Fisheries Science (2021) 87:411–423
https://doi.org/10.1007/s12562-021-01498-6

ORIGINAL ARTICLE

Social Science

Predicting the price of Vietnamese shrimp products exported to the US market using machine learning

Nguyen Minh Khiem[1,4] · Yuki Takahashi[2] · Khuu Thi Phuong Dong[3] · Hiroki Yasuma[2] · Nobuo Kimura[2]

Received: 7 August 2020 / Accepted: 1 March 2021 / Published online: 1 April 2021

Abstract

Accurately predicting the price of exported fishery products is an important task for fisheries because it will enable market trends to be determined, leading to the development of high-quality fishery products. In this study, we predicted prices in selected base periods (2, 3, 6, and 12 months) to investigate how historical data influenced the Vietnamese export price. A dataset (from May 1995 to May 2019) was collected from the US Department of Agriculture (USDA). We initially hypothesized that the dependent variable, Vietnamese export price, was affected by 33 independent variables, but ultimately used 15 key variables, which were chosen on the basis of Akaike information criterion (AIC) to train the models. A tree-based machine learning technique, including the random forest and gradient boosting tree algorithms, was applied for predictions. It was found that the random forest algorithm performed well for historical data for periods of more than 6 months, while the gradient boosting tree algorithm was better over short durations of less than 6 months.

Keywords Export price · Shrimp product · Machine learning

Introduction

Seafood has become a commercial food sector in the global marketplace. The total global value of seafood imports is currently more than 140 billion USD. The main import markets for seafood products are dominated by developed countries such as the USA (14.4% of global imports), Japan (10.6%), Spain (5%), France (4.7%), Germany (4.4%), Italy (4.4%), and Sweden (3.4%), all of which have a long history of seafood consumption (FAO 2018). Shrimp is the major seafood product in terms of daily food items, and accounts for 15.5% of the total value of seafood products in the world. There are around 60 countries that contribute to global shrimp production. The average growth rate of shrimp production was more than 20% from 2000 to 2016, making it one of the fastest growing commodities in the food industry (Flaaten 2018). Vietnam is the second largest global supplier, accounting for 9% of the total export value of shrimp products in the global market (FAO 2018). Most shrimp production worldwide is export oriented. The US, European, and Japanese markets are the major import destinations for shrimp products, accounting for 40% of the total import value of shrimp products in the world.

In Vietnam, the developing fishery industry is a key economic sector that plays an important role in rural development, income generation, and the improvement of livelihoods (Duc 2009; Phuong and Oanh 2010). According to the General Statistics Office of Vietnam (GSO) in 2018, the Vietnam Mekong Delta (VMD) region has the most shrimp aquaculture activity in the country, accounting for more than 70% of the total national production (Vietnam GSO 2018). Most (70–80%) farmed shrimp products are exported (Portley 2016; Tran et al. 2013). Shrimp fisheries account for a large proportion of the gross domestic product (GDP) of Vietnam, with production levels of 6.1 tons of product in 2015 and 6.7 tons in 2016 (COFI 2019). Vietnam exports

✉ Nguyen Minh Khiem
nmkhiem@cit.ctu.edu.vn

[1] Graduate School of Fisheries Sciences, Hokkaido University, Hakodate, Hokkaido 041-8611, Japan

[2] Faculty of Fisheries Sciences, Hokkaido University, Hakodate, Hokkaido 041-8611, Japan

[3] School of Economics, Can Tho University, Can Tho, Vietnam

[4] College of Information and Communication Technology, Can Tho University, Can Tho, Vietnam

fishery products to most regions of the world. According to the Food and Agriculture Organization (FAO), more than 50 countries and territories consume Vietnamese fishery products. Frozen shrimp is the main export product of the Vietnamese shrimp industry. In 2017, the Vietnam shrimp export value was 3.85 billion USD, which accounted for 44% of total national exports. The main export markets for Vietnamese shrimp products are nations with high living standards, including the USA, Japan, and Europe. The top ten import markets for Vietnamese shrimp products account for more than 95% of the total shrimp export value of the whole country, namely, the USA (17%), EU (22%), Japan (18%), China (17%), South Korea (10%), Canada (4%), Australia (3%), the Association of Southeast Asian Nations (ASEAN) countries (2%), Taiwan (1%), and Switzerland (1%). The three main import markets are the USA, Europe, and Japan, which together account for more than 50% of the total export value (VASEP 2018).

These markets have stringent traceability requirements, and quality assurance certificates are needed for imported seafood products. Previous studies (Suzuki and Nam 2013; Ha and Bush 2010; Duc 2010) have reported that these requirements in the US, European, and Japanese markets are not only aimed at food safety and the protection of consumer health but also act as trade barriers to protect their own domestic seafood production. Remarkably, the price of Vietnamese shrimp products exported to the USA, Europe, and Japan is 20% higher than the export price to other countries (VASEP 2018; FAO 2018). This has encouraged producers to export to these markets, and Vietnamese shrimp exporters have recently been challenged by high levels of competition from exporters from other countries (Flaaten 2018). Vietnamese producers must therefore seek ways to improve their competitive advantage to prevent their competitors from gaining a better price for the export of their products to these high-value markets (Dong and Duc 2012).

For Vietnamese fisheries to export fishery products successfully, we focused on key factors that could improve both the quality and quantity of shrimp production, such that more export markets could be targeted. Vietnamese shrimp producers must also meet the stringent food safety and product quality requirements of export markets.

Since Vietnam joined the World Trade Organization (WTO) in January 2007, it has had access to global export opportunities. Accordingly, the quality of Vietnamese shrimp production has been improved, allowing it compete with other countries. This is an important factor that has affected the export price of shrimp and promoted the development of shrimp farming in Vietnam.

To achieve this, accurate forecasting and identification of factors influencing the price of Vietnamese products in the export markets is essential. Furthermore, an understanding of the fluctuation of market trends is needed to determine how Vietnamese producers can overcome market challenges and develop strategies to grow the export of fisheries products. This will enhance quality and increase the quantity of fishery products.

In this study, we applied machine learning to solve some of the issues encountered in the shrimp farming process, including the prediction of market trends. Machine learning studies targeting shrimp production in Vietnam have been conducted previously to detect disease (Leung and Tran 2000; Khiem et al. 2020), predict fish stocks (Brander 2003), and predict the distribution of fishing activities (Soykan 2014). Many studies of aquaculture applications based on machine learning algorithms have also been published (Rahman 2014). Algorithms developed from learned models based on farm practices and environmental data sources have been implemented in applications such as shellfish farm closure predictions, algal bloom predictions, model relocations, and sensor data quality assessments. Pavlyshenko (2019) applied a machine learning model for sale forecasting. He used a linear regression to determine the bias in a validation test and applied the random forest algorithm as a supervised approach toward a time series of historical sales. The results revealed an accuracy of 3.9% for training set error and 11.6% for validation set error. Although the autoregressive integrated moving average method (ARIMA) is often used to predict data trends in time series modeling, the random forest method is better than ARIMA because it provides a superior prediction accuracy. Dudek (2015) showed that, in comparison with ARIMA, the random forest model was highly accurate in terms of seasonal time series forecasting of economic power generation and power system security. Stoll (2020) confirmed that machine learning approaches (including random forest, neural network, and K-nearest neighbor) were more powerful than traditional methods (including ARIMA, exponential smoothing, and moving average) in forecasting supply chain demand. Kamil (2020) found that, for water demand forecasting, random forest was the most accurate algorithm (90.4% accuracy), followed by ARIMA (90.0%).

Additionally, machine learning has been used in fishery economics to predict shrimp growth in commercial settings. There has been some analysis of commercial shrimp farms in Hawaii using logistic forms and artificial neural networks (Yu 2006). The results of a comparison showed that neural networks outperformed regression models and represented a reliable tool for predicting shrimp growth in commercial farms. To support decision making in aquaculture farm closure, a machine learning technique was applied to accurately predict the closure of shellfish farms (Shahriar 2014). In this study, an environmental time series was the main factor that was combined with a machine learning algorithm to suggest farm closure solutions. Manuel (2012) used a tree-based machine learning technique to extract the hidden information contained in large databases in fisheries research. As mentioned above, many studies have applied machine learning to fisheries, aquaculture, and economic analyses. However, the application of machine learning techniques has not been used to predict the fluctuation of Vietnamese aquaculture prices.

In this study, we used a machine learning technique to predict the price of shrimp products on the basis of case studies of Vietnamese frozen shrimp products exported to the US market. Predictions were made on the basis of factors influencing the import price of Vietnamese shrimp products in the target markets. Then, the relationship between predicted result and those factors was discussed. The results should assist policymakers to develop shrimp export strategies for global markets.

Material and methods

Dataset

Monthly price data for the import of Vietnamese frozen shrimp products to the USA and that of its competitors (e.g., Chile, Ecuador, China, India, and Thailand) from May 1995 to May 2019 were collected from the US Department of Agriculture (USDA). Data for the other factors influencing the import price of Vietnamese frozen shrimp products were obtained from the FAO, WTO, and International Monetary Fund (IMF). A description of all dataset variables is presented in Table 1.

The USA is one of the largest importers of shrimp products in the world. The USA accounts for 21.2% of the total exported value of Vietnamese shrimp products (VASEP 2020). Hence, any changes in demand from the USA would probably affect the quantity and price of Vietnamese shrimp products. Dong and Duc (2012) suggested that the total demand quantity for seafood products (i.e., catfish) imported from Vietnam depends on the consumer price and other demand drivers in the USA and rest of the world. Among these, the consumer price was found to be a direct determinant of the demand for Vietnamese shrimp products. Other factors driving demand included the income per capita of importing countries, the price of alternative goods, and improvements in production technology leading to increased product quality and thus higher prices. In addition to this, the prices of Vietnamese shrimp products exported to global markets are presented in US dollars. Thus, the prices of Vietnamese shrimp products depend on the exchange rate between the USD and VND, and the prices of Vietnamese shrimp products.

Besides Vietnam, the largest suppliers of shrimp products imported to the USA include India, Thailand, Indonesia, and Ecuador (International Trade Center 2020). These countries compete directly with Vietnam for shrimp imports to the US market. Thus, the imported quantity and prices of Vietnamese shrimp in the US market are expected to decrease if the prices of shrimp products imported from these other countries decrease (Tucker 2008). In the present study, the prices of shrimp products imported from other countries to the USA were included in the model to examine their effects on the price of Vietnamese shrimp products imported to the USA.

A number of recent food safety incidents have recently led to increased interest in food safety and quality assurance. There are various safety requirements and quality assurance issues, which protocol for shrimp exporters who wish to export their products to the USA and global markets (Dong et al. 2019). In 2019, approximately 90% of retailers in the market of the northern USA and more than 75% of retails in the EU market imposed those certifications on Vietnamese imported shrimp (VASEP 2020).

Dong et al. (2019) addressed that the mandatory regulations have been issued at the national level in the US market for imported shrimp products. Accordingly, the USA issued an official mandatory regulation, namely Country of Origin Labeling (COOL), regarding the traceability of fishery products in 2004 (which became effective in 2005). The regulation ensures consumer rights in terms of clarifying the country of origin and product traceability. Its aim was to provide information to consumers about the country of origin and method of production for both domestic and imported fishery commodities at the point of sale, including imported shrimp products (from USDA in 2016). Additionally, an antidumping tariff was imposed on Vietnamese shrimp products exported to the USA from 2003. The effects of the antidumping tariff on Vietnamese seafood products were examined by Duc and Kinnucan (2007), and their results indicated a negative effect on Vietnamese seafood price.

Besides the national regulations, various safety quality practices have been obligatorily imposed on Vietnamese shrimp products such as the Hazard Analysis and Critical Control Point (HACCP) system, the Safe Quality Food (SQF) program, and the Aquaculture Stewardship Council (ASC). Thus, Vietnamese shrimp exporters have to obtain the HACCP and ASC certifications for shrimp products to meet the mandatory practices in the US market. Additionally, in terms of requirements of quality certification in the global markets, Vietnamese shrimp exporters must achieve the Good Aquaculture Practices (Global GAP) certification to be eligible to export to the US market. Although a minimum set of standards as the guidelines for shrimp products to be accepted by various importing countries has not yet been determined (Bailey et al. 2018), Vietnamese shrimp exporters must obtain those certifications to meet the requirements of their targeted markets. However, shrimp

Table 1 Definitions and descriptions of the variables used in the model

Name	Definition	Description	Sources
PriceVietnam	Price of Vietnamese shrimp products in the US market	Dependent variable	USDA
LagPVietnam	Price of Vietnamese shrimp products in previous month	Continuous variable	
PChina	Price of shrimp products imported from China to the US market	Continuous variable	USDA
LagPChina	Price of shrimp products imported from China in previous month	Continuous variable	
PThailand	Price of shrimp products imported from Thailand to the US market	Continuous variable	USDA
LagPThai	Price of shrimp products imported from Thailand in previous month	Continuous variable	
PChile	Price of shrimp products imported from Chile to the US market	Continuous variable	USDA
LagPChile	Price of shrimp products imported from Chile in previous month	Continuous variable	
PIndonesia	Price of shrimp products imported from Indonesia to the US market	Continuous variable	USDA
LagPIndonesia	Price of shrimp products imported from Indonesia in previous month	Continuous variable	
PEcuador	Price of shrimp products imported from Ecuador to the US market	Continuous variable	USDA
LagPEcuador	Price of shrimp products imported from Ecuador in previous month	Continuous variable	
PIndia	Price of shrimp products imported from India to the US market	Continuous variable	USDA
LagPIndia	Price of shrimp products imported from India in previous month	Continuous variable	
Exchange	Exchange rate VND/USD	Continuous variable	www.oanda.com
LagExchange	Price of exchange rate from VND to USD in previous month	Continuous variable	
Yus	US income per capita	Continuous variable	BEA
LagYus	US income per capita in previous month	Continuous variable	
COOL	Country of Origin Labeling	COOL = 1 from the period this law was issued by the US Government for imported food products (1 January 2004), and 0 for otherwise	Dummy variable
WTO	WTO	WTO = 1 from the period after January 2007, and 0 for the period before	Dummy variable
ANTI	Antidumping tariff	ANTI = 1 for the period after the antidumping law of USA was officially imposed for Vietnamese shrimp products (December 1 2003), and 0 for otherwise	Dummy variable
HACCP	HACCP standard	Dummy variable. The value is 1 for the period after HACCP standard was imposed for Vietnamese shrimp products (January 2000), and 0 for otherwise	Dummy variable
GAP	Global gap	GAP = 1 for the period after Vietnamese shrimp processors applied for Global GAP (September 2007), and 0 for otherwise	Dummy variable
SQF	Safety Quality Food standards	SQF = 1 for the period after Vietnamese shrimp processors are required to meet the SQF standards for their products exported to the USA (July 2004), and 0 for otherwise	Dummy variable
EQ	Discovery of enrofloxacin antibiotics residue in Vietnamese shrimp products exported to global markets	EQ = 1 for the periods from January 2011 and afterward EQ = 0 for otherwise	Dummy variable
EMS	Early mortality syndrome	EMS = 1 from September 2011 and afterward, since the first case of EMS occurred for shrimp products EMS = 0 for otherwise	Dummy variable

Aumachine	Auto machine and technology started to apply to Vietnamese shrimp production to change from extensive farms to intensive farms	Aumachine = 1 from January 2009 and afterward
		Aumachine = 0 for otherwise
			Dummy variable

Table 1 (continued)

Name	Definition	Description	Sources
Cir_03	Circular 03/2011, issued by Vietnamese Government to specify the regulation of traceability for shrimp products in Vietnam	Cir-03 = 1 from January 2011 and afterward, Cir_03 = 0 for otherwise	Dummy variable
ASC	Aquaculture Steward Council certification: an international quality assurance applied for shrimp farms by the USA and EU from 2014	ASC = 1 from January 2014 and afterward, ASC = 0 for otherwise	Dummy variable
VietGAP	Vietnam Good Aquaculture Practices: quality assurance certificate issued by Vietnamese Government for shrimp products	VietGAP = 1 from January 2015, and afterward VietGAP = 0 for otherwise	Dummy variable
LENT	The vegetarian date in the USA, that is, every Monday of October	LENT = 1 for the LENT dates, LENT = 0 for otherwise	Dummy variable
Q1	First quarter of year	Q1 = 1 for first quarter of year, 0 = other quarters	Dummy variable
Q2	Second quarter of year	Q2 = 1 for the second quarter of year, 0 = other quarters	Dummy variable
Q3	Third quarter of year	Q3 = 1 for the third quarter of year, 0 = other quarters	Dummy variable

products that have been labeled with those certifications demonstrate the commitment of Vietnamese shrimp producers to safety, quality, and traceability. The shrimp products certified by the GLOBAL GAP certification are more likely to achieve the HACCP and ASC standards, and vice versa (Dong et al. 2019). Those certifications will increase consumer confidence in products; hence, the price of certified products will be higher than that of noncertified products. Thus, the issuing of safety and quality assurance certification for Vietnamese shrimp products not only acts as a passport for acceptance by the US and global markets but will also increase the long-term price of Vietnamese shrimp products (Suzuki and Nam 2018). Therefore, in our study, dummy variables have been added to the model to test the effects of these safety requirements on the price of Vietnamese shrimp products imported to the USA.

In the short term, these requirements are estimated to reduce the quantity of shrimp products imported to the US market, such that domestic Vietnamese shrimp products will be in a surplus. This, also, led to a decrease in the price of shrimp products exported to the US market, ceteris paribus (Linda and Barry 1998; Dong and Duc 2012). Anders and Caswell (2009) indicated that in the longer term, export countries that were able to adjust their production processes to satisfy the requirements of importing countries would ultimately see an improvement in export price and attain comparative advantages in the global marketplace.

In addition, the Vietnam Directorate of Fisheries issued a national traceability regulation, namely Circular No.03/2011/BNN-PTNT (Cir.03), in March 2011, and introduced the mandatory Vietnam Good Aquaculture Practices (VietGAP) in 2012 in an attempt to enhance traceability and manage the quality and safety of farmed shrimp products during distribution and processing. These regulations and practices are meaningful responses from the Vietnam Government that will ensure the eligibility of certified shrimp products into the USA and enable them to meet global market requirements, suggesting a positive effect on the price of Vietnamese shrimp products imported to the US market.

Antibiotic residues are one of the most important issues when evaluating the quality and safety of Vietnamese shrimp products in global markets. In January 2011, the first identification of antibiotic residues was confirmed in Vietnamese shrimp products in Japan. This incident impacted the reputation of Vietnamese shrimp products in global markets, and strict custom inspections were subsequently imposed in the import markets, including the USA (Suzuki and Nam 2003, 2018).

The acceptance of Vietnam into the WTO in 2007 was also added to the model to test its impact on the import price of Vietnamese shrimp products in the US. Factors related to production procedures, including diseases (i.e., early mortality syndrome, EMS) and the application of new technology to shrimp production in Vietnam, are estimated to have affected the quantity of production in the country. This has influenced the local price of shrimp farming inputs in Vietnam, which will in turn influence the price of Vietnamese shrimp products in global markets. Furthermore, seasonal influences were included as drivers of the supply and demand of Vietnamese shrimp products in the model.

All variables were standardized to ensure they had the same scale. The standardization procedure was mainly based on the mean and standard deviation of independent variables. This process made sense in terms of scaling the difference between the "spread" of the independent and dependent variables (Bring 1994).

Machine learning algorithm

The random forest and gradient boosting algorithms were applied to predict the price in selected base periods. The price of Vietnamese shrimp in the coming month was predicted from past data in four base periods: (1) previous 2 months, (2) previous 3 months, (3) previous 6 months, and (4) previous 12 months. The four predictions were used to indicate the trend in Vietnamese shrimp price from the short to long term, which enabled a determination of how long the price would be affected by nontariff factors, such as economic certificates and prices in other countries.

To select the subset that was most strongly related to the output response, the Akaike information criterion (AIC) score was used to identify potentially informative variables. This technique is based on an in-sample fit to estimate the likelihood of a model to predict/estimate future values (Akaike 1998). It can test how well a model fits a dataset without overfitting it. The AIC score was calculated on the basis of the independent variables to compare all models. The model with the lowest AIC score was considered optimal because of the balance between its ability to fit the dataset and its ability to avoid overfitting the dataset. In the first step, the AIC score for the model consisting of all 33 variables was calculated. The backward elimination process was used to remove variables from this model; variables that caused the AIC to become lower were eliminated from the model. This process is repeated until the model has an AIC score that is not lower than the one in the previous step, or until the preset number of variables has been reached. The best AIC score based on the random forest consisted of 15 variables, as presented in Table 2, while the best subset based on the AIC score for gradient boosting is presented in Table 3.

Table 2 The subset chosen by the random forest algorithm based on the AIC

No.	Variable
1	LagPVietnam
2	Yus
3	ANTI
4	HACCP
5	Aumachine
6	Cir_03
7	Q1
8	Q3
9	Price India
10	LENT
11	LagYus
12	LagPIndonesia
13	ASC
14	GAP
15	WTO

Table 3 The subset chosen by the gradient boosting algorithm based on the AIC

No.	Variable
1	LagPVietnam
2	Yus
3	LagPChina
4	LagPIndia
5	Price Thailand
6	Q1
7	Q3
8	Price Indonesia
9	Price Ecuador
10	Price India
11	LENT
12	PriceChile
13	LagPEcuador
14	LagExchange
15	LagPIndonesia

Random forest

This algorithm trained data for specific periods to learn how to predict price, i.e., 2, 3, 6, and 12 months. The random forest algorithm is based on a decision tree, which uses the shape of a tree to predict target values from input variables. The root node and multiple internal nodes are the inputs, while each leaf is an output. The random forest builds multiple decision trees and merges them to obtain an accurate and stable prediction. A random vector value is determined for each tree in the forest (Breiman 2001). In each tree, the internal nodes represent the value of features, while the leaf nodes are a label. The random forest algorithm selects samples randomly and uses features to build multiple decision trees. The final result is obtained by majority voting from decision trees, and therefore, the random forest is more flexible than a decision tree. Furthermore, the random forest may avoid overfitting because it creates small sub-trees, then combines them using an optimal process. This algorithm is based on a bagging technique that trains many individual models in parallel, with each model trained by a random subset of the data. We used the random forest algorithm supported by the sklearn Python package in this study (Pedregosa et al. 2011).

Parameters were set to increase the reliability of the predictions. The *n_estimator*, which indicates the number of trees in the forest, was set to 1000. The function for measuring the quality of a split in the forest, called "entropy," was set for the parameter *criterion*. The parameter *min_samples_split*, which indicates the minimum number of samples required to split an internal node, was set to 2, 3, 6, and 12, corresponding to 2, 3, 6, and 12 months, respectively. The maximum depth of the tree (*max_depth*) was set to 5. Other parameters, such as *min_samples_leaf, min_weight_fraction_leaf, max_features,* and *max_features, min_impurity_split,* etc., were set to the default values.

According to a time series analysis, a conventional cross validation approach could not be used, and therefore, the dataset (288 samples) was split into two subsets: 216 samples (75%) from the first time period to establish the model and 72 samples (25%) from the next time period to test the model using period splitting. The testing subset consisted of 72 continuous months during a time sequence of 6 years from May 2013 to May 2019. The mean absolute error (MAE) method was then applied to correct the multivariate linear regression model. The form of the MAE was as follows:

$$\frac{1}{m}\sum_{i=1}^{m}\left|y_i-\hat{y}_i\right| \tag{1}$$

where m is the number of test samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value. The formula was used to determine the average deviation between the actual and predicted values.

To assess how large the deviation was between the actual and predicted values, we also applied the mean square error (MSE) method. The form of the MSE was as follows:

$$\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2 \tag{2}$$

where m is the number of test samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

Gradient boosting tree

The gradient boosting tree is one of the most powerful machine learning techniques for building predictive models, and is also based on the concept of a decision tree algorithm. With the aim of making weak learners into strong learners, gradient boosting has been widely used in practical applications (Natekin 2013; Freund and Schapire 1996). It creates a decision tree in which each sample has an equal weight. After evaluating the first tree, the weights of each sample that are difficult to classify and lower than the preset weight are increased. The second decision tree is continuously built on this weighted data. The prediction of the first tree is improved in this way. The combination between the first and second trees will generate a new model, and this new model can be used to build the next model through the same process until the specified number of iterations is reached. The prediction of the final model is the sum of the predictions of previous tree models. This algorithm is based on a boosting technique that trains a group of individual models in a sequential way, and each individual model learns from the mistakes made by the previous model. Here, we used the gradient boosting tree approach supported by the sklearn Python package (Pedregosa et al. 2011).

The parameters in this algorithm were set to improve prediction accuracy. Similar to the random forest, *n_estimators* indicated the number of trees and was set to 1000. The parameter *max_depth* was set to 5. The loss function used was least squares regression. The parameter *min_samples_split* also was set to 2, 3, 6, and 12, corresponding to 2, 3, 6, and 12 months, respectively. The parameter subsample used to control variance and bias was set to 1. Other parameters, such as *alpha, max_features,* and *min_impurity_split,* were set to their default values.

As in the process used for the random forest model, the original dataset was divided into two subsets: a training subset (75%) and a testing subset (25%) that consisted of 216 and 72 monthly samples over a 6-year period (from May 2013 to May 2019), respectively.

The training subset was processed according to the desired purpose, and for each case one price was predicted from the values of all variables for the 2, 3, 6, and 12 previous months. For example, to predict the price on the basis of the two previous months, the price for March 2000 was determined using the 15 variables for January and February 2000. After applying the algorithm to the processed training data, the predicted model was obtained and used to predict values for the testing subset.

Results

The percentage error of each algorithm in each test case was calculated (for MAE, mean absolute percentage error (MAPE) = (MAE × 100)/average price, and for MSE, mean square percentage error (MSPE) = (MSE × 100)/average price^2). The average price in the dataset was 11.9 USD.

Table 4 Prediction errors for the different models based on the use of data from the previous 2 months

Model	Training error		Testing error	
	MAPE	MSPE	MAPE	MSPE
Random forest	1.34%	0.03%	3.76%	0.25%
Gradient boosting	1.05%	0.01%	3.55%	0.21%

Table 5 Prediction errors for the different models based on the use of data from the previous 3 months

Model	Training error		Testing error	
	MAPE	MSPE	MAPE	MSPE
Random forest	1.50%	0.06%	4.28%	0.31%
Gradient boosting	0.97%	0.02%	4.09%	0.27%

Table 6 Prediction errors for the different models based on the data from the previous 6 months

Model	Training error		Testing error	
	MAPE	MSPE	MAPE	MSPE
Random forest	1.49%	0.05%	4.36%	0.29%
Gradient boosting	0.85%	0.01%	4.42%	0.32%

Table 7 Prediction errors for the different models based on the data from the previous 12 months

Model	Training error		Testing error	
	MAPE	MSPE	MAPE	MSPE
Random forest	1.45%	0.04%	4.37%	0.28%
Gradient boosting	2.01%	0.06%	4.67%	0.33%

Prediction based on the 2 months

For the testing subset, the MAPE of the random forest algorithm was 3.76%, while for the gradient boosting tree the MAPE was 3.55%. The MSPE of the random forest and gradient boosting algorithms was 0.25 and 0.21%, respectively.

For the training subset, the MAPE of the random forest algorithm was 1.34%, while for the gradient boosting tree the MAPE was 1.05%. The MSPE of the random forest and gradient boosting algorithms was 0.03 and 0.01%, respectively; the values are also presented in Table 4.

Prediction based on the previous 3 months

The training subset was processed with one price value predicted by all features. Then, the model obtained from the algorithms was used to predict the testing subset.

The percentage errors are given in Table 5. The MAPE was better for the gradient boosting tree than for the random forest in the training subset, with values of 0.97 and 1.50%, respectively. The MSPE for the gradient boosting tree was less than that for the random forest, with values of 0.02 and 0.06%, respectively. In the testing subset, the gradient boosting model was slightly more accurate than the random forest. Its absolute percentage error was 4.09% compared with 4.28% for the random forest. The MSPE for gradient boosting tree (0.27%) was smaller than that for the random forest (0.31%).

Prediction based on the previous 6 months

The training subset was processed to predict the price based on the previous 6 months. A value for every month was predicted on the basis of the historical data from the previous 6 months. The predicted model was then used to generate the testing subset. The prediction results for the random forest are presented in Table 6. In the testing subset, the gradient boosting tree MAPE result was worse than that of the random forest (4.42% versus 4.36%, respectively). However, the random forest model was less accurate than the gradient boosting tree for the training set, with MAPE values of 1.49 and 0.85%, and MSPE values of 0.05 and 0.01%, respectively.

Prediction based on the previous 12 months

The training set was processed so that the price of every month was predicted by the previous 12 months. It was considered that a period of 1 year was sufficiently long to evaluate the effect of historical data on the export price, as well as to evaluate the effect of various external factors on the price. After the training model was established, the testing set was applied to evaluate the predicted model. In the testing set, the MAPE for the random forest was 4.37%,

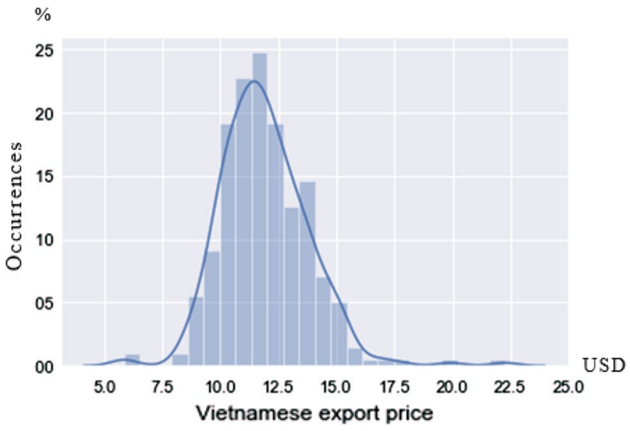

Fig. 1 Distribution of the export prices of Vietnamese shrimp

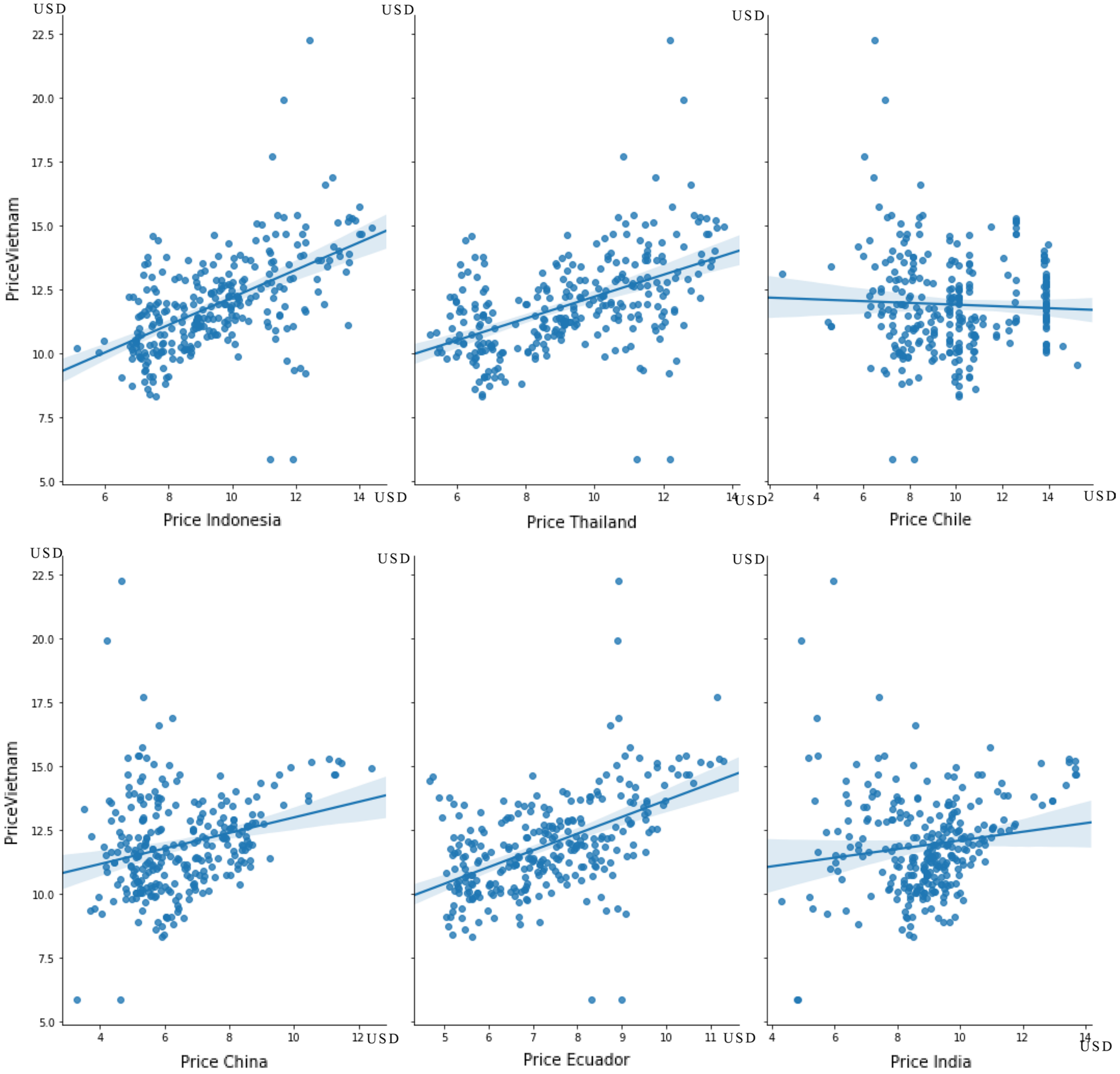

Fig. 2 Correlation between the Vietnamese export price and that of other countries

compared with 4.67% for the gradient boosting tree, and the MSPE was 0.28%, compared with 0.33% for the gradient boosting tree. In the training set, gradient boosting was prone to more error than the random forest, with values of 2.01 and 1.45% (for MAPE), and 0.06 and 0.04% (for MSPE), respectively, which are shown as in Table 7.

A comparison of the predicted values when using the different periods with the actual values in the 6 years (May 2013 to May 2019) using the random forest and gradient boosting algorithms is shown in Figs. 3 and 4, respectively.

Discussion

From the dataset, the price of Vietnamese exported shrimp was mainly distributed around 10.00–12.50 USD, as shown in Fig. 1. This was higher than for other countries (e.g., Indonesia, Thailand, Chile, Ecuador, China, and India), as shown in Fig. 2. The price changed frequently under the influence of factors such as the economic policies imposed by governments and competitive pricing from other countries. The prediction procedure was applied using two robust tree-based models, i.e., the random forest and gradient boosting algorithms, to obtain the predicted model and understand the

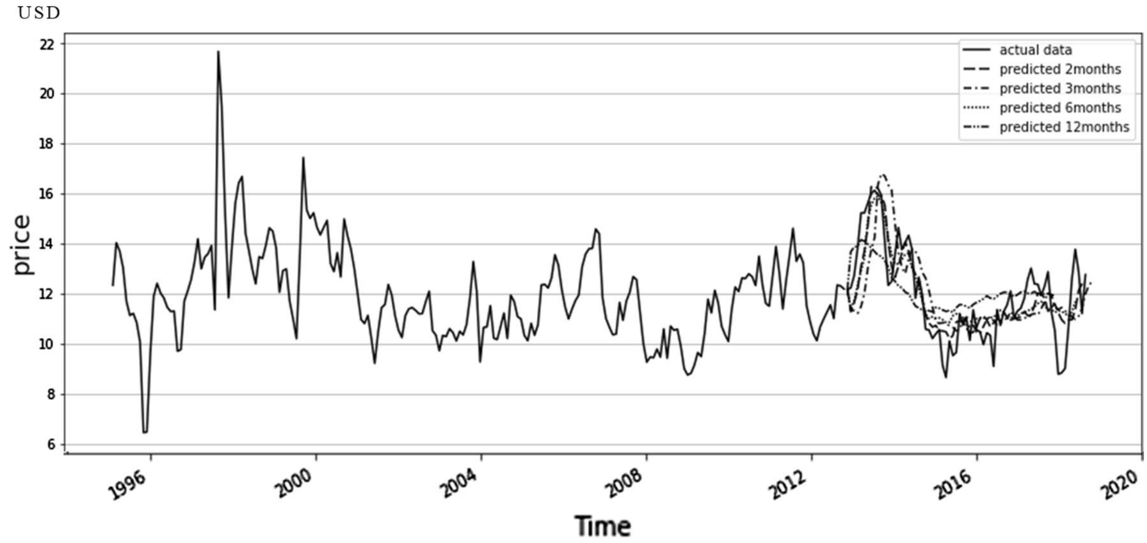

Fig. 3 Comparison of the predictions made by the random forest algorithm using data for different base periods

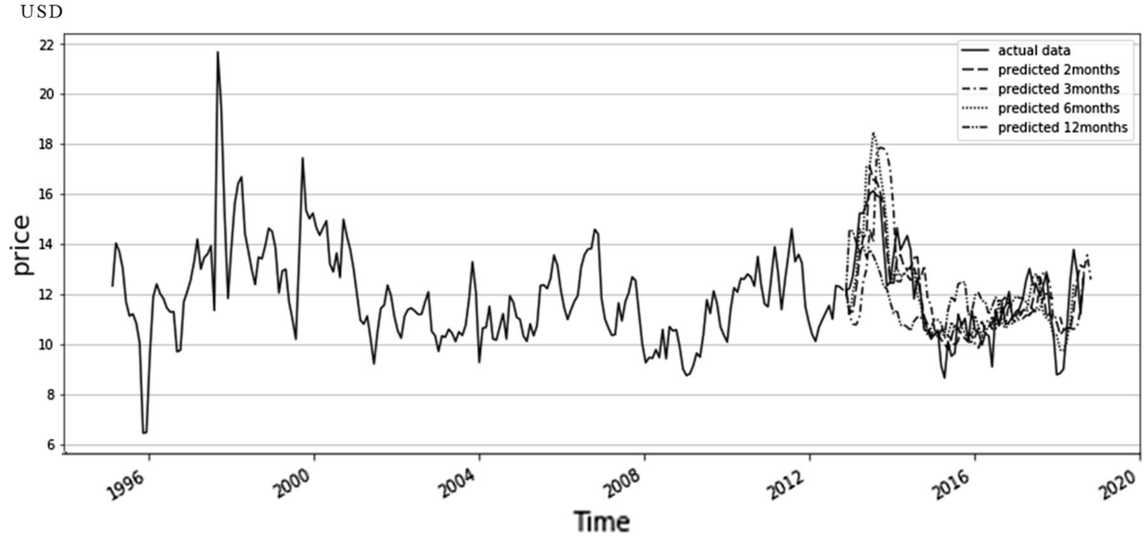

Fig. 4 Comparison of the predictions made by the gradient boosting algorithm using data for different base periods

export market trends for frozen shrimp (Figs. 3 and 4). The error was less than 5.00% for MAPE and less than 0.04% for MSPE for all testing cases, although it gradually increased from the short- to long-term testing periods. The prediction yielded a high accuracy (small error percentage) for the 2-month base period, which then decreased for the 3-month base period. It was slightly less accurate when using the 6-month base period and had the lowest prediction accuracy for the 12-month base period. This was likely because predictions based on long-term testing would project historical conditions that were no longer valid for the current situation. The 12-month base period was the longest period for which changes in relevant factors (economic certificates, requirements for safe food, trade laws, and competition from other countries) would be likely to affect the export price. These factors all had a strong impact, but this lessened over long time periods, which led to a low prediction accuracy. Similarly, the prediction based on the 6-month base period was also long enough to be affected by the fluctuation of associated factors. This was the main reason that the model had a low prediction accuracy. The predicted model based on a 2-month base period achieved a good result, with small errors for both algorithms. This indicated that the changes of any fairly stable factor would have a direct effect on the price only in the short term because producers were not able to change the resources and/or investments required for production. Conversely, the export price was probably adjusted in the long term because shrimp producers were

Table 8 Comparison of the error percentage (MAPE) among the full set and subsets for the two algorithms

Base period	Algorithm	All variables (original)	Subset of random forest	Subset of gradient boosting
2 months	Random forest	4.01	3.75	
	Gradient boosting	3.65		3.55
3 months	Random forest	4.06	4.27	
	Gradient boosting	4.67		4.09
6 months	Random forest	4.04	4.36	
	Gradient boosting	4.28		4.42
12 months	Random forest	4.76	4.37	
	Gradient boosting	4.48		4.67

able to improve their processes and acquire resources in response to production issues that may arise. To overcome any difficult production scenarios, Vietnamese producers want to increase the export price and increase the quantity of shrimp products entering the US market. For example, it is known that the Vietnamese price decreased when the requirement for an SQF certificate was imposed on producers (April 2004). Then, Vietnam joined the WTO, and the price increased again from January 2005. The appearance of EMS in September 2011 was also found to be a positive factor that increased the price of Vietnamese shrimp products. The average price before EMS appeared was 11.78 USD, which increased to 12.26 USD after EMS was observed. It is estimated that the disease reduced the production volume. However, the Vietnamese fisheries met the challenge this imposed on them by improving their product and guaranteeing the amount of shrimp produced for export. Since Vietnam joined the WTO (January 2007), the export volumes and prices of shrimp product have increased in comparison with other countries (the Vietnamese price was 11.78 USD in December 2006, which increased to 13.00 USD in January 2007, while for other countries the price decreased). However, the HACCP was then applied (January 2008), which was a drawback for the Vietnamese shrimp industry. The price was reduced from 11.46 USD in December 2007 to 10.69 USD when this factor was applied in January 2008.

The random forest and gradient boosting tree are powerful algorithms that are used in machine learning to make predictions. The gradient boosting tree achieved better prediction results for base periods of less than 6 months, while the random forest outperformed the gradient boosting tree for base periods of more than 6 months. Both algorithms have an optimal structure based on the tree principle and they avoid overfitting by using multiple decision trees to obtain results. In this study, each algorithm used 1000 subtrees to predict the result, and every sub-tree selected a small subset of independent variables to operate (for the random forest) or to learn from previous mistakes (for the gradient boosting tree). They could be applied reliably to subsets consisting of meaningful variables. Weak variables that made little contribution were ignored by the algorithms, as indicated by the trivial difference in accuracy between predictions made using all variables and selected subsets (Table 8). The random forest was able to solve long-term problems, while the gradient boosting tree was more robust over the short term. The mechanism of each algorithm determined

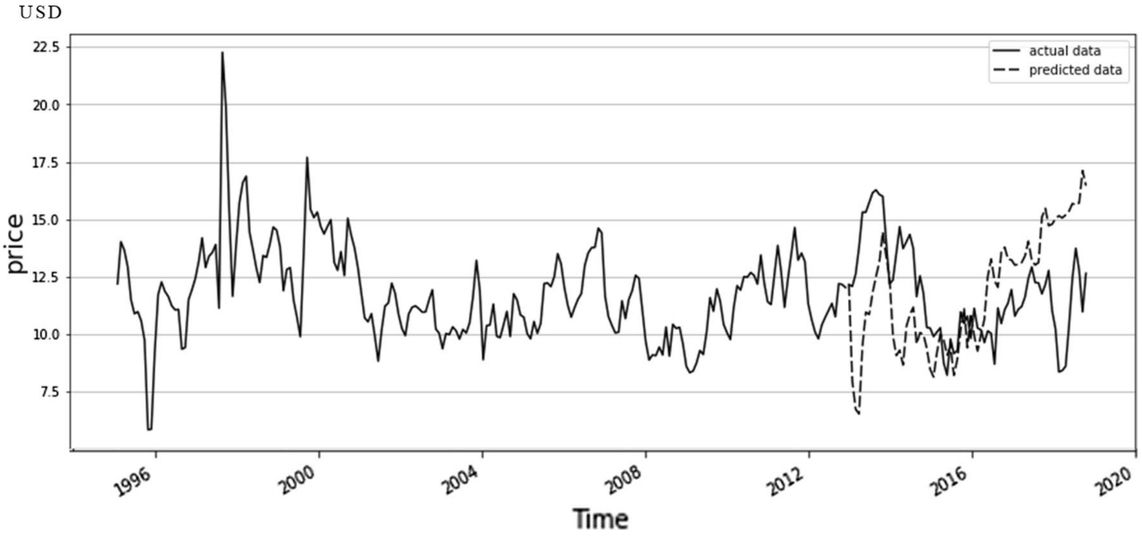

Fig. 5 Comparison between the predicted and actual values for the 2-month base period based on a linear regression

the way predictions were made, with the random forest creating multiple submodels at the same time and achieving the best results. Gradient boosting uses a step-by-step procedure to create models and ensure that the following model could fix the errors of the previous model. The predictions of other algorithms, such as a neural network and the K-nearest neighbor, were found to be less accurate (MAPE was greater than 7.00%) in our study. There was no significant merit in using a linear regression to achieve the predicted result (Fig. 5). Furthermore, ARIMA was not suitable because it does not work well for our situations with multiple variables.

Vietnamese shrimp producers have achieved product quality certifications to ensure that their shrimp products are safe and free of chemical residues, and to increase the probability to be accepted in the global markets (Dong et al. 2021). Automation and technological advances have been applied to guarantee that food products are wholesome. This is necessary to overcome the strong competition from other countries, including Thailand, Chile, Indonesia, Ecuador, India, and China, that also export fishery products. Accurate predictions will enable Vietnam to plan important actions to reduce the effect of negative economic factors, competition, and price dumping. The development of an intensive shrimp farming industry to improve the quality and increase the quantity of shrimp products has been considered in Vietnam because product quality can be easily controlled.

Our study could be extended to predict the trends in international markets for other seafood products. Vietnam has many potential exported fishery products, such as catfish, marine fish, mollusks, frozen seafood, and dried seafood. There is also the potential to develop further export items, including tuna, clams, and some other marine specialties. Accurate price predictions will enable the development of these exporting industries and generate more benefits for Vietnamese fisheries. By understanding market trends through such predictions, the importance of price dumping laws and food safety criteria can be evaluated. Then, Vietnamese products can satisfy the export markets where high standards are in place. This is an important issue in the development of the fishery industry, which significantly contributes to the GDP of Vietnam and currently employs millions of people. Many social issues could be resolved through the development of fisheries, especially poverty and education provision. The more seafood products that are exported, the more the fishery industry will benefit. This has motivated Vietnamese producers to develop and investigate the application of technology to shrimp farming. The success of seafood exports will encourage the establishment of a Vietnamese trademark for use in international markets.

Acknowledgements This study is funded in part by the Can Tho University Improvement Project VN14-P6, supported by a Japanese ODA loan. We also are grateful for the advice from anonymous reviewers, who have helped the authors develop this paper.

References

Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York

Anders SM, Caswell JA (2009) Standards as barriers versus standards as catalyst: assessing the impact of HACCP implementation on U.S. seafood import. Am J Agric Econ. https://doi.org/10.1111/j.1467-8276.2008.01239.x

Bailey M, Packer H, Schiller L, Tlusty M, Swartz W (2018) The role of corporate social responsibility in creating a Seussian world of seafood sustainability. Fish Fish. https://doi.org/10.1111/faf.12289

Brander K (2003) What kinds of fish stock predictions do we need and what kinds of information will help us make better predictions? Sci Mar. https://doi.org/10.3989/scimar.2003.67s121

Breiman L (2001) Random forests. Mach Learn. https://doi.org/10.1023/A:1010933404324

Bring J (1994) How to standardize regression coefficients. Am Stat. https://doi.org/10.2307/2684719

COFI (2019) Fishery and aquaculture country profiles: the Socialist Republic of Viet Nam. FAO, Rome

Dong KTP, Duc NM (2012) The impacts of non-tariff barriers on the export price of Vietnamese catfish. In: Toan HT et al (eds) IFS 2012: Sharing knowledge for sustainable aquaculture and fisheries in the Southeast Asia. Agriculture Publishing House, CanTho City, Vietnam, pp 315–326

Dong KTP, Saito Y, Hoa NTN, Dan TY, Matsuishi TF (2019) Pressure–State–Response of traceability implementation in seafood-exporting countries: evidence from Vietnamese shrimp products. Aquacult Int. https://doi.org/10.1007/s10499-019-00378-2

Dong KTP, Matsuishi TF, Duc NM, Hoa NTN, Saito Y, Dan TY (2021) Does application of quality assurance certification by shrimp farmers enhance feasibility of implementing traceability along the supply chain? Evidence from Vietnam. J Appl Aquac. https://doi.org/10.1080/10454438.2020.1856751

Duc NM (2009) Economic contribution of fish culture to farm income in Southeast Vietnam. Aquacult Int. https://doi.org/10.1007/s10499-008-9176-8

Duc NM, Kinnucan HW (2007) Effects of antidumping duties with bertrand competition: some evidence for frozen catfish fillets. Agric Appl Econ Assoc. https://doi.org/10.22004/ag.econ.9893

Dudek G (2015) Short-term load forecasting using random forests. In: Filev D et al (eds) Intelligent Systems'2014 Advances in Intelligent Systems and Computing. Springer, Cham. https://doi.org/10.1007/978-3-319-11310-4_71

The English in this document has been checked by at least two professional editors, both native speakers of English. For a certificate, please see: http://www.textcheck.com/certificate/GgCgWU

FAO (2018) The State of Food and Agriculture (2018) Migration, agriculture and rural development. FAO, Rome

Flaaten O (2018) Fisheries and aquaculture economics, 2nd edn. Aquaculture: plant and industry management, Norway

Freund Y, Schapire R (1996) Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, ***, pp 148–156

GSO (2018) Agriculture forestry and fishing. The General Statistics Office of Vietnam, Vietnam

Ha TTT, Bush SR (2010) The transformations of Vietnamese shrimp aquaculture policy: empirical evidence from the Mekong Delta. Environ Plann C Gov Policy. https://doi.org/10.1068/c09194

International Trade Center (2020) Trade statistics for international business development. International Trade Center, Switzerland

Kamil S, Barbara K, Wieslaw F, Witold R, Katarzyna S, Katarzyna K (2020) Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models. Urban Water J. https://doi.org/10.1080/1573062X.2020.1734947

Khiem NM, Takahashi Y, Oanh DTH, Hai TN, Yasuma H, Kimura N (2020) The use of machine learning to predict acute hepatopancreatic necrosis disease (AHPND) in shrimp farmed on the east coast of the Mekong Delta of Vietnam. Fish Sci. https://doi.org/10.1007/s12562-020-01427-z

Leung PS, Tran TL (2000) Predicting shrimp disease occurrence: artificial neural networks vs. logistic regression. Aquaculture. https://doi.org/10.1016/S0044-8486(00)00300-8

Linda C, Barry K (1998) Technical barriers to trade: a case study of phytosanitary barriers and U.S.-Japanese apple trade. J Agric Resour Econ. https://doi.org/10.22004/ag.econ.31191

Manuel M, Jose MB, Maria GP (2012) Tree-based machine learning analysis for fisheries research. Fishery Management, Chapter: Tree-based machine learning analysis for fisheries research. Nova Science Publishers Inc., pp 61–75

Natekin A, Knoll A (2013) Gradient boosting machines. Front Neurorobot 7:21. https://doi.org/10.3389/fnbot.2013.00021

Pavlyshenko BM (2019) Machine-learning models for sales time series forecasting. Data. https://doi.org/10.3390/data4010015

Pedregosa et al (2011) Scikit-learn: machine learning in Python. JMLR 12:2825–2830

Phuong NT, Oanh DTH (2010) Stripped catfish aquaculture in Vietnam: a decade of unprecedented development. In: De Silva SS, Davy FB (eds) Success stories in Asian aquaculture. Springer, Dordrecht, Netherlands, pp 131–147

Portley N (2016). Report on the shrimp sector: Asian shrimp trade and sustainability. Asian Shrimp Trade and Sustainability. Sustainable Fisheries Partnership, pp 1–74

Rahman A, Tasnim S (2014) Application of machine learning techniques in aquaculture. Int J Comput Trends Technol. https://doi.org/10.14445/22312803/IJCTT-V10P137

Shahriar MS (2014) A dynamic data-driven decision support for aquaculture farm closure. Procedia Comput Sci. https://doi.org/10.1016/j.procs.2014.05.111

Soykan CU, Eguchi T, Kohin S, Dewar H (2014) Prediction of fishing effort distributions using boosted regression trees. Ecol Appl. https://doi.org/10.1890/12-0826.1

Stoll F (2020) A Comparison of Machine Learning and Traditional Demand Forecasting Methods. All Theses. 3367. Clemson University, South Carolina

Suzuki A, Nam VH (2013) Status and constraints of costly port rejection: a case from the Vietnamese frozen seafood export industry. IDE-JETRO. Discussion Paper No.395

Suzuki A, Nam VH (2018) Better management practices and their outcomes in shrimp farming: evidence from small-scale shrimp farmers in Southern Vietnam. Aquacult Int. https://doi.org/10.1007/s10499-017-0228-9

Tran VN, Bailey C, Wilson N, Phillips M (2013) Governance of global value chains in response to food safety and certification standards: the case of shrimp from Vietnam. World Dev. https://doi.org/10.1016/j.worlddev.2013.01.025

Tucker BI (2008) Chapter 3: Market Demand and Supply. Economics for Today's World, 5th edn. Thomson South-Western. Transcontinental-Beauceville Quebec, Canada, pp 52–57

VASEP (2018) Report on Vietnam Shrimp Sector 2009–2018. Vietnam Association of Seafood Exporters and Producers. Vietnam

VASEP (2020) Report on Vietnam Seafood Exports 2020. Vietnam Association of Seafood Exporters and Producers. Vietnam

Yu R (2006) Predicting shrimp growth: artificial neural network versus nonlinear regression models. Aquacult Eng. https://doi.org/10.1016/j.aquaeng.2005.03.003