Time Series Regression and Instance-Based Learning for Bulk Shipping Daily Rate Prediction

MING-YU HUANG*, CHENG-HAN CHUA, HUNG-YI LO*** and MEI-HUI GUO****

** Transportation Department,*
****Green Energy and System Integration Research & Development Department,*
China Steel Corporation
*** Department of Applied Mathematics, National Sun Yat-Sen University*

China Steel Corporation ships raw material from producing countries to Taiwan by different chartered vessels. CSC spends about 200 million NTD per year on ocean freight costs. If we can predict the daily charter market rate of the future properly, CSC would be able to reduce their shipping costs through selective business operations. In this paper, we study two machine learning methods, time series regression model with Lasso and instance-based learning, for bulk shipping daily hire rate prediction. We developed some feature extraction methods for time series data. From the targeted daily hire rate index, we have observed a consistent time trend, called Chinese New Year effect. We designed a new feature for this effect and show that it is very useful. The best accuracy achieved by an average model on the test set is 65.7%.

Keywords: Bulk Shipping Daily Hire Rate, Time Series Analysis, Time Series Regression, Instance-Based Learning

1. INTRODUCTION

The China Steel Corporation (CSC) ships raw material from producing countries to Taiwan by different vessels. CSC spends about 200 million NTD per year on ocean freight costs. Some raw materials are transported by CSC-owned Capesize ships and the others are transported by Panama ships chartered from the market. If we can predict the daily hire market rate of the future properly, CSC would be able to reduce the shipping cost via some business operations.

In this paper, we study two machine learning methods, time series regression model with Lasso and instance-based learning, for bulk shipping daily hire rate prediction. Since most of the chartered Panamax vessels practice the route from Australia to Taiwan, our targeted daily hire market index is Baltic Panamax Index (BPI) P3A_03. We have gathered some factors or indexes which are correlated to P3A_03. These factors are exploited as input features to the machine learning methods. The remainder of this paper is organized as follows. In Section 2, we give an overview of the BPI P3A_03 and its related factors. Then, we present two methods for daily hire rate prediction in Section 3. We discuss the results of our experiments in Section 4. Finally, Section 5 contains some concluding remarks.

2. DATASET DESCRIPTION

2.1 P3A_03 Daily Hire Rate

In this paper, the targeted daily hire market index is P3A_03. The historical price of P3A_03 is shown in Figure 1.

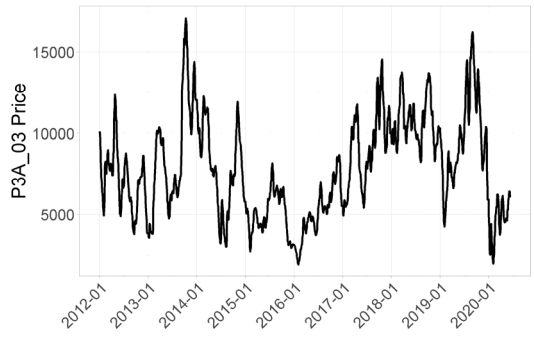

Fig.1. P3A_03 historical price

2.2 Related Factors of the Bulk Shipping Daily Hire Rate

Table 1 List of Related Factors of the Bulk Shipping Daily Hire Rate

Variable Notation	Variable Name	Description
x_1	BDI	Baltic Dry Index
x_2	BCI	Baltic Capesize Index
x_3	BPI	Baltic Panamax Index
x_4	BPI82 TCA	BPI82-TCA route rent price
x_5	BPI03 TCA	BPI03-TCA route rent price
x_6	C5	C5 route rent price.
x_7	C10	C10_14 route rent price
x_8	CRB	Commodity Research Bureau Index
x_9	Oil Price	Crude oil price
x_{10}	Oil Future	Crude oil futures price
x_{11}	P3A_03	P3A_03 route rent price
x_{12}	USA Dollar	US Dollar index futures
x_{13}	Newcastle Coal Futures	Coal futures prices

3. METHOD

3.1 Feature Extraction

Let P_t denote the rental price of P3A_03 at time t, where $\mathrm{t} = 1, 2, \ldots, \mathrm{T}$, and $R_{t+21}(21)$ be the 21-step simple return at time $t + 21$, that is

$$R_{t+21}(21) = \frac{P_{t+21} - P_t}{P_t}.$$

We will first build a time series regression model for $R_{t+21}(21)$ based on the aforementioned variables $(x_1, x_2, \ldots, x_{13})$ and their four kinds of transformations introduced below.

i. De-trend transformation of $x_j = (x_{1j}, \ldots, x_{Tj})'$

We estimate the trend of the variable x_j by the exponentially weighted moving average (EWMA) method [1]

$$e_{tj} = (1 - \lambda_j)e_{(t-1)j} + \lambda_j x_{tj},$$

and define the de-trend variable as

$$x^*_{\mathrm{tj}} = x_{\mathrm{tj}} - e_{tj}.$$

The smoothing parameter λ_{j} is chosen to maximize the absolute sample correlation between $\mathrm{R}_{\mathrm{t+21}}(21)$ and x^*_{tj} for $t \in \mathrm{T}_1$ where T_1 denoted the time index of the training set, that is $\hat{\lambda}_j = \mathrm{argmax}_{\lambda_j}\left|\mathrm{cor}_{\mathrm{t}\in\mathrm{T}_1}\left(R_{t+21}(21), x^*_{tj}\right)\right|$.

ii. Monthly Simple Return of x_j

Since there are around 21 transaction days (excluding the weekends) of P3A_03 in a month, we define the monthly simple return of x_j as

$$x^m_{tj} = \frac{x_{tj} - x_{(t-21)j}}{x_{(t-21)j}}.$$

iii. Weekly Simple Return of x_j

Similarly, since there are 5 transaction days in a week, we define the weekly simple return to be

$$x^w_{tj} = \frac{x_{tj} - x_{(t-5)j}}{x_{(t-5)j}}.$$

iv. Daily Simple Return of x_j

$$x^d_{tj} = \frac{x_{tj} - x_{(t-1)j}}{x_{(t-1)j}}.$$

We also include the monthly standard deviation of log rental price of P3A_03,

$$\sigma^m_{t,P} = \mathrm{sd}\left(\left\{\log\,(1 + P_j) : j = t - 21, \ldots, t\right\}\right),$$

and the Chinese new year effect introduced below as features in the model.

Figure 2 shows the time plots of $\mathrm{R}_{\mathrm{t+21}}(21)$ for eleven years, 2010-2020, which indicates there is a consistent time trend in the period Jan-01 to Mar-04

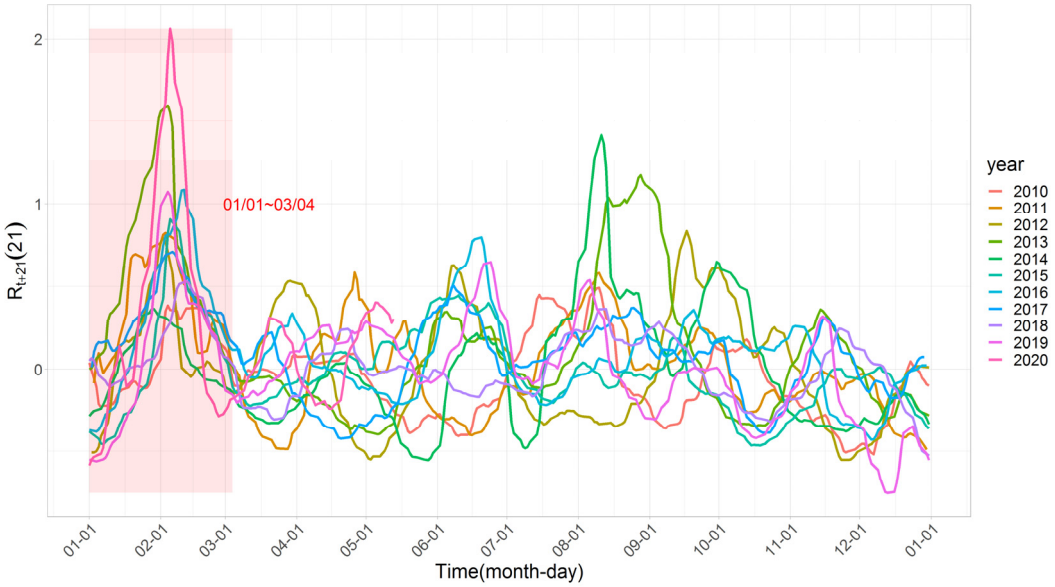

Fig.2. Time plots of $R_{t+21}(21)$ for eleven years, 2010-2020.

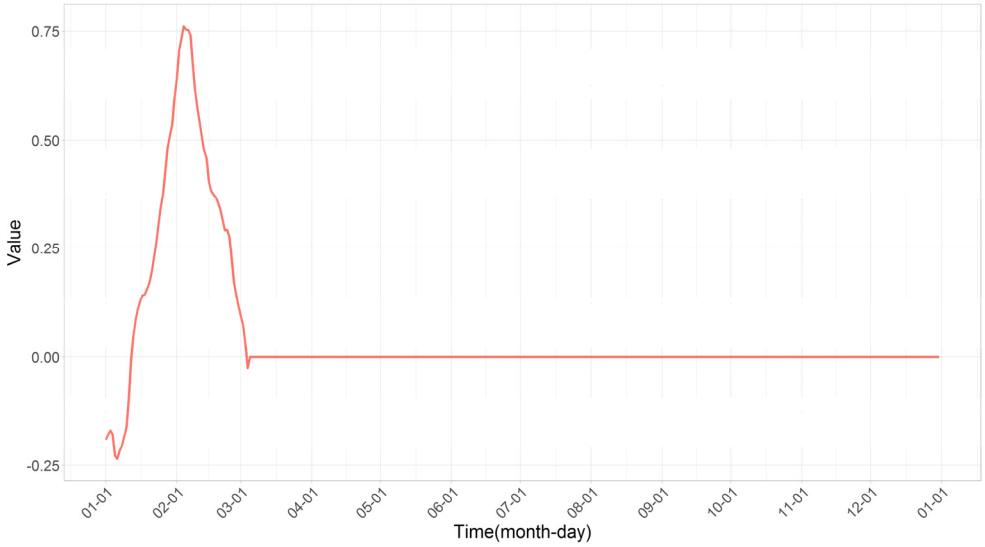

Fig.3. Chinese New Year feature $x_{t,C}$

for all years. We name this consistent time trend the Chinese New Year effect, since the time period includes the Chinese New Year period. According to the Economic Complexity Index (ECI) China is the largest importing and exporting country of Australia, Chinese New Year festival has a certain effect on the rental prices. For simplification of notation, we re-denote $R_{t+21}(21)$ as $R^{y}_{d+21}(21)$, where y is the year index and d is the day index in a year. We used the yearly average (2010-2017) curve of the period Jan-01 to Mar-04 to represent the Chinese New Year feature defined below.

$$x_{t,C} = \frac{1}{8}\sum_{y=2010}^{2017} R^{y}_{t+21}(21)\, \mathrm{I}_{t\in(\mathrm{Jan.01,Mar.04})}\,.$$

3.2 Time Series Regression Model and Lasso

Consider the following time series regression model with the lasso regularization[2]

$$R_{t+21}(21) = \beta_0 + \sum_{j=1}^{p} \beta_j z_{tj} + \varepsilon_t,$$

where z_{tj} denote the features introduced in previous section and the parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ are estimated by minimizing the mean squared error with the lasso regularization

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{T-21} \sum_{t=1}^{T-21} \left(R_{t+21}(21) - \beta_0 - \sum_{j=1}^{p} \beta_j z_{tj} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where the regularization parameter λ is decide by cross validation. The predicted value of $R_{t+21}(21)$ is

$$\hat{R}_{t+21}(21) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j z_{tj}.$$

3.3 Instance-based Learning

We also exploit an instance-based learning method[3] to predict $R^y_{t+21}(21)$. Assume the current year and day are $y = y^*$ and $t = d^*$. The following are the steps of the proposed instance-based learning method.

i. Let $S^y_d = \left(R^y_{d-7}(21), \ldots, R^y_d(21)\right)$ denote the vector of the seven-day 21-step simple returns on days $d-7, \ldots, d$ in the year y. Calculate the following Euclidean distance between $S^{y^*}_{d^*}$ and $S^y_{d^*}$,

$$D^{y^*,y}_{d^*} = ||S^{y^*}_{d^*} - S^y_{d^*}||_2 \text{ for } y = 2010, \ldots, y^* - 1.$$

ii. Let y_1, y_2 and y_3 denote the years which correspond to the three- smallest distances $D^{y^*,y}_{d^*}$.

iii. Predict $R^{y^*}_{d^*+21}(21)$ by the average of $R^y_{d^*+21}$ for years y_1, y_2 and y_3, that is

$$\hat{R}^{y^*}_{d^*+21}(21) = \frac{1}{3} \sum_{y=y_1,y_2,y_3} R^y_{d^*+21}(21).$$

3.4 Categorization of $R_{t+21}(21)$ and model evaluation

In order to reduce the total cost of chartering a vessel, when $R_{t+21}(21)$ becomes low (high), the next month (current) will be a better time to charter a ship. In practice, the thresholds of $R_{t+21}(21)$ to decide whether renting a ship or not are set to be ± 0.2. We introduce the following indicator function

$$c(x) = \begin{cases} 1, & \text{if } x > 0.2 \\ 0, & \text{if } |x| \le 0.2 \\ -1, & \text{if } x < -0.2 \end{cases},$$

to segment the values of x into three categories 1,0

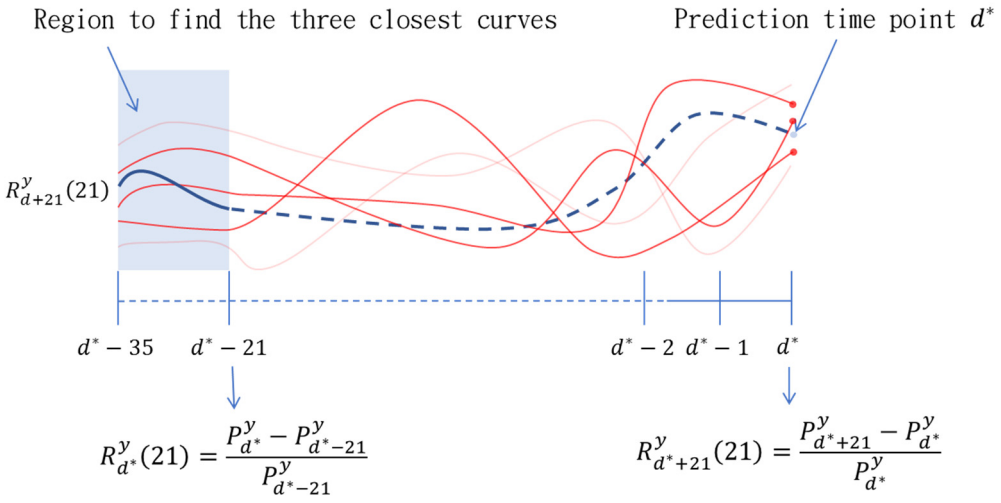

Fig.4. Time plot of $R^y_{d+21}(21)$ for different years, the blue curve is for the year y^*, the red curves are the three-closest curves and the light red curves are the curves for other years.

and -1. The prediction performance of the time regression model and the instance based learning method will be evaluated based on the accuracies and Macro-F scores of $c\left(\hat{R}_{t+21}(21)\right)$ and $c\left(\hat{R}^{y}_{t+21}(21)\right)$, respectively. Given a three-class classification model, we can obtain the following contingency table

		Predict		
		-1	0	1
True	-1	$n_{-1,-1}$	$n_{0,-1}$	$n_{-1,1}$
	0	$n_{0,-1}$	$n_{0,0}$	$n_{0,1}$
	1	$n_{1,-1}$	$n_{1,0}$	$n_{1,1}$

The accuracy of the model is defined as

$$\text{Accuracy} = \frac{\sum_{i=-1}^{1} n_{i,i}}{\sum_{i=-1}^{1}\sum_{j=-1}^{1} n_{i,j}}$$

Another criterion is the Macro-F score which is defined as the average F1-score of all classes, where the F1-score is the harmonic mean of precision and recall.

$$\text{Macro} - \text{F} = \frac{1}{3}\sum_{l=-1}^{1} F_l$$

$$= \frac{1}{3}\sum_{l=-1}^{1} \frac{2}{\frac{1}{\text{Precision}_l} + \frac{1}{Recall_l}},$$

where

$$\text{Precision}_l = \frac{n_{l,l}}{\sum_{i=-1}^{1} n_{i,l}} \quad \text{Recall}_l = \frac{n_{l,l}}{\sum_{i=-1}^{1} n_{l,i}}.$$

For multi-class imbalanced data, Macro-F is a more suitable criterion.

4. RESULTS AND DISCUSSION

The time series regression model trained by the training set is

$$\begin{aligned}\hat{\text{R}}_{\text{t+21}}\,(21) = {} & 0.06 + 0.15\,\text{x}^{*}_{\text{t,1}} - 0.04\,\text{x}^{*}_{\text{t,8}} + 0.03\,\text{x}^{*}_{\text{t,10}} \\ & -0.24\,\text{x}^{*}_{\text{t,11}} - 0.05\,\text{x}^{*}_{\text{t,12}} \\ & + 0.05\,\text{x}^{*}_{\text{t,13}} \\ & +0.05\,\text{x}^{\text{d}}_{\text{t,11}} + 0.03\,\text{x}^{\text{w}}_{\text{t,1}} + 0.02\,\text{x}^{\text{m}}_{\text{t,6}} \\ & +0.01\,\sigma^{m}_{t,P} + 0.17\,x_{t,C}\end{aligned}$$

We observed the following three features with the largest coefficients, the historical P3A_03 price, the Chinese New Year effect and the Baltic Dry Index.

The accuracies and Macro-F scores of the three models: the time series regression model, instance-based learning method and the average model, are presented in Table 2. The average model uses the average values of the time series regression model and instance-based learning method to predict $\hat{\text{R}}_{\text{t+21}}\,(21)$.

From Table 2, we observe that the instance-based learning performs better than the time series regression in terms of both accuracy and Macro-F on the training set. However, the time series regression has better generalization ability on the test set. The average model always performs better than each of the single models. The best accuracy achieved on the test set is 65.7%.

Forecasting future financial trends is a difficult task. The trend may be influenced by political events, natural disasters, or some other unpredictable events. Perfect predicting of the future bulk shipping daily hire rate is unrealistic. However, we believe that CSC transportation operations could benefit from this work for two reasons. First, the observation on the Chinese New Year effect can help CSC to avoid chartering too much in early February and try to charter more in early January. Second, current operations do not consider the future trend of the daily hire rate. We believe that the CSC transportation management may perform better than before by considering the prediction of the average model.

Table 2 Accuracy and Macro of the time series regression model, the instance-based learning method and the average model.

	Model	Accuracy	Macro-F
Training set	Time series regression model	0.613	0.570
	Instance-based learning method	0.626	0.592
	Average Model	0.672	0.622
Test set	Time series regression model	0.626	0.560
	Instance-based learning method	0.588	0.496
	Average Model	0.657	0.567

5. CONCLUSIONS

Raw material and its shipping are parts of the major costs in CSC. In this paper, we exploit machine learning methods for bulk shipping daily hire rate prediction. Furthermore, we discovered an interesting consistent time trend from the historical daily hire rate data and show that it is very useful for making predictions. We believe that CSC transportation operations could benefit from this work.

REFERENCES

1. Hunter, J. Stuart. "The exponentially weighted moving average." Journal of quality technology 18.4 (1986): 203-210.
2. Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.
3. Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." Machine learning 6.1 (1991): 37-66. □