization, the second-hand sailboat market has been growing in size and has broad commercial development prospects. To quickly seize the wealth of this market, it is necessary to accurately grasp the trading price of second-hand sailboats. The price fluctuation of second-hand sailboats is formed by the comprehensive influence of various factors, including macro market conditions and micro ship performance parameters^[1]. At the same time, the impact of regional economic development on the final pricing is also significant.

Based on big data, it is of great significance to establish a universal pricing model for sailboats considering the influence of multiple factors, and then apply it to specific regional practices to grasp the dynamic economic benefits of the sea in that area.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

-  Develop a mathematical model that explains the listing price of each of the sailboats in the provided spreadsheet.
-  Using the established model to explain the impact of geographic regions on the pricing of different types of sailboats.
-  Apply the model to the second-hand sailboat market in Hong Kong and elaborate on the pricing impact of different types of sailboats in the region.
-  Explore more interesting and meaningful conclusions from big data as much as possible.
-  Based on the model's results, prepare a one to two-page report for the Hong Kong (SAR) sailboat broker.

1.3 Literature Review

The pricing of second-hand sailboats falls within the scope of shipping market economics. This area has received extensive attention and considerable research since the last century. Pruyn et al^[2] reveal the trend of research directions in pricing models for second-hand sailboats in the past 20 years, which has shifted from macro-level influencing factors to micro-level influencing factors.

Regarding macro-level factors, the international research focus has been on the impact of freight rates on vessel pricing. Hawdon and Tsolakis have respectively established static^[3] and dynamic^[4] models for pricing and freight rates and pricing fluctuations and freight rates. Adland's research^[5] has further expanded the scope of macro-level factors to new-building

prices, which improves the universality and accuracy of the models.

Regarding micro-level factors, Koehn et al have used semi-parametric methods^[6] to reveal the relationship between many micro variables, such as DWT, years of use, and vessel size, and prices. The research results indicate that micro variables also have a significant impact on the pricing of vessels.

The transparency of transaction data and the availability of technical data provide new methods and ideas for using big data to analyze the influencing factors of second-hand vessel prices. However, research on micro-level factors, which started relatively late, has not yet been well resolved in terms of the global applicability of models and coupling effects among various influencing factors. At the same time, it also lacks practical verification and application in specific related regions.

This article hopes to provide a more optimized and specific solution to the pricing of second-hand sailboats by using the richness of data and combining the research results of predecessors.

1.4 Our Work

We have established three main models for the pricing of secondhand sailboats. Our core work involves the development of the PCSL model for regression analysis. We extracted eight indicators as the main influencing variables for price and conducted principal component analysis (PCA) for both catamarans and monohulls. The principal components were then used as new independent variables for stepwise regression, and the resulting equation was transformed back to the original variables.

To address the impact of regions on prices, we developed a regional impact model to demonstrate the significance of regional influence. We conducted data experiments to further demonstrate that the impact of regional changes on different types of boats varies, and we provided possible explanations for these differences.

Additionally, we expanded upon the conclusions derived from the PCSL model and used the AR autoregressive method to explore the impact of "year" on prices, ultimately establishing a dynamic pricing model.

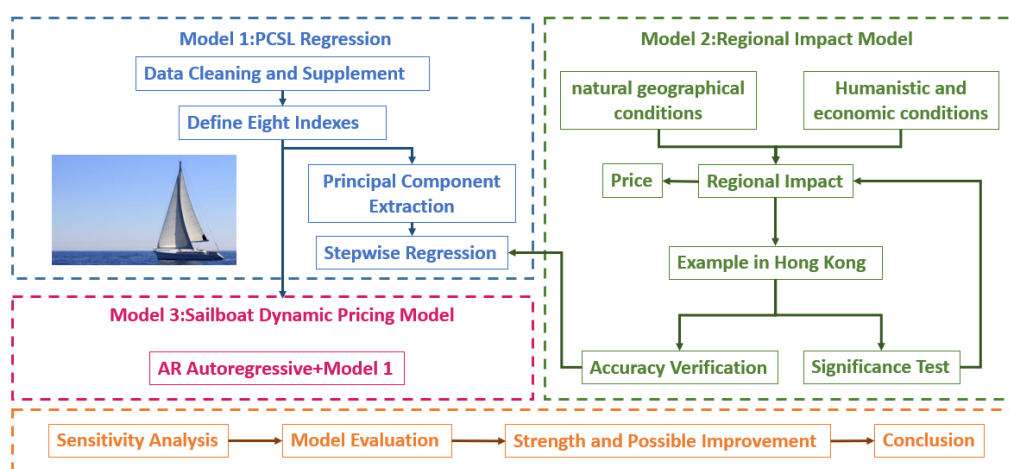


Figure 1: The structure of our work

2 Assumptions and Justifications

To simplify the problem, we make the following basic assumptions, each of which is properly justified.

Assumption 1: We assume that the price data we obtained represents the optimal price under the given conditions.

Our modeling mechanism is based on big data modeling, and the accuracy of price data directly affects the accuracy of our modeling results. Only when the original pricing data is scientific can we provide suitable selling prices to ensure profitability for the shipbuilding companies.

Assumption 2: We assume that the international trade environment is relatively stable, and the tariff rate index value is relatively stable.

For commodities such as ships with frequent imports and exports, the impact of tariff rates on prices is significant. To accurately quantify this index, we need to assume that the tariff rate fluctuations between regions are not significant within the time span under study and can be taken as a constant reference value by averaging the values over the past few years.

Assumption 3: We analyze the fluctuations of prices over time on an annual basis and ignore seasonal effects.

Seasonal changes have a relatively small impact on ship prices and have a complex fluctuation trend. In order to reduce the difficulty of obtaining data and simplify the dataset, we obtained the data for used sailboats over the years.

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description
X_i	Independent variables that affect prices
Y	Price of second-hand sailboats
X_i	Normalized independent variable
λ_{ij}	Pearson correlation coefficient
a_i, c_i	Fitting undetermined parameters
α, β	Characterizing the effect of an independent variable on the dependent variable
γ	Representing the cross-influence effect of multiple independent variables on dependent variables
ε	random error

4 Model I: PCSL regression model of price

The price of second-hand sailboats is influenced by various factors. To address the issue of having a large number of independent variables and the need to establish a specific mathematical relationship, we use a combination of principal component analysis and stepwise regression. The former is used to reduce the dimensionality of the many correlated indicators, while the latter involves iterative testing and optimization during the regression process to obtain a more accurate fitting equation.

4.1 Data Description

Our pricing model is based on the accuracy and richness of data. Therefore, data cleaning and enhancement are crucial for establishing the model.

4.1.1 Data cleaning

For the raw data, we conducted data cleaning work as follows:

(1) Outlier handling

We observe the distribution of the original data points and used the classical indicators of mean and variance to remove data outliers and smooth the dataset. After processing, most of the original data is retained.

(2) Missing value handling

For the few missing values in the original data, we conduct screening and processing. We believe that considering all manufacturers and sailboat variants is not appropriate for building a more universally applicable model because the data volume for some manufacturers and sailboat variants is insufficient to provide reliable results. Therefore, we finally select the top 6 manufacturers and top 22 sailboat variants by frequency of appearance as the objects of big data analysis for monohulls/catamarans. We defined them as typical analysis objects.

We also perform the same analysis for catamarans, but due to space limitations, we are only showing the results for monohulls here:

Table 2: Typical analysis objects for monohulls

Typical analysis objects	Name list
Makers	Bavaria, Beneteau, Catalina, Hanse, Island Packet, Jeanneau
Sailboat variants	53, 400, 415, 445, 461, 495, 505, 540e, Cruiser 46, Oceanis 40, Oceanis 41, Oceanis 41 Owners Version, Oceanis 43, Oceanis 45, Oceanis 46, Oceanis 48, Sun Odyssey 39i, Sun Odyssey 409, Sun Odyssey 55, Sun Odyssey 469, Sun Odyssey 50DS, Sun Odyssey 509

4.1.2 Data supplement

Lun et al conducted a comprehensive and detailed discussion on the factors that affect the pricing of second-hand sailboats. Referring to their research, we divide the factors that affect

pricing into two main levels: macro factors and micro factors. Macro-level indicators generally focus on overall objective price factors such as global maritime trade volume, cross-regional tariff rates, regional economic development levels, and freight market prices. Micro-level indicators, on the other hand, focus more on the characteristics of the boat itself, such as production year, draft, and hull size. Based on the descriptions in *Shipping and Logistics Management* [7], we have selected the following eight indicators as the core pricing factors for second-hand sailboats.

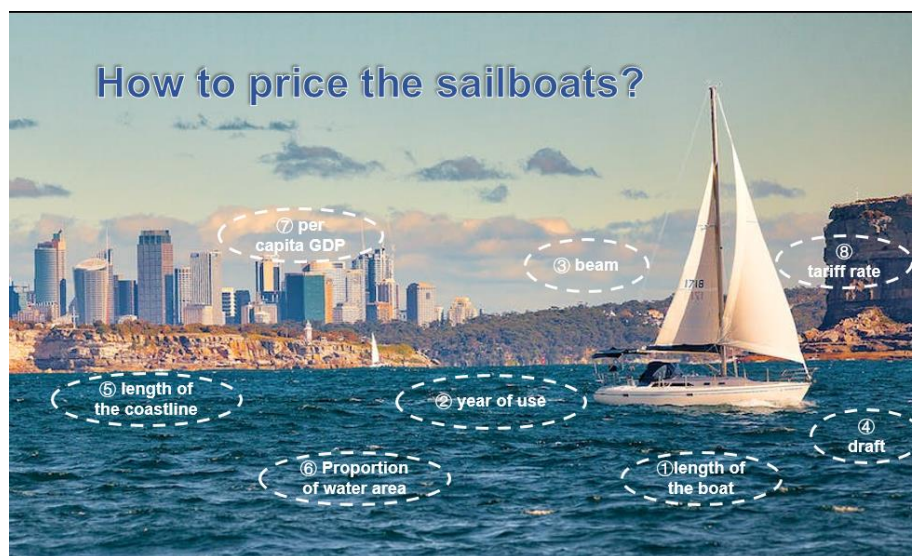


Figure 2: Defining regression independent variables

There are many factors that affect sailboat pricing, so it is difficult to accurately represent them with limited data. Therefore, it is necessary to expand the original data.

(1) micro-level factor: *Beam and Length*

The size of a vessel largely determines its cost. Beam and Length together determine the vessel's lateral and longitudinal distances. Therefore, for the selected variants of sailboats from typical manufacturers, we extracted their Beam and Length from the related website and compiled them into a table.

(2) macro-level factor: *Proportion of water area and Length of coastline*

The second-hand sailboat market strongly depends on the natural geographical environment. We believe that macro factors brought by geographical regions should occupy a considerable weight in sailboat pricing.

According to the supply and demand relationship theory in economics, we believe that the proportion of water area has a close relationship with sailboat pricing. In arid areas, the narrow application background of sailboats and insufficient purchasing power drive up prices. In humid areas, prices often decrease due to increased demand and mature production technology.

At the same time, in regions where sea transportation is more developed, the length of the coastline is also an important factor affecting the supply and demand relationship of sailboats. It is not difficult to imagine that more trading space stimulates greater demand for transportation tools.

To accurately extract the length of the coastline of typical analysis areas, we used the **Google Earth Engine** platform to perform edge extraction on remote sensing images. At the

same time, using the NDWI (Normalized Difference Water Index) value to measure the water area of the area of interest further enriched the macro natural geographical data.

(3) macro-level factor: *per capita GDP*

Sailing boats are a luxury item, even second-hand ones are priced at around 200-300 US dollars per ton, which is not affordable for the ordinary income class. Considering this factor, the economic development status of the region naturally becomes an important factor in pricing. The sailing boat market targets the middle and high-income class, and their development status and purchasing power largely determine the buying and selling prices. For simplicity, we collected the per capita GDP data of typical regions in 2020 to measure the economic development status of the region, reflecting the purchasing power and purchasing willingness.

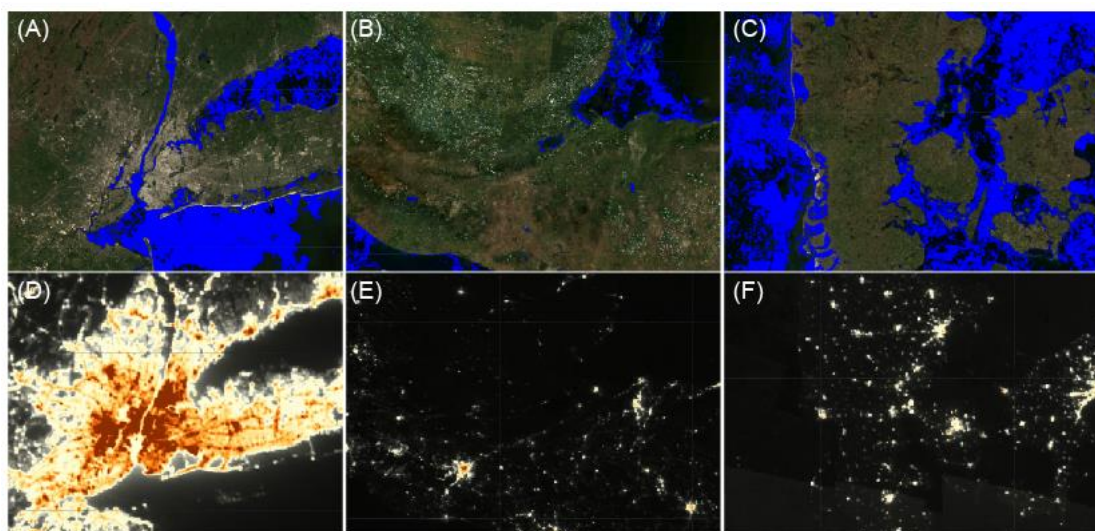


Figure 3: A, B, C represents GEE remote sensing images of New York, Panama, and the Netherlands. The blue color in the image represents the rendered water area. D, E, F represent their night light remote sensing images, reflecting their economic development status.

(4) macro-level factor: *tariff rate*

Sea transportation generally involves the flow of goods across national borders. As mentioned earlier, the freight rate market often plays a crucial role in determining the pricing of sailboats. In the increasingly complex international trade environment, transport tariff rates hold the key to the freight rate market. Additionally, due to the developed shipbuilding industry in Europe, the sailboats produced there are often purchased as imported goods by other regions, and the introduction of tariffs adds to their added value. We have also obtained tariff rate data for typical regions to quantify the pricing impact of this macro factor.

4.1.3 Data collection

The supplementary data we used mainly include specific performance parameter data of ships from different variants and manufacturers, such as draft and length, trade data (including tariffs) between different countries, and waterbody geographic information and economic development status of various typical regions extracted based on the GEE platform. The data sources are summarized in Table 4.

Table 3: Data source collation

Source Names	Database Websites/Platforms
Beam and Length	https://www.sailboatlistings.com/ https://www.yachtworld.com/ https://www.boats.com/
Length of coastline	Google earth engine
Proportion of water area	Google earth engine
Tariff rate	https://www.usitc.gov/
Per capita GDP	https://www.census.gov/ https://www.worldbank.org/
Hong kong related data	https://www.oceanway.org/
COVID-19 related data	wx.wind.com.cn
Pictures used in the paper	Google

4.2 The Establishment of PCSL regression model

4.2.1 Definition of independent variable

Based on the previous analysis, we define the main indicators that affect pricing as shown in the following table.

Table 4 Definition of independent variable

Independent variables	length	year	beam	draft	LOC	POWA	pc GDP	tariff
Symbols	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8

Where “LOC” refers to “Length of coastline”, and “POWA” refers to “Proportion of water area”.

We also define the unique dependent variable the price as : Y .

To facilitate regression analysis, it is necessary to normalize each independent variable in the dataset, namely:

$$X_i = \frac{X_i - \mu}{\sigma} \quad (1)$$

Where μ and σ are the mean and variance of the corresponding dataset.

4.2.2 Analysis of correlation between independent variables

Prior to conducting regression analysis, we consider the possibility of strong multicollinearity among the independent variables, which could affect the final regression results. Therefore, we first use the Pearson coefficient to analyze the correlations among the eight independent variables, to determine an appropriate regression method.

$$\lambda_{ij} = \frac{\sum_{i=1}^n (X_i - \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_i)^2} \sqrt{\sum_{i=1}^n (X_j - \bar{X}_j)^2}} \in [-1, 1] \quad (2)$$

In equation 2, \bar{X}_i is the mean value of normalized indicators.

Based on this formula, we can construct a correlation coefficient matrix to further analyze the autocorrelation coupling between the independent variables.

4.2.3 Principal component analysis

Since the selected indicators may inevitably have autocorrelation, directly using the original independent variables for regression analysis can lead to an increase in the confidence interval of regression coefficients, and a decrease in precision and significance level. Therefore, we use principal component regression to reduce the dimensionality of the original independent variables.

Defining the covariance matrix, we have:

$$C = \begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \vdots \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \vdots \\ \dots & \dots & \text{cov}(X_n, X_n) \end{bmatrix} \quad (3)$$

In equation 3, cov is the covariance coefficient that characterizes the correlation.

For solving the characteristic equation, we have:

$$(C - \lambda E)\Phi = 0 \quad (4)$$

The eigenvectors ϕ_i obtained from the formula are the corresponding principal components.

We calculate the contribution rate of each principal component to the variance and select the top 4 principal components with a cumulative contribution rate of 80%.

4.2.4 Stepwise linear regression

Based on the modeling process above, we use the 4 identified principal components as new independent variables along with the original dependent variable: price, to establish a multiple linear regression model.

$$\begin{cases} Y = a_0 + \sum_{j=1}^4 a_j \phi_j + u \\ u \sim N(0, \sigma^2) \end{cases} \quad (5)$$

In equation 5, $a_j, j=1, 2, \dots, 4$ are the regression fitting coefficient, u is the assumed random error term based on the Gaussian-Markov hypothesis.

To further improve the accuracy of regression fitting, we adopt the idea of stepwise regression. That is, we introduce variables into the model one by one, and after introducing each variable, we test it and calculate its P-value, which reflects the significance level of each independent variable's impact on the dependent variable. when the P-value of a certain independent variable is less than 0.05, we consider it to have a significant impact on the dependent variable.

After obtaining the fitting equation of the dependent variable with the main component independent variables, we then transform the main component independent variables back to the original independent variables to obtain the fitting equation of the original independent variables and the pricing.

The overall modeling process can be represented by the following flowchart.

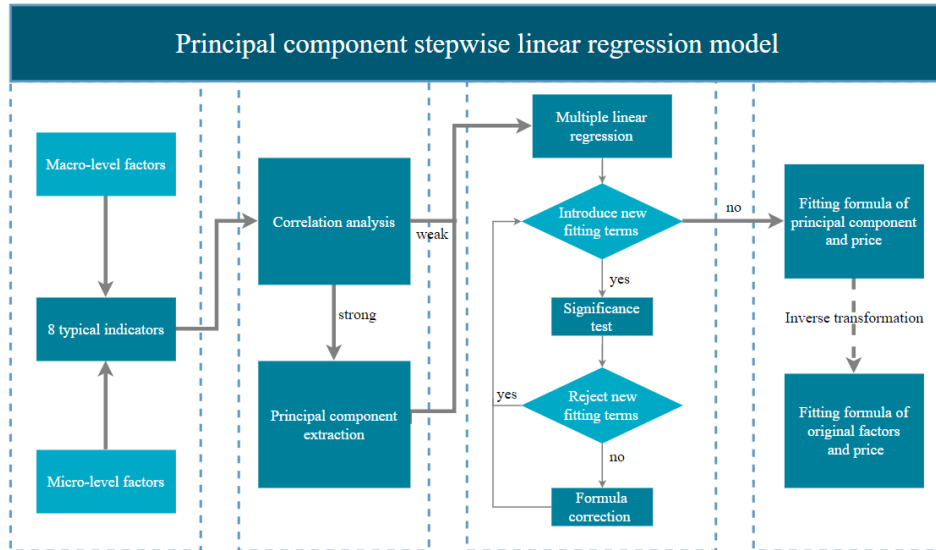


Figure 4: Flowchart of PCSL regeneration model

4.3 The Solution of PCSL regression model

4.3.1 Correlation Analysis and Principal Component Extraction

We conduct correlation analysis and principal component analysis on the 8 independent variables, and the results are as follows.

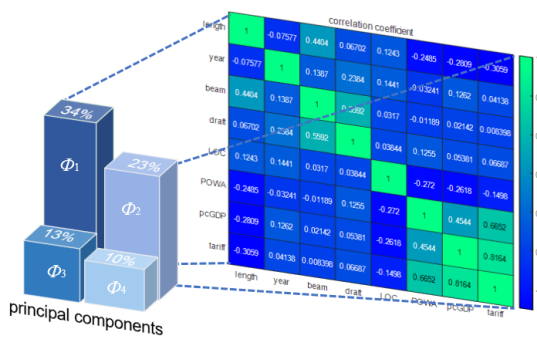


Figure 5: Results of monohull boat

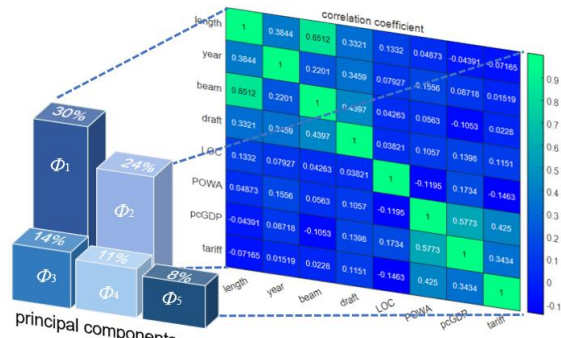


Figure 6: Results of catamarans

From the results, for monohull boats, there is a strong positive correlation between tariff rates and per capita GDP. This is not surprising, as countries with higher per capita GDP tend to have more developed foreign trade. We also found that boat manufacturers are mainly concentrated in Europe, while buyers often come from developed countries such as the United States. Trade exchanges between these two regions are frequent, and conflicts and frictions are correspondingly higher, so tariff rates are often higher as well.

For catamarans, there is also a strong correlation between beam and length. We believe this is reasonable because catamarans have a larger hull, and generally have higher structural requirements. The overall width and length of the boat need to be better matched to maintain structural stability and safety.

In summary, there are strong correlations between some of the independent variables, both for monohull and catamaran boats, so it is necessary to extract principal components.

4.3.2 Regression Results and Analysis

We used Matlab for regression analysis. The core logic and steps of the algorithm are presented below in pseudocode.

Algorithm 1: Stepwise regression

Input: $\{x_{i=1,2,3\dots n}, y\}$

Output: $\{x_{i=1,2,3\dots m}, y\}, b_{i=1,2,3\dots m+1}, R^2, \text{RMSE}$

for $i = 1$ to m **do**

 Calculate the correlation coefficient: $\text{cov}(x_i, y)$

 Calculate R^2 , F -value, P -value of the variable with the largest correlation coefficient.

While R^2 in the previous step > 0.001

 Add the independent variable with the largest current correlation coefficient to the regression model, and delete the value of the independent variable in the original correlation coefficient series

 Calculate the R^2 , F -value, P -value of the modified model

If F -value $>$ given value

 Add the new variable

else reject the new variable

end

end

The final regression equation for the selected 8 independent variables and price is presented below.

(1) Monohull sailboats

After conducting principal component analysis and inverse transformation of the independent variables, we find that the fitting coefficient of the cross-term is quite small for the regression formula of the price of a single-hull vessel. To further simplify the result, we choose a quadratic fit form to construct the regression equation. The result is as follows:

$$\begin{aligned}
 Y = & 25574.3X_1 - 26469.4X_2 + 69319.7X_3 + 19224.9X_5 + 40995.2X_7 \\
 & + 18485.6X_8 + 11713.9X_2^2 + 20276X_3^2 + 6533.5X_4^2 - 9827.7X_5^2 \\
 & - 4176X_6^2 - 19409.5X_7^2 + 15695.9X_8^2 + 179095.7
 \end{aligned} \tag{6}$$

(2) Catamarans

The result for Catamarans is more complicated than that for monohull sailboats. We find

that the coefficient of the cross-term cannot be ignored when we reverse the original independent variables. If we still expect to construct the relationship between the original independent variables and the price, the resulting regression equation will be very long and complicated. Therefore, we directly provide the fitting coefficients of the principal component independent variables and explain their possible practical meanings. The result is as follows:

$$\begin{aligned}
 Y = & 39003.2\phi_1 - 14969.7\phi_2 - 135028\phi_4 - 34946.3\phi_5 + 7323.5\phi_2^2 \\
 & + 13133.6\phi_4^2 - 31825.9\phi_1\phi_4 + 52891\phi_2\phi_3 + 24013\phi_2\phi_5 - 3017.43\phi_3\phi_4 \\
 & - 30476\phi_3\phi_5 + 458170.2
 \end{aligned} \tag{7}$$

(3) Accuracy analysis

We conduct precision analysis and data validation for the regression results. We use two indicators to describe the precision: the coefficient of determination R^2 and root-mean-square error (RMSE), and the results are shown in the table below.

Table 5: Accuracy analysis results

Type of the sailboats	R^2	RMSE
Monohull sailboats	0.814	32300
Catamarans	0.751	83400

- From the perspective of R^2 , we can see that whether for monohull or catamaran, the R^2 values are relatively high, indicating that the model has a good fitting effect.
- From the perspective of RMSE, we can see that the RMSE values for both monohull and catamaran are relatively small compared to their sale prices, indicating that the error of the model is small.

In addition, we also perform cross-validation between real data and predicted data. We randomly select 75% of the original data to establish the regression equation and use the remaining 25% of the data to evaluate the precision of the fitted formula. The comparison and residual plots are shown below.

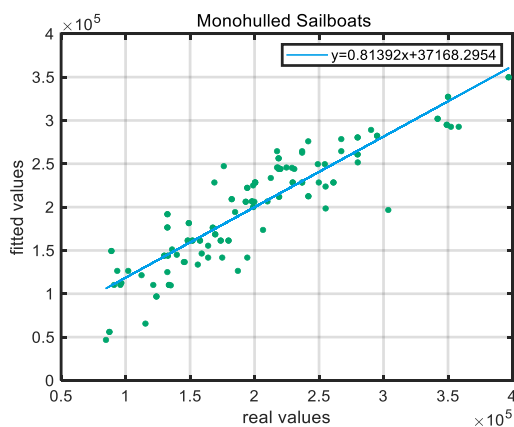


Figure 7: Comparison of fitting effects

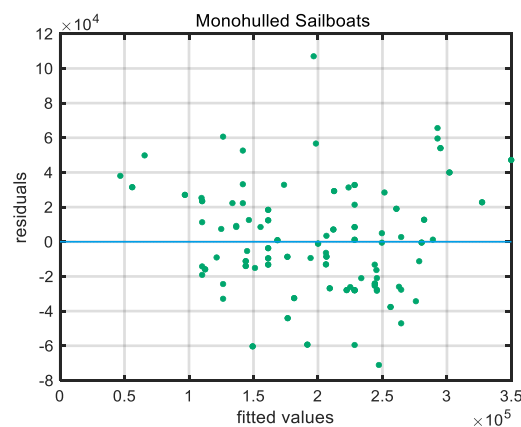


Figure 8: Residual plot

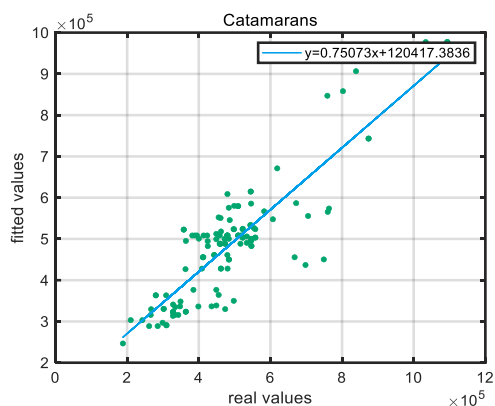


Figure 9: Comparison of fitting effects

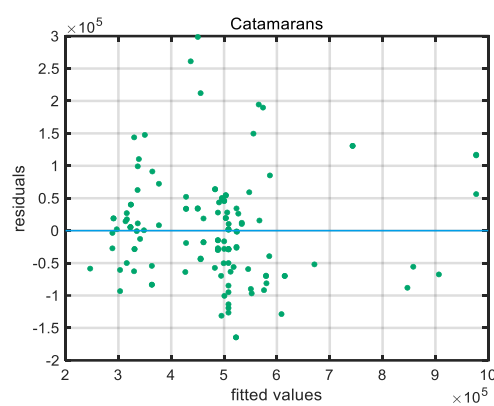


Figure 10: Residual plot

The data results are relatively concentrated, indicating a good fitting effect.

(4) Discussions

We attempt to explore interesting conclusions from the fitting terms and coefficients of each variable.

Regarding **monohull sailboats**, we find that:

- Length, beam, and draft are all positively correlated with pricing, with beam having the greatest impact. Increasing the volume of the hull undoubtedly leads to a higher cost of construction. Draft measures the weight and buoyancy of the hull, and as this value increases, the quality and quantity of materials required for construction also increase. As for the significant impact of beam, generally, the larger the beam value, the stronger the stability of the hull, which is of great positive significance for monohull sailboats that are often used for commercial transportation or recreational activities.
- Water surface area ratio has a small and negative impact on pricing. We think this may be because the impact of water surface area ratio on ship pricing is not clear enough. Because water surface area includes both inland water surface area and coastal water surface area, and inland lakes often account for the majority of water surface area. This gives us the insight that we can further refine the definition of water surface area ratio to optimize this indicator.
- Per capita GDP and tariff rates have a relatively large impact on pricing and exhibit some volatility. We will further discuss the sensitivity of each indicator later.
- In addition to the above basic conclusions, we also find **two interesting conclusions**. The impact of usage time on pricing shows a trend of first increasing and then decreasing. The possible explanation is that as a means of transportation and luxury goods, sailboats have both use value and collection value. For sailboats with lower pricing, use value predominates, so the longer the usage time, and the poorer the performance, the lower the pricing. However, the situation is reversed for sailboats with higher pricing, as the rarity of sailboats increases with time. In addition, the impact of coastline length on pricing also shows a trend of first increasing and then decreasing. We speculate that there is a saturation threshold of promoting the price. In undeveloped areas, even with a longer coastline, sailboat pricing cannot be promoted due to a lack of market demand.

Regarding **catamarans**, starting with the principal components as independent variables,

we find that:

- Out of the five principal components selected, three of them possibly have practical meanings. Component Φ_1 is mainly composed of micro-parameters such as length, beam, and draft, which comprehensively reflect the impact of the catamaran's own parameters on its price. Component Φ_3 is mainly composed of the coastline length and the proportion of water area, reflecting the impact of natural geographical conditions on the price. Similarly, component Φ_2 is composed of per capita GDP and tariff rates, revealing the impact of regional cultural and economic conditions on the price.
- For component Φ_1 , it has a significant positive impact on pricing, and its coefficient of influence is larger than that of a monohull.
- For component Φ_3 , it participates in the generation of multiple interaction terms. We can speculate that natural geographical conditions are highly coupled with other indicators such as GDP and performance requirements of the catamaran, and there are multiple constraint relationships.
- For component Φ_2 , it has a significant fluctuating impact on the price, just like a monohull, further reflecting its non-stationarity.

5 Model II: Regional impact model

5.1 Regional impact analysis

5.1.1 Establishment of variance analysis model

(1) One-way ANOVA

To further investigate the impact of region on the pricing of used sailboats, we directly add region as a new independent variable. Since we are only interested in the effect of region on price, we try to keep other factors constant or with small fluctuations during the data screening process. Therefore, we assume that during the experiment, all factors except for region remain unchanged, and only random errors exist, in order to explore whether the impact of region is significant.

For each price dataset in each region, we establish the following model:

$$\left\{ \begin{array}{l} P_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \sum_{i=1}^r n_i \alpha_i = 0 \\ s.t. \left\{ \begin{array}{l} \varepsilon_{ij} \sim N(0, \sigma^2) \\ i = 1, 2, \dots, r, j = 1, 2, \dots, n \end{array} \right. \end{array} \right. \quad (8)$$

In equation 8, μ is the total mean value of regional data, n_i is the data amount for the region i , P_{ij} is the price estimate, ε_{ij} is a random error term, α_{ij} represents the impact of region i on prices.

We set the confidence level at 95% and use the F -value to determine the significance of the results.

(2) Two-way ANOVA

The regional effect on the interaction between different variants of ships is significant, as different regions may have collective preferences for specific ship variants. Therefore, we also conduct two-factor analysis of variance on the combined effects of ship variants and regions on prices.

Modifying the above model as follows:

$$\left\{ \begin{array}{l} P_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \sum_{j=1}^s \alpha_j = 0, \sum_{r=1}^r \beta_r = 0, \sum_{i=1}^r \gamma_{ij} = \sum_{i=1}^s \gamma_{ij} = 0 \\ s.t \left\{ \begin{array}{l} \varepsilon_{ijk} \sim N(0, \sigma^2) \\ i = 1, 2, \dots, r, j = 1, 2, \dots, s, k = 1, 2, \dots, t \end{array} \right. \end{array} \right. \quad (9)$$

In equation 9, γ_{ij} represents the cross impact effect of region i and ship variants j on prices.

We believe that the two independent variables: region and boat variant are mutually exclusive, meaning there is no interaction effect between them. Therefore, we have

$$\gamma_{ij} = 0 \quad (10)$$

5.1.2 Solution of the module

(1) Results of One-way ANOVA

We first extracted price data of ships from some typical regions for visualization. It can be clearly seen that the impact of region on price is significant, whether it is for catamarans or monohulls.

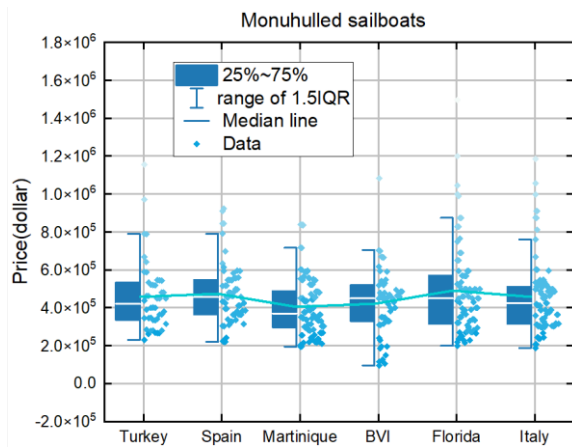


Figure 11: Typical data of Monohull

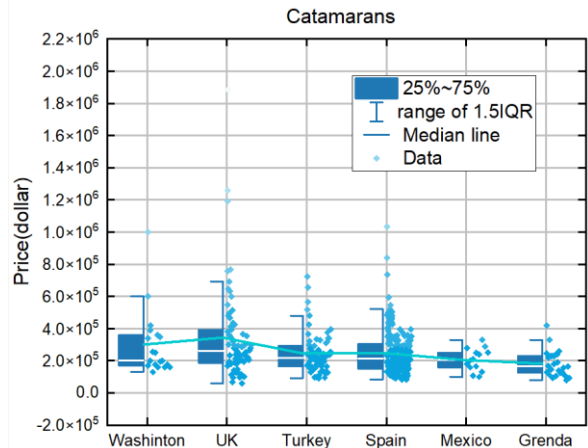


Figure 12: Typical data of catamaran

We conduct one-way ANOVA using SPSS, and the results are shown in the table below:

Table 6: Results of one-way ANOVA

Type of the sailboats	η^2	F-value	P-value
-----------------------	----------	---------	---------

Monohull sailboats	0.061	5.245	0.0013
Catamarans	0.027	2.491	0.0031

Here, η^2 indicates the intensity of impact, F -value and P -value characterize the reliability of the results.

We can see that:

- The result's P -value is very small, indicating that we have full confidence in accepting our conclusion.
- The strength of their impact indicator, eta-squared (η^2), is relatively large, statistically explaining the importance of region for price.
- Monohull prices are more sensitive to region, which may be related to the wider range of applications for monohulls.

(2) Results of Two-way ANOVA

The results are shown in the table.

Table 7: Results of Two-way ANOVA

Type of the sailboats	η^2	F -value	P -value
Monohull sailboats	0.855	22.148	0.0008
Catamarans	0.705	19.921	0.0015

- The effect size measure η^2 significantly increased. However, at this point, it reflects the interaction effect between the region and the boat variant, meaning that the impact of changing regions on the prices of different boat variants is significantly different. The specific relationship can be obtained based on the regression model we previously established.
- We have statistically answered the consistency issue of the region's impact on different boat variants.

5.2 Practical application of the model: Hong Kong

Hong Kong is one of the largest and busiest port cities in the world with highly developed sea transport and trade. Second-hand sailboats have a broad market in this region, and we apply the model to this area to further validate its accuracy while providing practical predictions for sailboat prices.



Figure 13: The landscape of Hong Kong

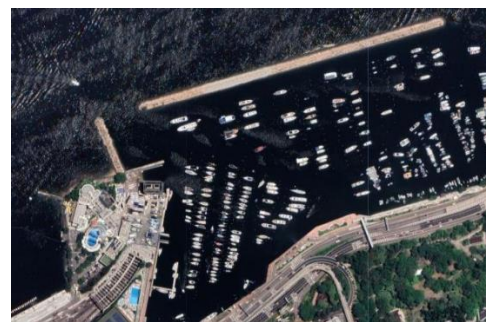


Figure 14: Port of Hong Kong

5.2.1 Price prediction

We use our price model to search for relevant data in the Hong Kong ship market in 2020 and predict the pricing under the influence of 8 independent variables, and then compare it with the actual market selling price.

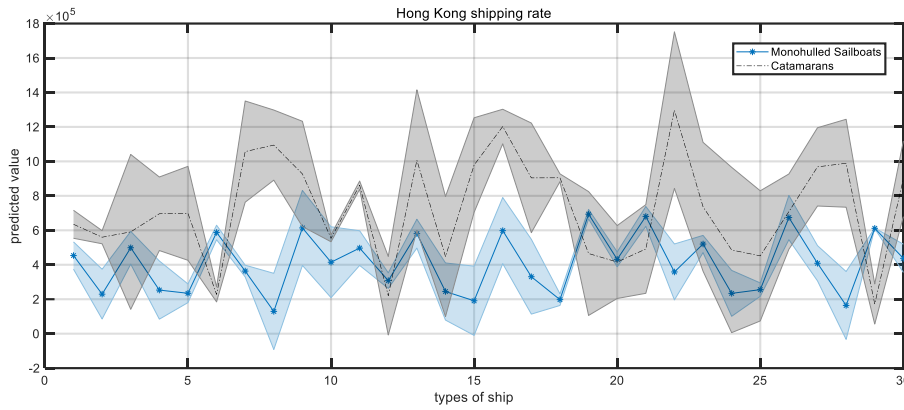


Figure 15: Error band chart of Hong Kong ship price prediction

We find that the error bars of the predicted results are relatively small for both catamarans and monohulls, indicating that our model has strong applicability in the Hong Kong area and achieved satisfactory results.

5.2.2 Analysis of “Hong Kong” Effects

We use Hong Kong as an example to focus on the impact of regional effects on sailboat pricing. We replaced the region-related data in the original dataset, such as coastal length and water area ratio, while keeping other indicators such as length and beam unchanged. In this way, we guarantee the single variable principle and compare pricing in different regions based on the regression model.

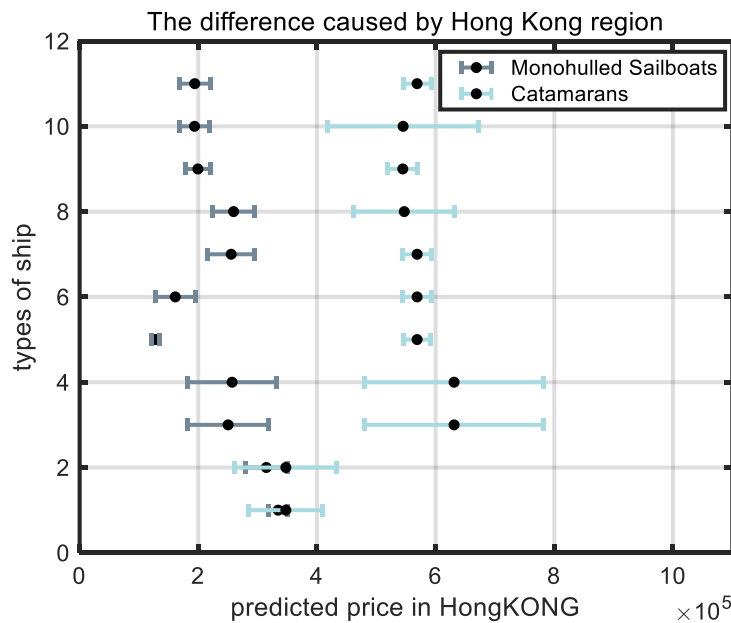


Figure 16: Errors in the impact of the Hong Kong region

We use a T-test for significance analysis.

Table 8: Results of T-test

Type of the sailboats	<i>T</i> -value	<i>P</i> -value
Monohull sailboats	5.548	0.0002
Catamarans	-3.103	0.0003

We conduct a detailed study and discussion on the statistical and practical significance of the results.

(1) Statistical significance

From the test results, there was a certain decrease in selling prices after changing the region to Hong Kong, which further verifies the significant impact of region on pricing. In addition, for monohull sailboats, the selling price decreased more significantly.

(2) Practical significance

We attempt to provide practical explanations for the above results from the perspective of economic principles.

- Hong Kong is one of the world's most important trading ports and economic centers. Countries often implement tax exemption policies for transactions related to the Hong Kong region for their own development considerations. The significant reduction in tariff rates has significantly reduced the selling price cost. We believe this is the main factor for the decrease in selling price.
- As a typical port city, Hong Kong has a large proportion of water area, and there is a high demand for sailboats. However, the shipbuilding market in Hong Kong has been developing for many years and is approaching saturation. In recent years, the transfer of shipbuilding industry has stimulated more ship supply, which has suppressed ship prices.
- Single-hull boats, as a traditional ship type, have a wider range of applications and are more susceptible to the above factors, so the selling price has decreased more significantly.

6 Model III: Sailboat Dynamic Pricing Model

We have discovered some interesting findings in Model I:

The impact of production year on pricing varies depending on the price range. Generally, older sailboats are sold at a lower price. However, for some expensive sailboats, the older the production year, the higher the price.

The promotion effect of coastline length on pricing may have a saturation threshold.

We are particularly interested in the impact of production year on pricing. In fact, Model I only establishes the relationship between various macro and micro factors and sales price and is a static model. Considering the multidimensional impact of time on other factors, it is meaningful to establish a dynamic pricing model for production year. The year affects the age of the ship, and also has a corresponding relationship with the economic development of the region and the adjustment of trade policies of various countries. In addition, establishing a dynamic

model to predict sailboat pricing can also establish a certain advantage for future market competition.

6.1 The establishment of the model

Year as an independent variable has a significant impact on determining prices. However, at the same time, per capita GDP, an important indicator affecting prices, also exhibits significant fluctuations over time. Based on this, we use the AR autoregressive method to establish a time-series model for the prices of second-hand sailboats. where

$$Y_t = c_1 Y_{t-1} + c_2 Y_{t-2} + \varepsilon_t \quad (11)$$

Where c_1 , c_2 are the parameter items to be determined, ε_t is a random perturbation term.

The parameters can be determined using the least squares method.

After obtaining the temporal relationship between per capita GDP and year, combined with the regression equation in Model I, we can establish an overall dynamic pricing model.

6.2 Prediction results and analysis

We take the dataset of Hong Kong as an example, select several typical variants of catamarans and monohulls, and plot their price curves with time as the independent variable. For the prediction of prices in the next 5 years, we assume that the international trade environment will remain relatively stable during this period, so the tariff rate remains unchanged.

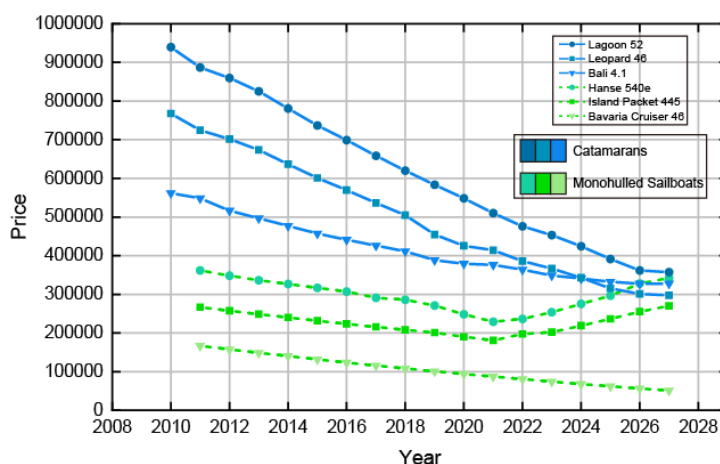


Figure 17: Typical ship variant price prediction curve

We have discovered some interesting findings.

- For the monohull boats, the variant Island Packet 445 and the variant Hansen 540e's price decreases first and then increases with time, while the prices of variants Bavaria Cruiser 46 decrease with time. Island Packet 445 and the variant Hansen 540e has a higher initial price and a higher overall price range than Hansen 540e. We speculate that Island Packet 445 and the variant Hansen 540e belong to a certain expensive sailboat brand with a greater collection value than utilitarian value, while the other one is generally used in the conventional shipping industry with higher utilitarian value.
- For the catamarans, the prices of all variants decrease with time, and the prices are generally higher than those of monohulls. We speculate that catamarans belong to a new type of

vessel with better cost-effectiveness than traditional monohulls, and therefore, their utilitarian value dominates. The improved performance naturally demands higher fees.

- We also found that the boat pricing in Hong Kong increased significantly less from 2019 to 2021 but then showed signs of recovery thereafter. We collected GDP data for these three years and found that the main factor may be the decline in GDP. The COVID-19 pandemic is undoubtedly the main behind-the-scenes culprit that drove this change. We also collected monthly epidemic data for Hong Kong to support the results, which showed a strong similarity in trend with the boat pricing.

7 Sensitivity Analysis

We mainly conducted sensitivity analysis on the core model, the PCSL model, which selects four indicators: Per capita GDP, length, draft, and proportion of water area.

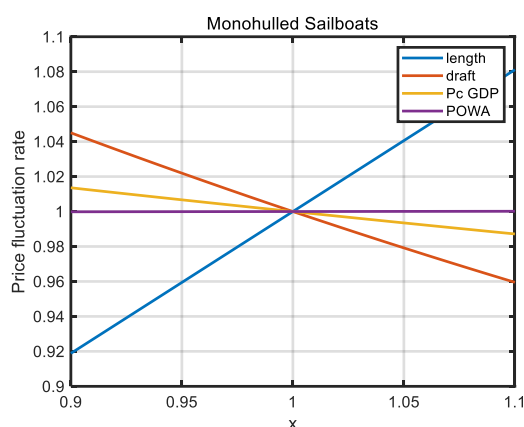


Figure 18: Sensitivity Analysis of Monohull Ships

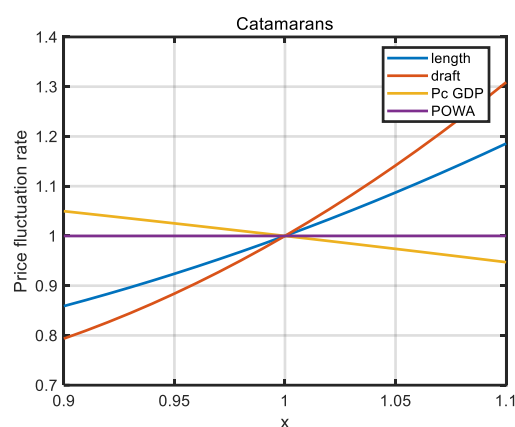


Figure 19: Sensitivity analysis of catamaran

- (1) For monohull sailboats:

Our study shows that for a 10% change in the independent variables, the maximum fluctuation in price will not exceed 8%. This indicates that the model has strong stability.

- (2) For catamarans:

The price of catamarans is more sensitive and less stable than monohull sailboats.

- (3) Common conclusions:

The relationship between price and each variable is consistent with our analysis in 4.3.2, which verifies the correctness of the regression results.

8 Model Evaluation

8.1 Strengths

- We combined two commonly used methods, principal component analysis and stepwise linear regression, to improve the accuracy of our model. Principal component analysis can avoid the coupling effect between independent variables, while stepwise linear regression can optimize the regression process independently, which improves the accuracy of the results. Our model validation based on the Hong Kong dataset also proved that our model

had a good regression performance, enabling a scientific and accurate evaluation of the pricing of second-hand sailboats.

- We conducted a detailed and comprehensive examination of the influence of region on pricing. Based on big data, we evaluated the impact of region on pricing, revealing objective laws of second-hand boat pricing in a statistical sense.
- We established both static and dynamic models for second-hand boat pricing and discovered some interesting conclusions, such as the hidden collectible value of sailboats may lead to longer usage time and higher prices.
- We conducted sensitivity analysis on the model and the results showed that the model was relatively stable. The model is beneficial for operators to formulate scientific marketing strategies in the boat market.

8.2 Possible improvement

- For different price indicators, the number of data collected varies, which may affect the representativeness of the results. This poses higher requirements on our data collection capabilities.
- In the analysis of the impact of regions on prices, we can further delve into which specific indicators in the regions affect prices. We preliminarily consider dividing them into natural geographical indicators and cultural and economic indicators, which can correspond to factors such as coastline length in the regression model and further optimize the regression model.

9 Conclusion

9.1 Basic conclusion

Among the many factors that affect sailboat pricing, we defined 8 independent variables: length, year, beam, draft, water area proportion, per capita GDP, tariff rate, and length of the coastline. We used principal component analysis combined with stepwise linear regression to establish regression equations. The results are shown in formulas 6 and 7.

Analyzing the regression equation for monohull sailboats, we found that length, beam, and draft have a significant positive correlation with price, and draft has the largest promoting effect. Due to the influence of the calculation definition, the water area proportion has little impact on price. Per capita GDP and tariff rates have a significant impact on pricing and show a certain degree of fluctuation over time.

Analyzing the regression equation for catamaran sailboats, we found that the principal component reflecting the ship's parameters has a greater impact on pricing than that of monohull sailboats. The principal component reflecting the natural geography of the region has more complex limiting conditions with other principal components. The principal component reflecting the human economy of the region also shows a clear time fluctuation.

The influence of the region on sailboat pricing is significant, and we have demonstrated this through data visualization and statistical analysis. The results of the experiments also show

that the effect of region change on different sailboat variants is different.

The application of the model to the Hong Kong region further demonstrates the applicability of the regression model and the significance of regional impact. The sailboat selling price in Hong Kong is often lower than that in other regions, and this is more significant for monohull sailboats than for catamarans.

9.2 Further conclusion

Our constructed price dynamic model has conducted a rich and meaningful exploration on the impact of the independent variable "year" on prices, and the main conclusions are as follows:

- Sailing boats have a dual attribute of use value and collection value. Sailing boats in different initial price ranges have different pricing patterns over time. Sailing boats with higher collection value will increase in price over time, while those with higher use value will decrease.
- The slowing of sailing boat price growth from 2019 to 2021 may be due to the impact of the pandemic.
- Based on our model, we have also predicted pricing for the next few years, which can provide reference for some commercial companies.

References

- [1] PENG W. Shipping market economics: knowledge evolution, second-hand ship price and freight rate[J/OL]. 2020[2023-03-31]. <https://theses.lib.polyu.edu.hk/handle/200/11613>.
- [2] Second hand vessel value estimation in maritime economics: A review of the past 20 years and the proposal of an elementary method | SpringerLink[EB/OL]. [2023-03-31]. <https://link.springer.com/article/10.1057/mel.2011.6>.
- [3] HAWDON D. Tanker freight rates in the short and long run[J/OL]. *Applied Economics*, 1978, 10(3): 203-218. DOI:10.1080/758527274.
- [4] Econometric Modelling of Second-hand Ship Prices | SpringerLink[EB/OL]. [2023-03-31]. <https://link.springer.com/article/10.1057/palgrave.mel.9100086>.
- [5] Asset Bubbles in Shipping? An Analysis of Recent History in the Drybulk Market | SpringerLink[EB/OL]. [2023-03-31]. <https://link.springer.com/article/10.1057/palgrave.mel.9100162>.
- [6] Generalized additive models in the context of shipping economics.[EB/OL]. [2023-03-31]. https://figshare.le.ac.uk/articles/thesis/Generalized_additive_models_in_the_context_of_shipping_economics_/10084175.
- [7] LUN Y H V, LAI K hung, CHENG T C E. *Shipping and logistics management*[M]. London New York: Springer, 2010.

Sail the price up!

With our mathematical regression model

profit



Dear broker,

Hello! Thank you very much for hiring our team to help you solve the issue of second-hand sailboat pricing in Hong Kong. We are honored to introduce our independently developed pricing system, SAILUP. SAILUP is a regression system based on big data. Since receiving your task, we have been working day and night to search for ship data from around the world, especially in the Hong Kong area. We conducted rigorous mathematical deductions

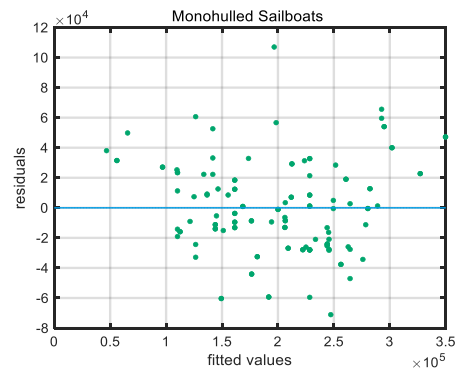
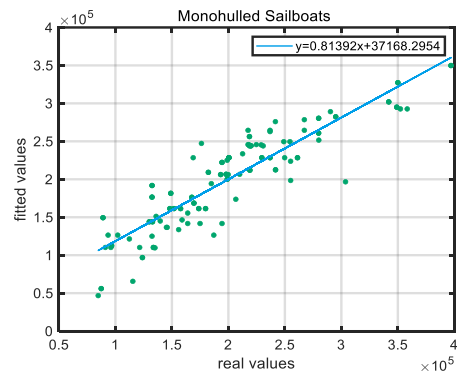


and statistical tests to develop this second-hand sailboat pricing system. We will present

our research results in the form of the following report, hoping to help you better understand my team's work and provide an initial solution to your problem.

During this period, we put a lot of effort into data mining. We finally collected a total of 5,158 valid data records, which were divided into different subsets based on factors such as different ship variants and years. Based on our research on market data,

we cleaned and supplemented the original data, and 98% of the data was retained. Based on this data, we defined eight independent variables that influence pricing, which are length, year, beam, draft, water area proportion, per capita GDP, tariff rate, and length of the coastline. We understand that you are particularly concerned about the pricing of monohulls and catamarans. We will use monohulls as an example to show the results of the regression analysis using the SAILUP system.





SAILUP: Your Dream Choice!

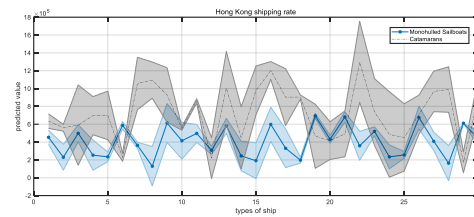


We conducted accuracy testing and error analysis of our pricing, and the results showed that our system has a very high matching accuracy. We also conducted extensive research on the ship market in Hong Kong. SAILUP system's results indicate that the selling price of ships in the Hong Kong area should be lower than the world average to be more competitive in the market, and the discount for monohulls should be greater than that for catamarans. This is a statistical conclusion we found. We also demonstrated that regional changes have different effects on different ship variants, and we can provide an analysis report on this part of the content. In addition, we can give optimal predictions for future sailboat pricing to help you gain a competitive advantage and increase profits.

Based on our report results, we have listed some feasible pricing strategies to help you achieve better profitability:

- (1) Due to preferential tariff policies for Hong Kong, you can price second-hand sailboats, especially monohulls, slightly lower than international market prices to maintain competitiveness.
- (2) For different types of ships, you can adopt different pricing strategies. For high-value collectible ships, you can wait for the right time to sell, while for ships with high utility value, we suggest that you sell them as soon as possible to prevent a decrease in their highest selling price.
- (3) Considering the impact of COVID-19 on the Hong Kong economy in recent years, we recommend that you expand your market share and transaction volume. China is in the economic recovery phase after the end of the epidemic, and the recovery of the

tourism industry is driving the growth in demand for ships. You can take this opportunity to increase revenue.



We understand that you may have some doubts about the prediction accuracy of our system. We have also conducted rigorous self-testing. First, based on mathematical principles, we conducted multiple method tests on the significance of regional impact, and our system passed all of the tests. We also compared actual price data with our predicted data, and you can see that our error margin is very small, which confirms the high applicability of our system for predicting second-hand sailboat prices in the Hong Kong area.

Dr. Hannah Fry once said that mathematics is the key to unlocking the secrets of big data. Our team firmly believes in this. All of our report findings and strategy recommendations are based on scientific data sets and rigorous mathematical logic. If there is an opportunity for offline communication, we would be happy to further explain the usage of the SAILUP system and more marketing strategies. Thank you again for your trust, and we look forward to hearing back from you.

Yours sincerely
Team 2332142