## 因子分析

## 基于BP神经网络回归的价格预测模型

## 5.1 The Foundation of Model

Back Propagation Neural Network (BP Neural Network) is a common artificial neural network with strong learning ability and adaptability. Its basic principle is to iteratively calculate the connection weights between neurons by using input and output data, in order to achieve approximation and prediction of nonlinear patterns.

The basic structure of BP neural network consists of input layer, hidden layer and output layer. The input layer accepts external input data, the hidden layer is an intermediate layer composed of multiple neurons, and the output layer outputs the final result. During the training process, the network uses forward propagation to pass the input data from the input layer to the output layer, and then calculates the output error using the back propagation algorithm, and adjusts the connection weights between neurons according to the error. This process is repeated many times until the network error reaches the preset threshold or the maximum number of iterations is reached.

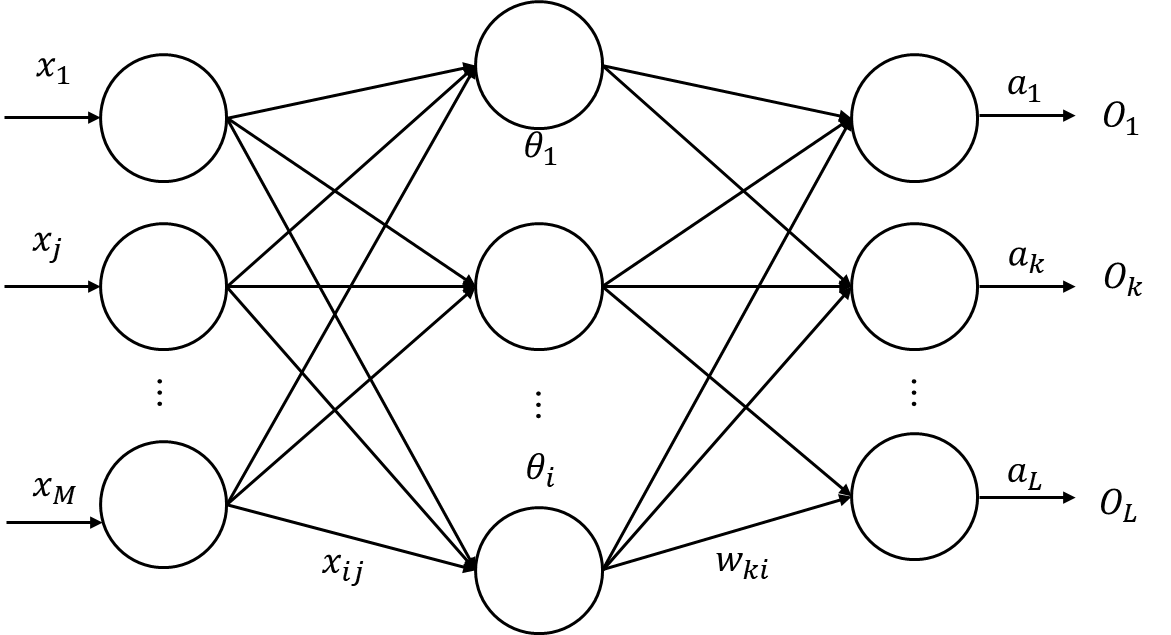


Figure :Three-layer Neural Network Structure

## 5.2 Algorithms used in the model

We used the Levenberg-Marquardt (LM) algorithm to train the neural network. LM algorithm (Levenberg-Marquardt algorithm) is a commonly used nonlinear least squares optimization algorithm, mainly used in the training process of neural networks.

The LM algorithm is based on the following objective function:

|  |  |
| --- | --- |
|  | () |

Where denotes the Euclidean norm, m is the number of residuals, and

is the predicted value of the input of function at .

In the LM algorithm, we first use the Gauss-Newton method to obtain an initial value for . Then, we calculate the gradient and the Hessian matrix at . Next, we use the following formula to calculate the update for :

|  |  |
| --- | --- |
|  | () |

where is the identity matrix and is a parameter that controls the step size. If is small, the update is close to the Gauss-Newton method, and if is large, the update is close to the gradient descent method. Therefore, the value of should be balanced between the two.

Finally, update the value of :

|  |  |
| --- | --- |
|  | () |

This process iterates until the error reaches the predetermined threshold or the maximum number of iterations is reached.

## 5.3 Training Process

1.我们使用因子分析和相关性分析得到的降维后的具有强相关性的数据作为输入矩阵矩阵的列变量，不同的品牌及变种的帆船种数作为矩阵的行变量，构成输入矩阵；以不同的品牌及变种的帆船作为矩阵的行变量，给定的价格作为矩阵的列变量，构成输入矩阵。

2.我们根据相关性分析的结论，选择了161个与帆船价格相关性较强的特征，并将补充收集的32984条数据进行划分，70%作为训练集，15%作为Validation Data，15%作为Test Data，构建含有1层输入层，10层隐藏层和1层输出层的神经网络。

3.我们使用MATLAB软件来编程实现上述操作。Training will continue until a pre-defined stopping condition is met (Epoch greater than or equal to 1000 or the validation error of the iteration continues to increase).

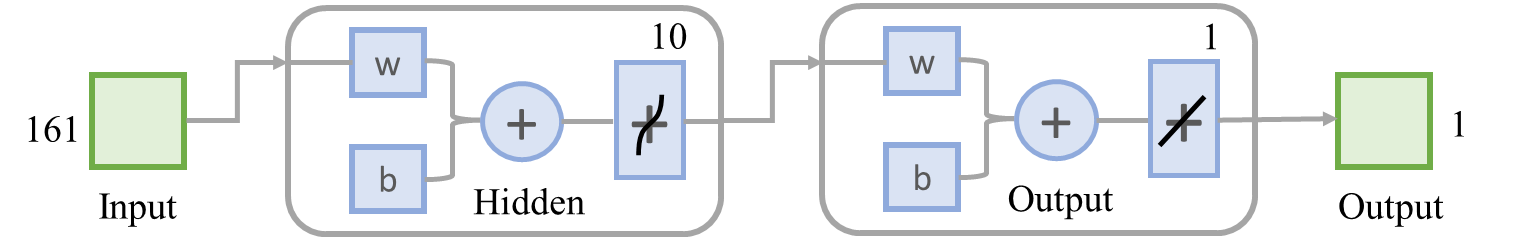


Figure : Neural Network Construction

## 5.4 模型准确性分析

Linear regression plots were generated to display the corresponding (target) network predictions (output) for the training set, validation set, and test set. If the fit is perfect, the data should fall along a 45 degree line where the network output is equal to the response. It can be seen from the figure below that the dataset fits very well. To obtain more accurate results, the dataset can be trained again with different initial weights and biases for the network, and the network can be improved after retraining.

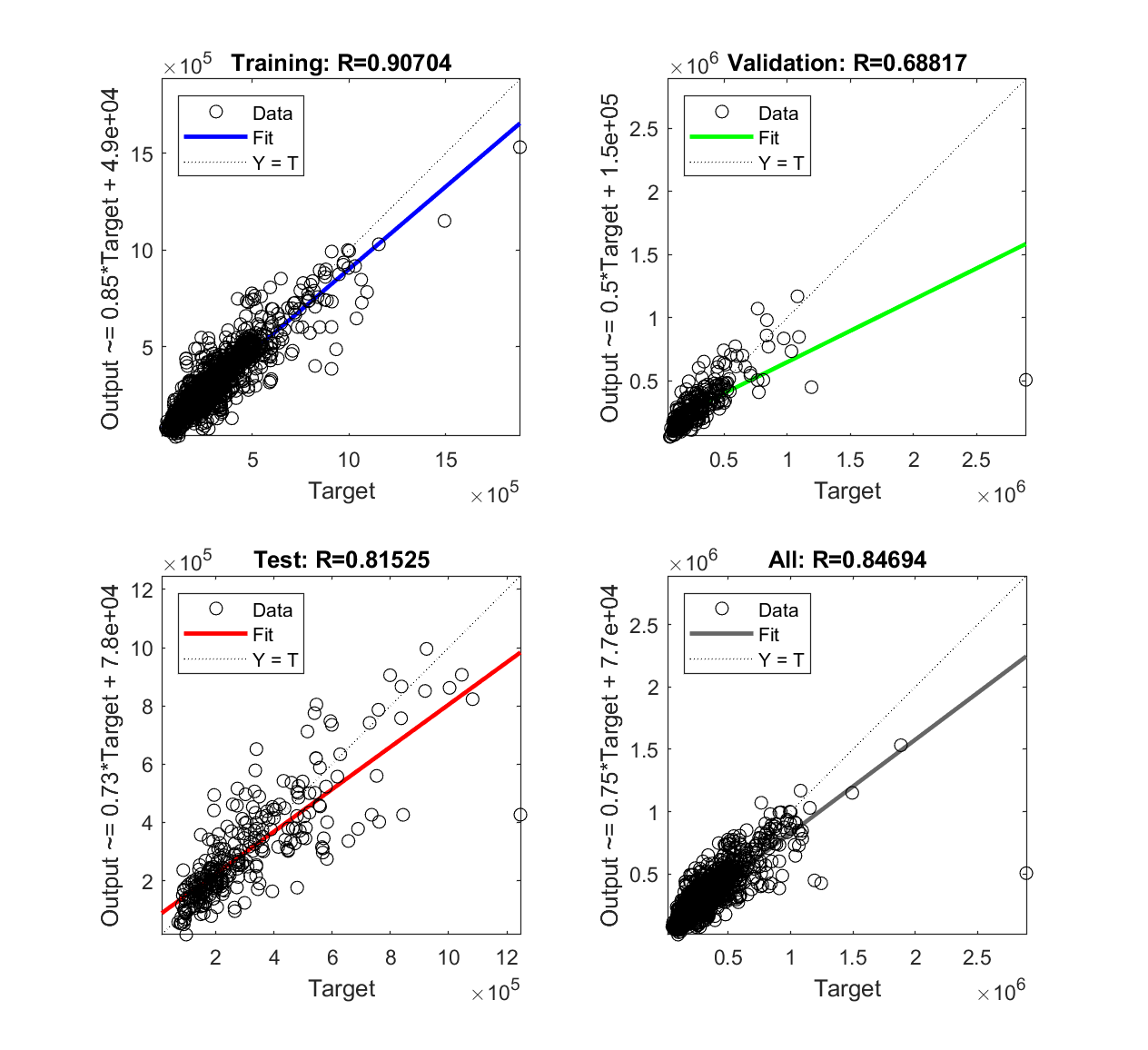


Figure :Regression plot

The results of the mean square error (MSE) and R-value indicate that the network has high accuracy in predicting the prices of used sailboats, reflecting the efficiency of the network model.

Table : Parameter Index Chart

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observations | MSE | R |
| Training | 23088 | 1.8087e+10 | 0.7237 |
| Validation | 4948 | 1.7412e+10 | 0.6873 |
| Test | 4948 | 2.3298e+10 | 0.6736 |

This training used conditional stop training, which stopped at 6 epochs.

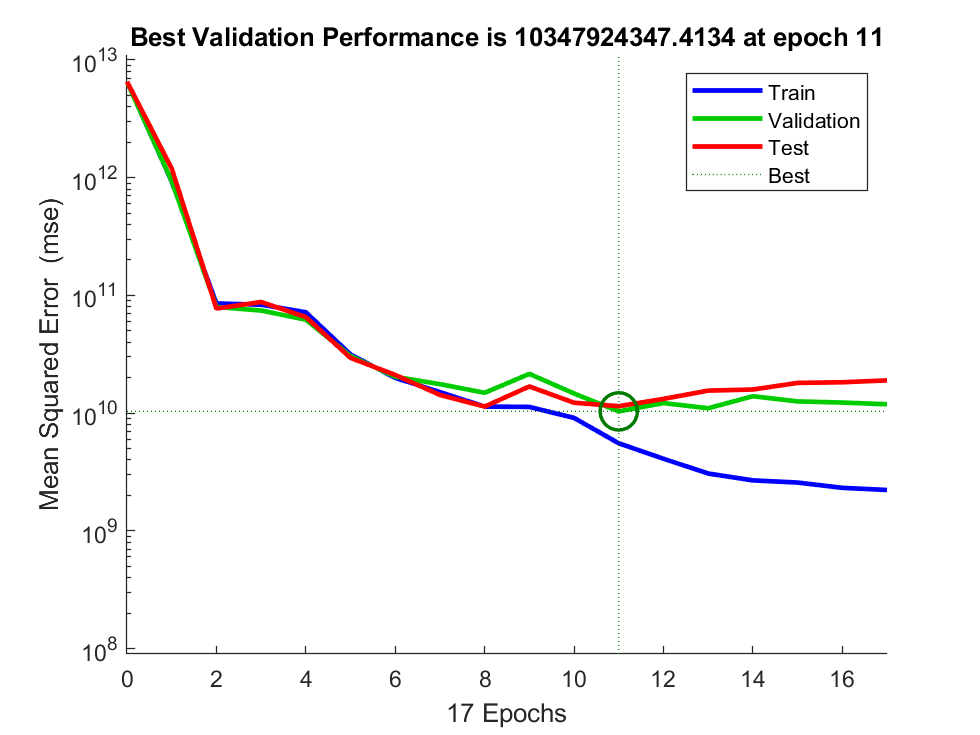


Figure :Optimal state chart

View the error histogram for additional verification of network performance.

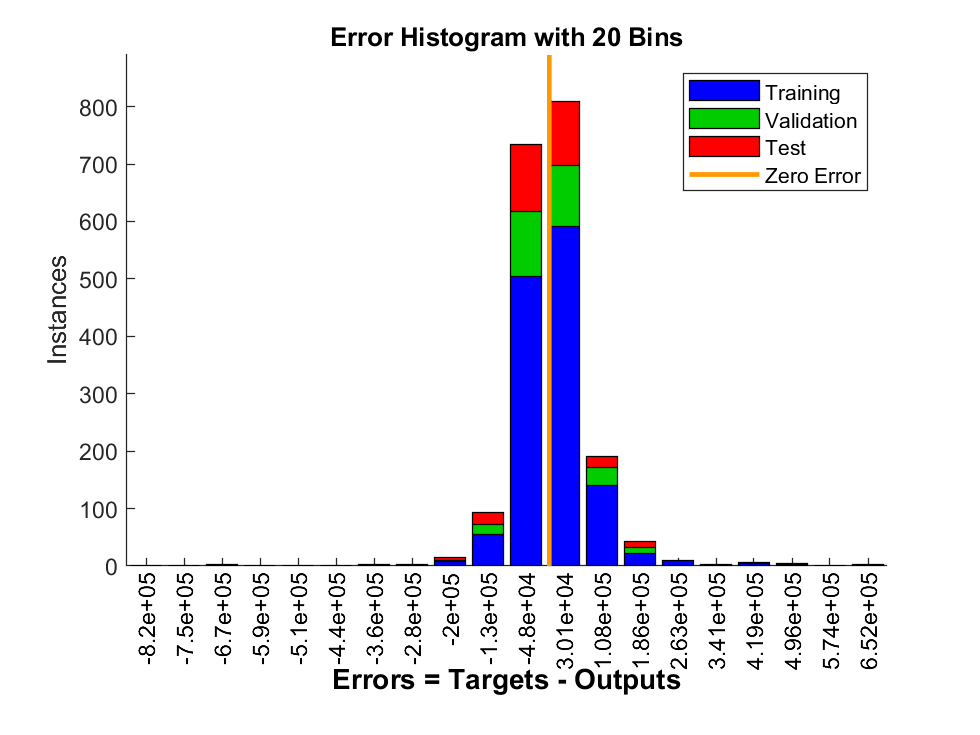


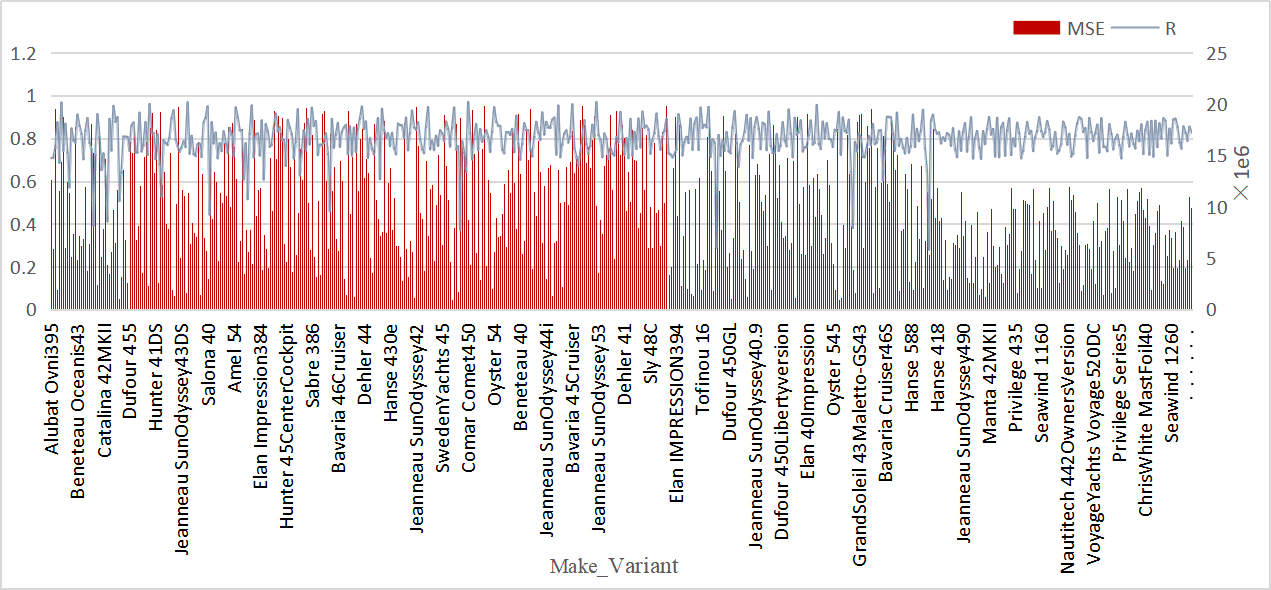
Figure : Error Histogram

The blue bars indicate training data, the green bars indicate validation data, and the red bars indicate test data. The histogram can indicate outliers, which are data points that fit significantly worse than the majority of the data. The outliers are examined to determine if the data are poor within the error range or if any of these data points differ from the rest of the data set, and the data points with valid outliers are interpolated using the network.

## 5.5 结果分析：

**5.1.1：对每一种帆船价格估计的精确度**

通过建立的BP神经网络模型，我们对题目给出的原始表格的品牌和变体进行了价格预测，下图是我们估计的价格与实际价格的偏差值的可视化图形（由于变体数量过多，只显示了限定长度截图，完整图片详见附录），采用R值与MSE值衡量。



（对每个变体估计的MSE和R变差）

由上图可以看到，该模型预测的价格与实际价格的R值偏差在0.7周围上下浮动，MSE控制在千万数量级，因此模型的预测具有较好的准确性。

观察发现，针对一些变体，模型预测的价格与实际价格偏差较大，通过查找发现，这些变体部分为私人订制帆船品牌，如Baltic 67，Oyster Yachts 545等。以及由于二手帆船使用状况，新旧情况未知等因素，对部分帆船的价格预测结果与实际有较大偏差。

综上，总体上模型对不同的变体的价格预测较为精确，误差在可接受范围内。但针对某些不同变体，精确度有较大偏差。

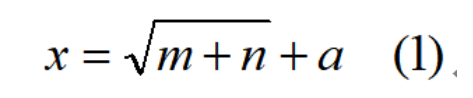
**5.1.2：区域对价格的影响及实际意义\***

5.1.2.1 数据集分割与处理

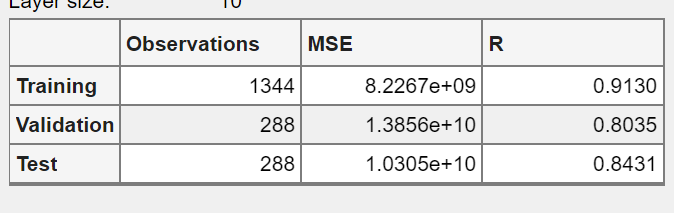
我们从补充完善后的数据集中筛选出了在三大区域都有售卖的变体，并对变体在同一区域的多个价格进行取平均处理，这样之后共计数据12861条。接着，我们将每个变体的三条数据按照地域分割再整合。分别经过因子分析降维后得到127维Factor作为列变量。至此，完成三个输入矩阵的构建。

5.1.2.2 针对三大区域分别建立三个神经网络

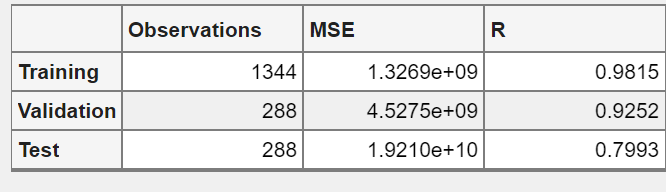
对处理好的数据进行BP神经网络训练，构建1层输入，10层隐藏层，1层输出层的神经网络模型，将数据集按70%、15%、15%随机划分为训练集、Validation data、test data。隐含层节点的范围由公式（1）计算得出，m代表输入层节点数，n代表输出层节点数，a的取值范围为1~10之间的整数，根据公式（1）求出10个可选择的隐含层节点数，依次对BP网络设置这10个隐含层节点并通过计算训练集均方误差MSE的方式，最终得出误差最小所对应的隐含层节点，即为最佳隐含层节点。



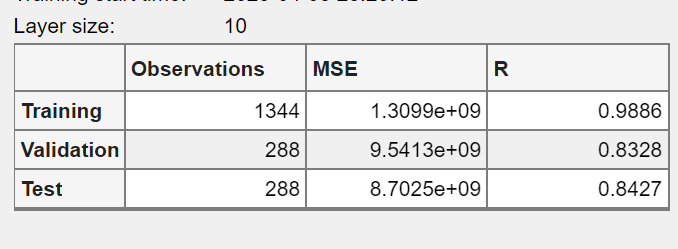
经过调试和训练，最终三个分区域训练的神经网络模型的对价格的预测与实际值相近。



（加勒比的模型误差）



（美国的模型误差，最右下数据改成0.8466）



（欧洲的模型误差）

（把这三张表画出来合成一张表（三张也可以），格式随便，清晰好看就可以Obs那一列是 9003, 1929，1929）

———————————————————————————————————————

得到上述三个神经网络模型后，我们再将原数据去除地域这一维变量，用因子分析进行降维，得到127个Factor。将该行列变量的矩阵作为输入，分别用建立的三个地域神经网络进行价格预测，得到三个价格预测数据数据集U\_pred, E\_pred, C\_pred。

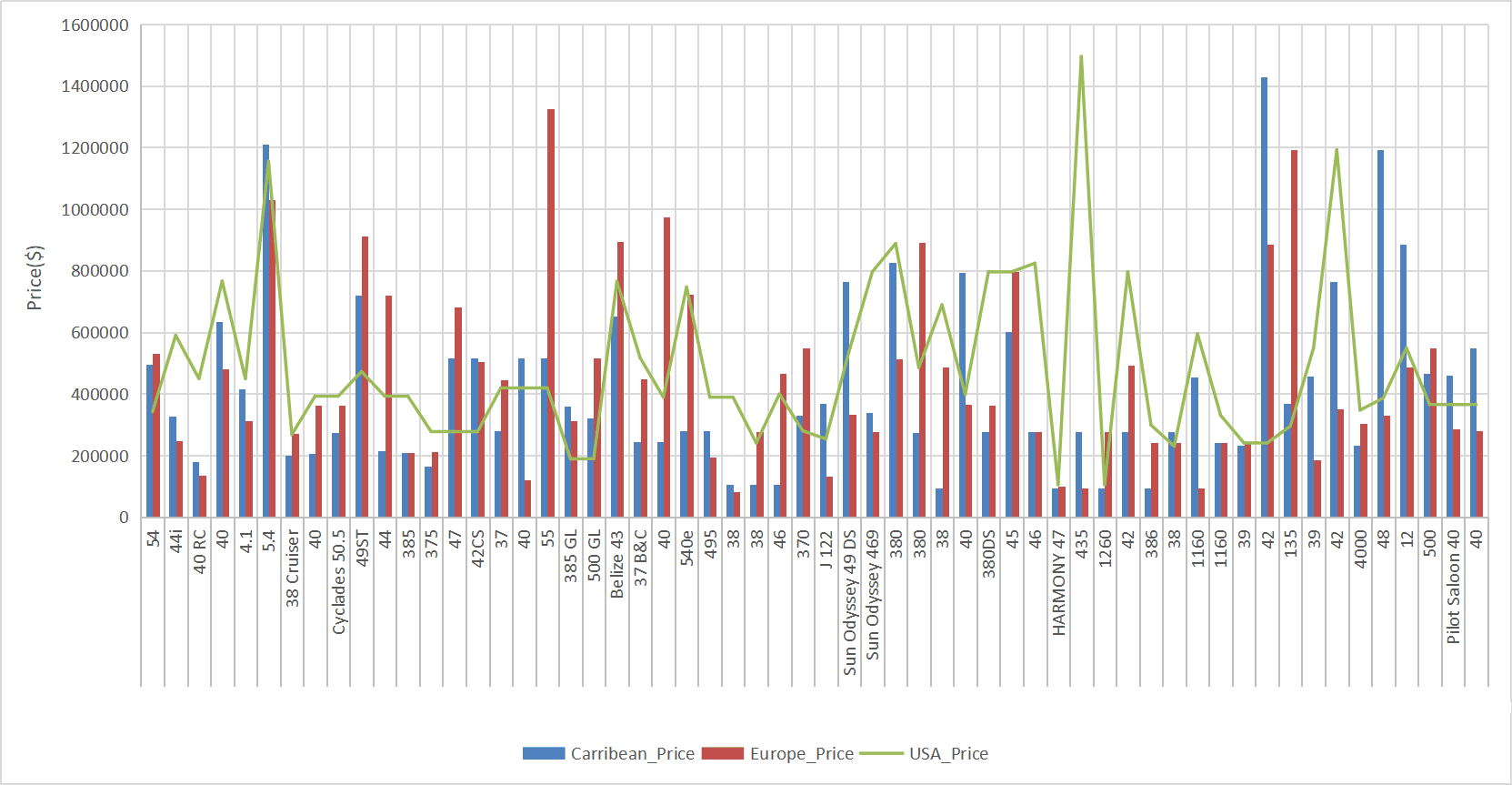
5.1.2.3 F检验分析结果

对得到的三个数据集分别画箱线图进行总体分析：

（箱线图，篇幅超的话可能要去掉这个图和上面这句话）

并以帆船变体为横轴，价格为纵轴，绘制三大区域的价格-变体曲线：

（插入折线图）



原理：  
要证明区域对每个变体影响一致，只需看每个变体在三个区域的价格分布是否具有一致性，因此使用F检验分析三个区域数据的方差是否相等。

F = Σni \* (x¯i - x¯)^2 / (k - 1)/ Σ Σ(xi - x¯i)^2 / (n - k)

其中，ni表示第i组的样本容量，x¯i表示第i组的平均值，k表示组数，n表示总样本容量，xi表示第i组中的观测值，x¯表示所有样本的平均值。。

统计意义：

假设H0：方差没有显著差异；H1：方差没有差异，在显著性水平\alpha=0.05，自由度为180下，我们计算的F检验结果为1.054<F0.025(180)=2.118，故拒绝原假设，认为方差没有显著性差异，即总体来说，区域对所有变体影响一致。

但由上述组合图我们观察发现，对于部分变体如435等，三个区域的预测价格相差相当大。推测与文化、港口数等因素相关。

实际意义：

上述结论说明，我们根据三大区域数据建立的BP神经网络模型价格预测模型对于香港经纪人预测香港市场的价格情况有具有一定的参考性。而针对地域性偏爱的帆船变种或是私人订制变种等非一般情况，则需另外分析讨论。

Task3：在香港区域的作用及对单双体船的影响

我们首先在主办方提供的电子表格中选择了变种种数大于等于3的记录作为数据子集，分为单体船和双体船。接着，我们从https://www.asia-boating.com/ 等相关网站（详见附录）上查到了香港的帆船变体的length、beam等上述提取出的重要特征参数和上市价格数据，并进行缺失值、异常值等处理，然后将这些参数输入上述训练的BP神经网络，得到香港地区各帆船的预测上市价格。

（插入单、双体船的图片）

如图所示，预测数据与实际数据略有出入，但总体上拟合效果较好，误差在可接受范围内。

据此，我们推断：

1.我们对区域影响的建模较为合理，在香港市场上可提供较好的预测价格参考，可以帮助卖家合理定价、帮助买家规划购买帆船开支、有利于稳定市场价格。

//(IMG_256https://gdp. IMG_257gotohui.com)提供的GDP数据，港口数量情况和气候情况（放附录）//

2.区域影响对单体船和双体船的影响不一样，对双体船的影响更稳定。通过文献查阅，我们分析认为，双体船与同样吨位的单体船相比，有更大的甲板面积和舱容，因此一般用于货船。由于货运市场的饱和度较高，双体船的市场需求也往往接近饱和，所以价格对需求变化的反应灵敏度低，价格稳定性较强。而且，双体船是将两艘船横向连接在一起，在海上航行不容易翻船，可以承受较大的风浪。