Print ISSN: 2765-6934 / Online ISSN: 2765-7027
AJBE website: http://www.ajbe.or.kr/
Doi: 10.13106/ajbe.2022.vol12.no3.1

Export-Import Value Nowcasting Procedure Using Big Data-AIS and Machine Learning Techniques

Jimmy NICKELSON[1], Rani NOORAENI[2], EFLIZA[3]

Received: September 12, 2021. Revised: November 21, 2021. Accepted: July 05, 2022.

Abstract

Purpose: This study aims to investigate whether AIS data can be used as a supporting indicator or as an initial signal to describe Indonesia's export-import conditions in real-time. **Research design, data, and methodology:** This study performs several stages of data selection to obtain indicators from AIS that truly reflect export-import activities in Indonesia. Also, investigate the potential of AIS indicators in producing forecasts of the value and volume of Indonesian export-import using conventional statistical methods and machine learning techniques. **Results:** The six preprocessing stages defined in this study filtered AIS data from 661.8 million messages to 73.5 million messages. Seven predictors were formed from the selected AIS data. The AIS indicator can be used to provide an initial signal about Indonesia's import-export activities. Each export or import activity has its own predictor. Conventional statistical methods and machine learning techniques have the same ability both in forecasting Indonesia's exports and imports. **Conclusions:** Big data AIS can be used as a supporting indicator as a signal of the condition of export-import values in Indonesia. The right method of building indicators can make the data valuable for the performance of the forecasting model.

Keywords : export-import in Indonesia, AIS Data, forecasting, ANN, ARIMA

JEL Classification Code : C18, C32, C53, F17

1. Introduction

Economic growth is an important thing that needs to be considered by various parties. This is because economic growth is an indication of the success of economic development in a region (Dewi & Wulansari, 2021). One of the efforts to monitor the increase in international trade is to pay attention to the growth of exports and imports. Improved export-import performance can have an impact on increasing economic growth. This means that exports and imports have an important role in the economy of a country. The availability of export-import data is very important for decision-making regarding foreign trade policies by the government.

In Indonesia, international goods trade data are collected by Statistics Indonesia or Central Bureau of Statistics (BPS) obtained from compilations and surveys. The main data is collected based on export-import information documents produced by the Directorate General of Customs and Excise every month. Export data also comes from PT Pos Indonesia, other records on the border, and the results of a cross-border trade survey (Badan Pusat Statistik, 2020). However, in the process of publishing the official statistics, there is a time lag from the

* Authors would like to thank to Politeknik Statistika STIS, BPS Statistics Indonesia, UNSD for their support of this research.

[1] First Author, Politeknik Statistika STIS, Indonesia. Email: 221709765@stis.ac.id

[2] Second Author and Corresponding Author, Assistant Professor, Department of Applied Statistics, Politeknik Statistika STIS, Indonesia. Email: raninoor@stis.ac.id

[3] Third Author, Director of the Directorate of Distribution Statistics, BPS-Statistic Indonesia. Email: efliza@bps.go.id

data collection until data publication, resulting in delays in the publication of official statistics. For example, data published in May 2021 is temporary data for March 2021.

Publicity delays can occur due to several factors: the complexity of the data collection stages, the complexity of compiling the results of several surveys, and the length of data processing required. Therefore, the data collected will not be released or disseminated at that time. To help reduce these problems, forecasting or nowcasting is needed to describe the current situation and to give an early signal for monitoring these statistics.

There are challenges in predicting the current condition of Indonesia's export-import value. One of them is to obtain supporting information that can describe the current state of export and import. In addition, the supporting information can produce a fairly good predictive accuracy of export-import data. However, it is difficult to obtain or define indicators and information that can fulfill both conditions.

The Automatic Identification System (AIS) is an automatic tracking system that uses transceivers on vessels and is used by vessel traffic services (VTS). AIS data are periodically transmitted by vessels and received by AIS terrestrial stations or via satellite receivers. Because they are sent continuously, AIS data can be used to monitor seaborne trade patterns in real-time.

Several previous studies have tried to use the ais data as a new data source, one of which is to use the ais data to predict crude oil exports (Adland, Jia, & Strandenes, 2017). Other studies also use AIS data to develop indicators of trade and maritime activity in Malta (Arslanalp, Marini, & Tumbarello, 2019) or indicators of world seaborne trade using raw AIS data (Cerdeiro, Komaromi, Liu, & Saeed, 2020). The AIS task team on the UN Global Platform also analyzes AIS data for use in various research fields, such as migration, environment, maritime and fisheries, as well as economics and trade activities (UN Global Working Group, 2019). From these studies, it can be seen that AIS has a great potential and can be used as a new data source, such as to develop new indicators of international trade, which includes export and import activities. However, neither of these studies has the objective of predicting export-import statistics. Also, there are not many studies using AIS in Indonesia. Especially those that use AIS data as an auxiliary variable in forecasting export and import statistics. Indonesia is a maritime and archipelagic region where most of the trading activities occur in the maritime area. Therefore, it is important to identify and explore AIS data, for which the data is more up-to-date.

As an alternative solution, AIS can be a new indicator and a proxy variable for forecasting export-import statistics. AIS data is transmitted by the vessels every minute, so it can be made available in real-time. This makes AIS data potential to be used in nowcasting. In addition, most of Indonesia’s export-import commodities use sea transportation (Badan Pusat Statistik, 2020), so that AIS data becomes relevant for use in forecasting export-import statistics in Indonesia.

In general, forecasting time series data can use traditional statistical methods or machine learning methods. The ARIMA model is one of the popular traditional statistical methods and produces good predictions when used for linear patterned data (Rahkmawati, Sumertajaya, & Aidi, 2019). Meanwhile, machine learning methods, such as Artificial Neural Network (ANN), can recognize patterns from past data, even though there are nonlinear patterns or noise in the data (Neves & Cortez, 1998). In addition, ANN is also more flexible to use than traditional methods (Zissis, Xidias, & Lekkas, 2016).

Based on this background, the objectives of this study are the following: first, explore the use of AIS data by performing the appropriate preprocessing steps stated in this paper; second, formulate the formation of composite indicators that can describe export-import statistics using clean AIS attribute data; third, apply several statistical methods to get the best predictors that can produce the smallest forecasting error; and finally, compare the results of export-import forecasting between conventional statistical forecasting methods and machine learning techniques.

2. Literature Review

2.1. Automatic Identification System (AIS)

AIS is an international maritime communication system that is transmitted by each vessel and is used to track the movement of vessels. AIS was introduced by IMO (International Maritime Organization) in 2004 to aid vessels to avoid collisions and to help port authorities control sea traffic efficiently. All major international shipping vessels (over 300 gross tonnages) and all passenger ships are required to install AIS transceivers on their ships (International Maritime Organization, 2021). AIS messages are automatically sent via Very High Frequency (VHF) radio waves equipped with a GPS (Global Positioning System) every two to ten seconds when the ship is moving, or every six minutes when the ship is stationary (Arslanalp, Marini, & Tumbarello, 2019).

2.2. Variable Selection Methods

Variable selection is one of the important stages in building a predictive model. The focus of this method is on selecting several predictor variables that can describe the overall input data efficiently and still provide good predictive results for the output data, so it can build a simpler and more comprehensive model (Ahani, Salari, & Shadman, 2019). In this study, the variable selection methods used are permutation importance, forward stepwise selection, and correlation value.

Permutation importance measures the importance levels of a variable by shuffling the value of the variable. If the error of a model increases after the variable is shuffled, then the variable is classified as important in the forecasting model because the model requires it to make predictions (Ahmed, Cui, Fu, & Chen, 2021). Furthermore, these variables are selected as predictor variables in the forecasting model. There is also a consideration to exclude the variable if the randomization results in a smaller error.

Forward stepwise selection chooses variables by considering the increase in a criterion for each variable in the forecasting model. This method starts with a model without variables. Then for each iteration, the variable that gives the greatest increase in the criteria to the model will be included in the forecasting model. However, there are also considerations to exclude variables if they don't provide an improvement in the model (Ahani, Salari, & Shadman, 2019).

Variable selection methods are also done based on significant variables in a model, especially in a traditional statistical model. Variables that have a p-value less than the level of significance (α=0.05) in the model will be selected and included in the model. Variable selections are also done based on the correlation value between the predictor variables and the response variables. Variables that have a high correlation will have a linear relationship and have almost the same effect as the response variables. Variables that are correlated beyond a threshold value will be selected as predictor variables.

2.3. Pearson Correlation

Correlation is a value that indicates the strength of the relationship between two or more variables. The correlation coefficient on the variables *X* and *Y* on n data can be calculated using a formula such as equation (1):

$$r_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \qquad (1)$$

where *n* is the number of data, *X* is the independent variable or the first variable, and *Y* is the dependent variable or the second variable. The correlation value will have a range of values between -1 to +1. The negative sign (-) indicates that the relationship between the variables is inversely proportional, while the positive sign (+) indicates the opposite direction. The closer the correlation value to number one, the stronger the relationship between these variables.

2.4. Forecasting Methods

2.4.1. AutoRegressive Integrated Moving Average (ARIMA)

ARIMA model is a time-series data forecasting model. This model is a combination of autoregression and moving average models with differencing values, so this model uses information from past observations to make predictions. The equation of this model can be written as equation (2):

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (2)$$

where y'_t is the differenced series of y_t; $y'_{t-1}, \ldots, y_{t-p}$ are the lag value of y_t; and $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ are the lag of error. This model is called the ARIMA model (p,d,q), where *p* is the order of the autoregressive values, *d* is the degree of differencing used, and *q* is the order of the moving averages. The ARIMA model can include other relevant information to improve prediction accuracy, which is done by adding exogenous variables to the model. The model is called the ARIMAX model (Nooraeni, Sari, & Yudho, 2019).

There are three steps to build an ARIMA model. First, identify ARIMA's tentative model order by analyzing past data. Second, estimate the model parameters, and third, perform diagnostic tests on the former model (Montgomery, Jennings, & Kulahci, 2007).

1. Identify the stationarity of the data by visually determining the plot of the data series. Another method is by using statistical tests, such as the Dickey-Fuller (DF) test. If there is a non-stationary indication, the data must be differenced. Once stationary, identify order *p* and order *q* by looking at the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. The ACF plot considers the relationship between y_t and y_{t+k}, while the PACF plot considers the relationship between y_t and y_{t+k} after the linear dependencies $y_{t+1}, y_{t+2}, \ldots$, and y_{t+k-1} are removed.
2. Parameter estimation: There are several methods for estimating model parameters, such as methods of moments, maximum likelihood, and least squares. These methods are used to estimate the parameters in the model that have been tentatively identified in the previous stage. Parameter estimation can be done automatically with statistical software such as Minitab, SAS, and JMP.
3. Diagnostic test: After obtaining a parameter model, the feasibility test of the model was carried out using

residual analysis. If the model formed is adequate with the appropriate *p*, *d*, and *q* values, then the residual must have a white noise pattern, i.e., residuals that are identical (have homogeneous variance) and independent (no correlation between lag residuals). The residual must also be normally distributed and have a stationary pattern.

2.4.2. Artificial Neural Network (ANN)

ANN is a machine learning modeling algorithm that is used to process information and analyze data. ANN is built by following the analogy of the nervous system, which is a set of interconnected units or neurons. ANN has three main layers, i.e., the input layer as a layer for input data, the hidden layer as a layer for processing data, and the output layer as a layer for producing predictions/estimations, as shown in Figure 1. If the ANN architecture has more than one hidden layer, the architecture is called a multilayer feed-forward neural network. Fig 1 shows the process in a neuron.

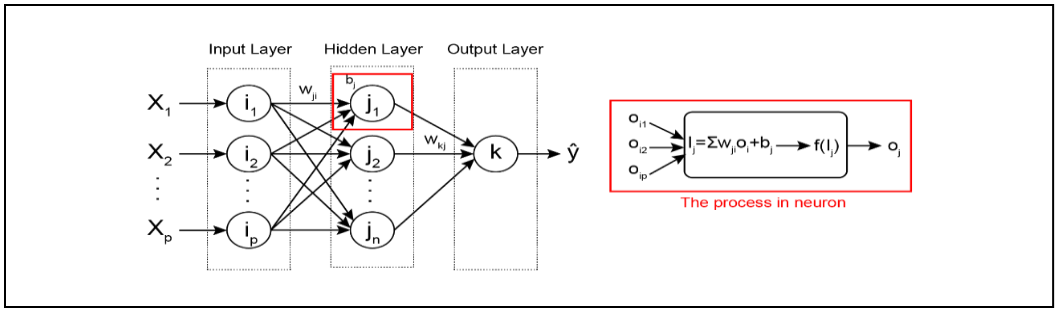


Figure 1: ANN's architectural topology and layer configuration in the *p* input layer, *n* hidden layer neurons, and 1 output variable

One of the ANN learning algorithms is the backpropagation algorithm. This algorithm learns the data by processing the training data repeatedly. For each iteration, the weights will be adjusted to get the smallest possible error between the predicted results and the actual target data, which can be class label data or numeric data. This weight modification is carried out in a backward direction (Han, Kamber, & Pei, 2006).

Table 1: Activation Function in ANN

Function Name	Activation Function f(x)
Linear	*cx*
Sigmoid	$\frac{1}{1+e^{-x}}$
Tanh	tanh(*x*)
ReLU	max (0, *x*)
Leaky ReLU	max (α*x*, *x*)

The ANN method requires parameters to form the model. Some of these parameters are as follows:

1. Numbers of hidden layers and neurons.
2. Activation function, which is a function that converts the input that has been added weight and bias into an output value. This function is used to control the output of the neuron so that the output results can be within the specified boundary. Some examples of activation functions are shown in Table 1.
3. Learning rate, a parameter that controls the amount that the weights are updated during training. This parameter also controls the learning speed of the model.
4. Momentum, a parameter that can smoothen algorithm learning so that it can also speed up the training process.

2.4.3. Parameter Tuning ANN

Determining the ANN parameter values is important because these values can affect the prediction results and model performance. Therefore, parameter tuning is needed to determine the appropriate ANN parameter value. One of the parameter tuning methods is the random search method. Random search is a search algorithm that involves generating and evaluating random values to an objective function, which, in this case, is the ANN function model. The value of parameters in the model will be generated by random numbers, and the error value is calculated. The process is repeated with a certain number of iterations and saves the model that has the smallest error value in each iteration. At the end of the iteration, there will be one model that has the smallest error value and gives the best parameter results (Torres, Gutiérrez-Avilés, Lora, & Martínez-Á lvarez, 2019).

2.5. Forecasting Evaluation Methods

To determine the accuracy of the forecasting results from the model, it is necessary to evaluate the results. Forecasting evaluation methods can also be used to help find an optimal method and compare the accuracy of two or more different modeling methods. There are several techniques to evaluate forecasting results, one of which is using the RMSE (Root Mean Squared Error) value.

The RMSE method calculates the square root of the mean squared difference between the predicted value and the actual value. A model is said to have good performance if the resulting RMSE is minimum. The RMSE function can be written as equation (3):

$$RMSE = \sqrt{\sum_{t=1}^{n}(\frac{\hat{y}_t - y_t}{n})^2} \quad (3)$$

where n is the number of data, $\hat{y}_t$ is the predicted value, and y_t is the actual value.

3. Research Methodology

3.1. Data Collection Methods

In this study, the data that will be used are AIS data and Indonesia's international trade data, which includes the value and volume of exports and imports. The AIS data used is from the AIS data provider exactEarth via UN Global Platform, which is ship movement reports data. The scope of the data used is AIS data sent by vessels inside the Indonesian bounding box area, especially in certain ports listed on the Maritime Safety Information (MSI) website (National Geospatial-Intelligence Agency, 2021), i.e., as many as 123 ports. The data range used is between December 2018 to December 2020, with a total of 35 features and 661.8 million records. The features contained in the AIS data are as shown in Table 2.

Table 2: Features in AIS Data

Features Name	Explanation
MMSI	Vessel unique number, *Maritime Mobile Service Identity* (MMSI)
DTG	Observation date (yyyy-MM-dd'T'HH:mm:ssZ)
Vessel_type	Vessel type, like 'cargo', 'tanker', 'tug', 'passenger', etc
Nav_status	Vessel navigation status, like 'at anchor', 'moored', 'aground', etc
SOG	Speed Over Ground (knots)
COG	Course Over Ground (degree)
Draught	The vertical distance between the waterline and the bottom of the hull (meters)
Longitude	Longitude coordinates in WGS 84 (decimal degrees)
Latitude	Latitude coordinates in WGS 84 (decimal degrees)

Indonesia's international trade data used are export and import data sourced from the BPS website (Badan Pusat Statistik, 2021). The data consists of the value and volume of Indonesia's exports and imports in US dollars and kilograms. The data used is monthly data with the same period, which is between December 2018 to December 2020.

3.2. Data Processing Methods

3.2.1. Data Preprocessing Methods

In AIS data, it is possible to have messages that have noises or outliers. For example, there is a message that has an MMSI that does not match the rules. Some messages are not related to export-import activities, such as messages sent by passenger ships and military ships, among others. For this reason, an appropriate preprocessing stage is needed to reduce noises in AIS data and obtain data related to export and import, as well as to remove features that contain the missing data. The preprocessing stage is done by performing several steps/rules to filter AIS data.

- Filter 1: valid vessel's MMSI
  A valid vessel's MMSI has the number of nine digits (Raymond, 2021), so AIS messages that have an MMSI of less than or more than nine digits, in other words, MMSI < 100000000 or MMSI > 999999999 will be eliminated from the data.
- Filter 2: moving vessels
  Vessels that report positions within a small area over two years are eliminated from the data. These filters are used to track vessels that do not leave port or vessels that only sailed in a small area during this period in which did not carry out the trading activity. Thus, they are excluded from units that contribute to exports and imports (Noyvirt, 2019). The area size used is 0.1 decimal degrees (11.1 km).
- Filter 3: navigation status of anchored vessels
  The navigation status in the AIS data indicates the status of the ship when the message was sent. Several navigation statuses, such as 'under way using engine', 'at anchor', 'moored', 'sailing', 'engaged in fishing', 'not under command', 'restricted maneuverability', 'aground', and 'not defined', are available. The ship's navigational statuses on the AIS data are filtered related to export-import activities; namely, 'at anchor', 'moored', and 'restricted maneuverability'. Vessels with navigational status can unload or transfer cargo from/to ships related to import-export activities (USCG, 2021).
- Filter 4: non-zero draught
  Draught is the distance between the water surface and the bottom of the hull (keel). The vessel's water load can be used to estimate the potential weight of goods transported by vessels (Arslanalp, Marini, & Tumbarello, 2019). The draught value on AIS messages has ranged from 0.1 to 25.5 meters. A value of 25.5 indicates a draught value of 25.5 m and over. Whereas the zero value represents the default value or if it is not available (Raymond, 2021). Then messages that have a

draught value of zero will be removed from the data.

- Filter 5: relevant vessel type
  Types of ships in the AIS data are fishing, passenger, sailing, military, pilot, cargo, tanker, etc. AIS data is filtered based on the type of vessel that is related to export-import activities, i.e., cargo vessels and tanker vessels. This type of ship can indicate the commodity of goods transported by the vessels (Arslanalp, Marini, & Tumbarello, 2019).
- Filter 6: vessels inside a port
  From the AIS data sent by a vessel, it can be identified whether the vessel is in a port or not. This filter is used to get data on vessels that carry out export-import activities at the port. The vessel is in a port and is included as valid data if an AIS message is sent inside the bounding box of that port. There are also considerations to eliminate data if the vessel is not in any port.

3.2.2. Derive Indicator from AIS Data

Indicators related to export-import activities will be derived and aggregated from the AIS data. Some of the features in the AIS data can be seen in Table 2. Of these several features, not all features can be used to form an indicator related to export-import. As such, the related features will be selected, and they can represent the export-import indicators. The selected AIS data features are 'MMSI', 'DTG', 'draught', 'longitude', and 'latitude'. The method for calculating each indicator is explained in Table 3. This indicator will be a predictor variable to predict export-import statistics.

Table 3: Derived AIS Indicator

No	Indicator Name	Explanation
1	Length of time vessels in port / *timeInPort* (X1)	This indicator calculates the average time spent by vessels in a port in seconds. The method is done by calculating the time difference between AIS messages sent by the ships while in the port area. Then the difference in time is summed up by each vessel per port. If the time sent by the vessel exceeds three days, then the message will be eliminated from the data because it will be assumed that the message was sent by a vessel being repaired or something else so that it can't contribute to trade. Features used: MMSI, DTG, Longitude, Latitude
2	Unique numbers of vessels in port / *numVessel* (X2)	This indicator calculates the unique number of ships that enter or are in the port area based on their MMSI number (vessel identity number), so this indicator is calculated in units of the number of vessels. Features used: MMSI, DTG, Longitude, Latitude
3	Numbers of vessel visits / *numVisit* (X3)	The number of vessel visits can be calculated using the coordinates of the AIS messages sent by the vessel. If the previous message is sent outside the port area and the next message is sent in it, then the message is counted as one visit to that port. Features used: MMSI, DTG, Longitude, Latitude
4	Numbers of changes in the vessel's draught / *numDraught Diff* positive (X4) and negative (X5)	This indicator calculates the number of vessels that have changed in the vessel's draught while in port. This indicator will be divided into positive and negative vessel draught changes. Features used: MMSI, DTG, Draught, Longitude, Latitude
5	The amount of change in the ship's water draught / *sumDraughtDiff* positive (X6) dan negative (X7)	This indicator calculates the amount of change in the vessel's draught while in port in meters. The greater the indicated draught value, the deeper and heavier the vessel is. The change in vessel draught is calculated by differentiating the amount of the vessel's draught in the next AIS message with the draught in the previous message. These changes will be divided into vessels that have an increase and vessels that have a decrease in the draught. Vessels that have an increase in draught can be shown from a positive value difference and indicate an increase in load/weight on the vessel, and this indicator will be used to monitor the value/volume of exports and vice versa. Features used: MMSI, DTG, Draught, Longitude, Latitude

3.2.3. Data Transformation

Growth rate export-import data and indicators from AIS data will be calculated every month with a growth formula in equation (4):

$$\Delta y_t = \frac{y_t - y_{t-1}}{y_{t-1}} \times 100 \qquad (4)$$

where y_t is the growth of y_t against y_{t-1}, y_{t-1} is the data in the previous period, and y_t is the data for the current period. Growth rate data is used in this study because it provides a better correlation value between AIS indicators and export-import data. In addition, this transformation is also used to see and compare the movement patterns of the two data and make the data more stationary. Furthermore, the data is standardized using equation (5):

$$Z_i = \frac{X_i - \underline{X}}{S_x} \qquad (5)$$

where $\bar{X}$ and S_X are the mean and standard deviation of the X attribute. Standardization is useful for giving equal weight to all attributes, normalizing input values, and accelerating the learning process of the model, especially in the Neural Network algorithm (Han, Kamber, & Pei, 2006).

3.2.4. Predictor Variable Selection

Several predictor variable simulations have been carried out to determine the input variables of the forecasting model and to obtain optimal results. There are several simulations of these variables, which are as follows:

1. All AIS indicators that have been derived.
2. AIS indicators that have been selected based on the permutation importance of the ANN model.
3. AIS indicators that have been selected using forward stepwise selection.
4. AIS indicators that have been selected based on the level of significance in the ARIMA model.
5. AIS indicators that have been selected based on the correlation value with export-import data, which is more than 0.25.

3.3. Analysis Methods

3.3.1. Allocation of Train and Test Data

The data used are AIS indicators and Indonesia's export-import data. These data are monthly data, starting from the period December 2018 to December 2020. The data are 24 series and will be divided into 80% as training data and 20% as test data. The training data will use a series from January 2019 to July 2020, which is 19 data. Meanwhile, the test data will use series from August to December 2020, or 5 data.

3.3.2. Forecasting Models

In this study, export-import statistics will be predicted with indicators from AIS data using the ANN and ARIMA models. The ARIMA model will use indicators from AIS data as exogenous variables to forecast export-import data. This model will not consider seasonal effects because the data used are insufficient to see seasonal patterns, and there is no indication of seasonal patterns in the data period used. The parameter d in the ARIMA model will use the value 0 because the data series is already stationary at the level. Meanwhile, the parameters p and q will be determined based on the ACF and PACF plots, and a tentative model will be obtained later. From the several tentative models, the best model will be selected using the smallest AIC and BIC values generated by each model (Rahkmawati, Sumertajaya, & Aidi, 2019).

For the ANN model, the architecture used in this research is Multilayer Feed-forward Neural Network known as Multilayer Perceptron (MLP), which has more than one hidden layer. The learning algorithm used is the backpropagation algorithm, where the weight of each layer will be adjusted to get the smallest possible error between the prediction results and the actual target data. MLP can have good results for classification and prediction models and is a popular structure (Ahani, Salari, & Shadman, 2019). In addition, the MLP and backpropagation architectures have a simple process and structure (Achkar, Elias-Sleiman, Ezzidine, & Haidar, 2018), so it was chosen as the architecture in the ANN model. To get prediction results with the ANN model, several parameters need to be determined, as shown in Table 4. Because the parameter values will affect the results and performance of the model, the determination of the parameters and network topology of the ANN model will be done by tuning the parameters using random search. Random values will be generated for each ANN parameter with a range of limit values as contained in Table 4. The process will be repeated with a total of 100 iterations.

Table 4: ANN Parameter Value

Parameter Name	Parameter Value Boundary
Numbers of hidden layer	[1, 4]
Numbers of neuron	[1, 100]
Activation Function	{'identity', 'tanh', 'sigmoid', 'relu', 'leaky_relu'}
Learning rate	[0.0, 1.0]
Momentum	[0.0, 1.0]

4. Results and Discussion

4.1. AIS Data Preprocessing Results

The number of AIS data used during the period from December 2018 to December 2020 is 661.8 million records. Several filtering steps were carried out on the data to reduce noise and obtain data related to export and import, as explained in Chapter III. Table 5 shows the number of AIS messages at each filter stage, from before filtering until after the filtering stage.

In Table 5, the third filter stage, i.e., the filter based on the status of vessels at anchor, reduces a lot of data or by 75% from the previous data. The third filter uses the vessel's navigational status to determine the status of the vessel at anchor.

Table 5: Total AIS Data at Each Filter Stage

Filter Stages	Total AIS Data
Whole AIS messages	661,847,517
Filter 1: valid vessel's MMSI	655,794,385
Filter 2: moving vessels	640,138,932
Filter 3: navigation status of anchored vessels	154,253,335
Filter 4: non-zero draught	145,645,162
Filter 5: relevant vessel type	98,837,324
Filter 6: vessels inside a port	**73,576,052**

If viewed based on the navigation status, the distribution of AIS data based on the data in the previous filter is shown in Figure 2. The status of 'under way using engine' dominates the data by 53.57%, while the status of vessel navigation that related to export-import activities, i.e., 'restricted maneuverability', 'moored', and 'at anchor', only cover about 25.09% of the data. This causes a lot of data to be eliminated. There are also 'unknown' and 'not defined' statuses which cover around 12.08% of the data.

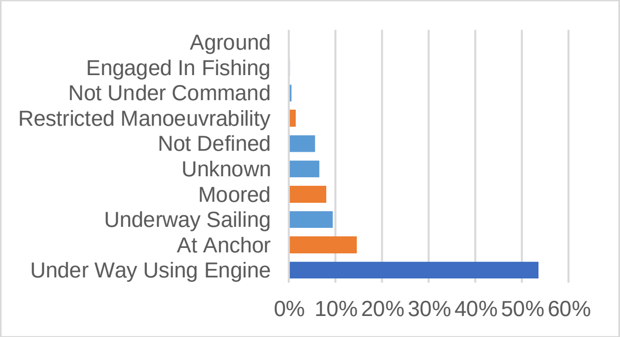


Figure 2: Percentage of AIS data after the second filter by vessel's navigation status

The fifth filter stage uses vessel types to determine the types of vessels that are related to export-import. Figure 3 shows the distribution of AIS data at the fourth filter stage when classified by vessel types. It can be seen that the types of vessels related to export-import activities, i.e., cargo and tanker vessels, cover 67.9% of the data.

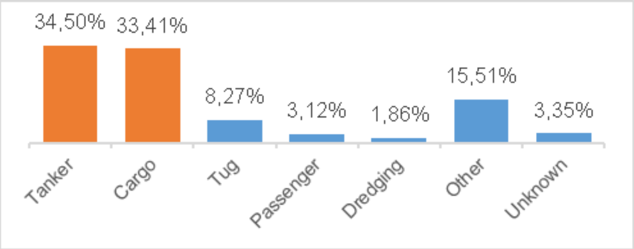


Figure 3: Percentage of AIS data after the fourth filter by vessel types

The rest are tugs, passenger ships, dredgers, and other types of vessels. Some messages are not identified or classified as 'unknown', which covers 3.35% of the data. From all the six filter stages, the amount of clean AIS data obtained is 73.6 million from 661.8 million records. Furthermore, this clean data will be used for further analysis.

4.2. Indicators Derived from AIS Data

From the clean AIS data, indicators related to export-import statistics are derived. The indicators that are derived are as shown in Table 3. If the indicator–series are compared with the export-import data series for the period of December 2018 to December 2020, it will look like Fig 4.

The AIS data indicator and export-import data have the same pattern of increase in June 2019, but the 'timeInPort' indicator does not show the pattern of increase. In addition, the indicators 'timeInPort', 'numVessel', and 'numVisit' have the same movement patterns as the export data. Meanwhile, the 'numDraughtDiff' and 'sumDraughtDiff' indicators have almost the same patterns on the import data, especially at the beginning of the period. However, there was a high increase for the indicators 'timeInPort', 'numDraughtDiff', and 'sumDraughtDiff' in December 2019, while the export-import data remains stable.

Indicators related to export-import activities have been derived and aggregated from AIS data as described in chapter IV. To find out the relationship between the AIS indicator and export-import data, it is necessary to calculate the correlation value between the two data. Figure 5 shows the Pearson correlation value between export-import data and the AIS indicator on the growth rate. It can be seen that doing preprocessing by filtering AIS data that is relevant to export-import activities, as in Table 5, can provide a better correlation between AIS indicators and export-import data.

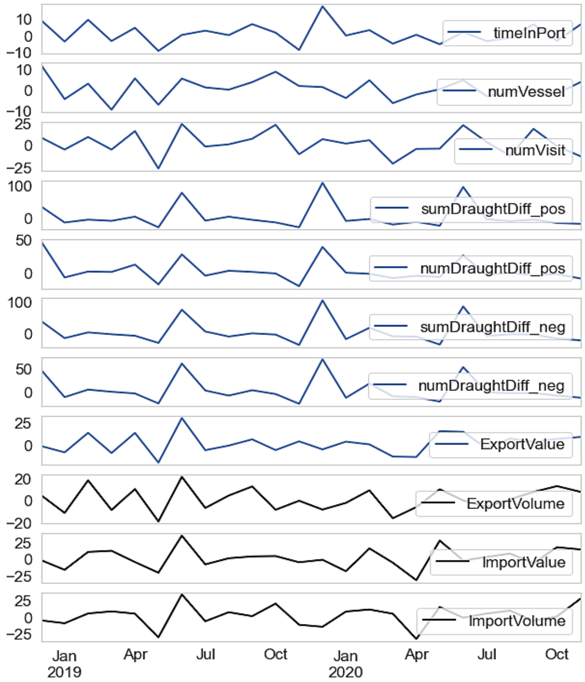
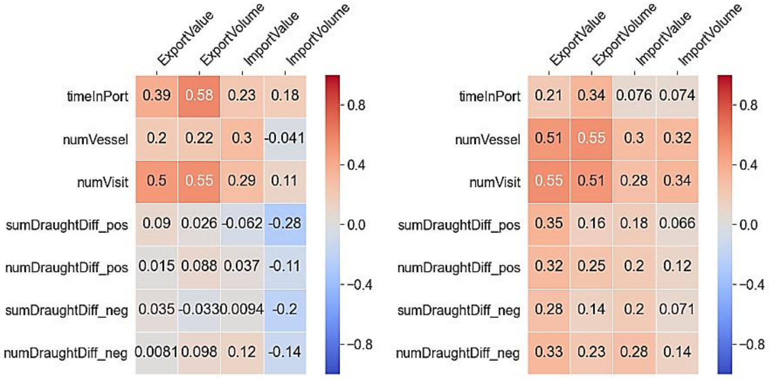


Figure 4: Plot series of AIS indicators and export-import data from December 2018 to December 2020

A positive correlation value indicates that the growth rate of the AIS indicator has a positive relationship with export-import data. The AIS indicators 'numVisit' and 'numVessel' have a fairly strong correlation (≥0.50) with the value and volume of exports, while the value and volume of imports have a moderate correlation (≥0.25). These two indicators have the highest correlation rate when compared to other AIS indicators.



(a) (b)

Figure 5: AIS indicator correlation matrix with export-import data (a) before filtering AIS data and (b) after filtering AIS data

4.3. AIS Indicator as Predictor Variable

All indicators derived from AIS data are used as predictor variables to predict export-import data. Several variable selection simulations are also carried out on these indicators to determine the appropriate predictor variables for the forecasting model. These variable selections are based on the permutation importance of the ANN model, forward stepwise selection, significant variables in the ARIMA model (α=0.05), and a correlation value of more than 0.25. The predictor variables selected for each simulation can be seen in Table 6, where each indicator AIS is denoted by a notation as shown in Table 3.

Table 6: Predictor Variable Selection Simulation

Response Variables	Without Selection	Permutation Importance	Forward stepwise selection	Sig. Var.	Corr
Export Value	X1, X2, X3, X4, X6	X2, X6	X3	X1, X2, X3	X2, X3, X4, X6
Export Volume	X1, X2, X3, X4, X6	X2, X3, X4, X6	X2	X3	X1, X2, X3, X4
Import Value	X1, X2, X3, X5, X7	X1, X2, X7	X2	X1, X3	X2, X3, X5
Import Volume	X1, X2, X3, X5, X7	X1	X3	X3	X2, X3

4.4. Export-Import Forecasting Results with ANN and ARIMA

The indicators that have been derived will be used as predictor variables to predict the value and volume of imports and exports using the ANN and ARIMA models. The RMSE value predicted by the forecasting model is shown in Table 7, which shows that the ANN model has a smaller error value than ARIMA if it is used to predict the value and volume of export and import volumes. As for the import value, the ARIMA model is more suitable to be used, because it produces a smaller error value.

Table 7: RMSE Value of Predicted Results with ANN and ARIMA Models

Response Variable	Predictor Variables	ANN		ARIMA	
		Train	Test	Train	Test
Export Value	X1^{a}, X2^{a}, X3^{a}, X4, X6	0.1068^{b}	1.2118	0.3769	1.6388
	X2^{a}, X6^{a}	0.2934	0.9310	0.4905	1.4856
	X3^{a}	**0.7223**	**0.7852**^{b}	0.4569	1.2496^{b}
	X1^{a}, X2^{a}, X3^{a}	0.2020	1.3935	0.4334	1.4063
	X2, X3, X4, X6	0.4870	0.7950	0.3494^{b}	1.8891
Export Volume	X1, X2, X3^{a}, X4, X6	1.1913	1.4106	0.7685^{b}	1.8301
	X2, X3^{a}, X4, X6	1.1937	1.3347	0.9694	1.9395
	X2^{a}	1.3341	1.2974	1.1668	1.2391^{b}
	X3^{a}	1.3618	1.2938	0.9768	1.7833
	X1, X2, X3^{a}, X4	**1.0408**^{b}	**1.2305**^{b}	0.7886	1.8389
Import Value	X1^{a}, X2, X3^{a}, X5, X7	0.0055^{b}	1.0299^{b}	0.4112^{b}	1.0004
	X1, X2, X7	0.0685	1.2387	0.7100	0.9077
	X2	0.6839	1.4943	0.7173	0.9944
	X1^{a}, X3^{a}	0.3946	1.2611	0.5232	2.2097
	X2, X3^{a}, X5	0.3344	1.5844	**0.5655**	**0.9061**^{b}

Import Volume	X1, X2, X3[a], X5, X7	0.0323[b]	1.4515	0.6563	1.7721
	X1	**0.8067**	**0.7581[b]**	0.5663[b]	1.6655
	X3[a]	0.7839	1.2394	0.7072	1.4441[b]
	X3[a]	0.7697	1.1988	0.7072	1.4441[b]
	X2, X3[a]	0.6436	1.4102	0.6822	1.6684

a. Predictor variable that is significant in ARIMA models
b. The smallest value for each response variable

In Table 7, it can also be seen that several AIS variables are not significant when used as predictor variables in the ARIMA model. In addition, some variables are not significant at all. In the ANN model, several models have indications of overfitting on the model. This can be seen from the error values. The error value is small for the train data but in contrast, it is large for the test data especially in the ANN model with import data. The RMSE value on the training data is 0.0055, while the test data is 1.0299. This makes the ARIMA model chosen as a forecasting model on the import values because it has a smaller error value. However, for other models, the ANN model has a smaller error value so that it is superior to ARIMA.

The best model is selected based on the smallest error value generated in the test data. The predictor variables used in each model and the best model parameters are shown in Table 8. It can be seen that the best variable selection is using forward stepwise selection, correlation value, and permutation importance. The ANN model was chosen to forecast the value of exports, export volumes, and import volumes. As for the import value using the ARIMA model.

Table 8: Best Model Parameters and Predictor Variables Used

Response Variables	Variables Selection	Predictor Variables	Best Model Parameters
Export Value	Forward stepwise selection	numVisit (X3)	'hidden_layer': [75, 50, 100], 'activation':'leaky_relu', 'learning_rate': 0.7348, 'momentum': 0.5662
Export Volume	Correlation Value	timeInPort (X1), numVessel (X2), numVisit (X3), numDraughtDiff positif(X4)	'hidden_layer': [60, 9, 69], 'activation':'leaky_relu', 'learning_rate': 0.1513, 'momentum': 0.4370
Import Value	Correlation value	numVessel (X2), numVisit (X3), numDraughtDiff negatif (X5)	ARIMA (2,0,0)
Import Volume	Permutation importance	timeInPort (X1)	'hidden_layer': [62, 88, 91], 'activation': leaky_relu', 'learning_rate': 0.9829, 'momentum': 0.5484

The best model that has been obtained is used to forecast each response variable. Figure 6 shows a plot of prediction results with actual values showing the same pattern and small error, that is, the forecasting of import values with indicators from AIS. As for export data, the forecasting results are not the same. Because the import value forecasting model uses ARIMA, it can be said that the ARIMA model can produce a better forecasting pattern than ANN, especially on import values with the AIS indicator as a predictor variable.

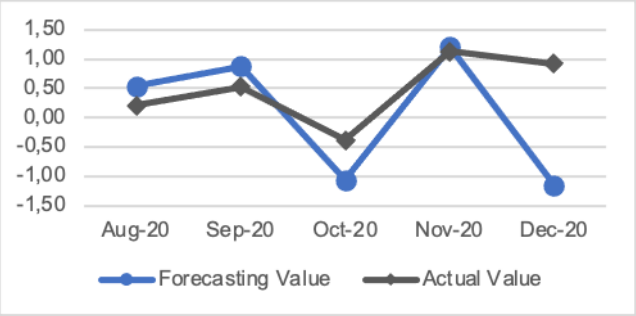


Figure 6: A plot of prediction results and actual value with small error and the same pattern in import value

This research can be classified as a new research on the use of AIS data in Indonesia, so, there are a few things that limit this research. One of them is that the monthly AIS data and export-import data are only available for two years. Therefore, they are not able to capture patterns of data movement for the previous period. In addition, access to AIS data is still limited, so not everyone can access the data.

In this study, the volume of export and import is estimated using the draught feature of the AIS data. The volume of export and import can be estimated better if there is information on the weight of the vessel. For example, information about cargo load indicators using the draught and deadweight tonnage vessel features (Arslanalp, Marini, & Tumbarello, 2019). The AIS data used in this study is ship movement report data, and there is no information about the weight of the ship/vessel. This information can be obtained from the AIS port calls data (Arslanalp, Marini, & Tumbarello, 2019) or by combining shipping information/data in Indonesia.

The variable selection method used in this study uses a simple method. Other variable selection methods, such as MCP, SCAD, or LASSO, can be used to produce more optimal performance or prediction results (Ahani, Salari, & Shadman, 2019). ANN forecasting models can also use other architectures such as Recurrent Neural Network (RNN) or Long Short Term Memory (LSTM), which can study long-term dependencies and can understand data structures over time. As such, they have a high predictive capacity on time series data (Achkar, Elias-Sleiman, Ezzidine, & Haidar, 2018; Gharibshah, Zhu, Hainline, & Conway, 2020)

5. Conclusions and Future Work

Based on the research results, it can be concluded that AIS data can be used to derive indicators related to export-import statistics. By carrying out the proper preprocessing on the data, indicators derived from AIS data can have a better relationship pattern with export-import statistics. The indicators derived are the length of time vessels in port, the number of unique vessels, the number of vessel visits, and the number and amount of changes in the vessel's draught. The AIS indicator can also be used as a predictor variable to forecast export-import data. By selecting predictor variables, it can provide better forecasting results, such as selection based on forward stepwise selection, correlation values, or permutation importance. The ANN model is the best model for forecasting export value, export volume, and import volume. As for the value of imports using the ARIMA forecasting model, the ANN model is superior because it provides a small error value for three of the four export-import statistics, while the ARIMA model provides a better forecasting pattern than the ANN model on the import value statistics.

The recommendations for further research are as follows:

1. Using a longer range of AIS data and export-import data to identify export-import patterns in previous periods and create AIS indicators with more Using a longer range of AIS data and export-import data to identify the patterns in the previous period. Also create AIS indicators with more granular, such as on a weekly or daily basis.
2. Using a further developed step of preprocessing or cleaning AIS data. For example, by combining it with ship weight information to help estimate the volume of export-import.
3. Using variable selection methods or forecasting methods and other ANN architectures, such as LSTM or RNN.

References

Achkar, R., Elias-Sleiman, F., Ezzidine, H., & Haidar, N. (2018). Comparison of BPA-MLP and LSTM-RNN for Stocks Prediction. *International Symposium on Computational and Business Intelligence (ISCBI).*

Adland, R., Jia, H., & Strandenes, S. P. (2017). Are AIS-based trade volume estimates reliable? The case of crude oil exports. *Maritime Policy & Management*.

Ahani, I. K., Salari, M., & Shadman, A. (2019). Statistical models for multi-step-ahead forecasting of fine particular matter in urban areas. *Atmospheric Pollution Research, 10*(3), 689-700.

Ahmed, F., Cui, Y., Fu, Y., & Chen, W. (2021). A Graph Neural Network Approach for Product Relationship Prediction. *ASME IDETC*.

Arslanalp, S., Marini, M., & Tumbarello, P. (2019, December). Big Data on Vessel Traffic: Nowcasting Trade Flows in Real Time. *IMF Working Paper.*

Badan Pusat Statistik. (2020, November). *Perdagangan Luar Negeri*. Retrieved from Web Badan Pusat Statistik: https://www.bps.go.id

Badan Pusat Statistik. (2021, February 8). *Ekspor dan Impor.* Retrieved from Situs Badan Pusat Statistik : http://www.bps.go.id/exim/

Cerdeiro, D. A., Komaromi, A., Liu, Y., & Saeed, M. (2020, May). World Seaborne Trade in Real Time: A Proof of Concept for Building AIS-based Nowcasts from Scratch. *IMF Working Paper*.

Dewi, D. M., & Wulansari, I. Y. (2021). The Impact of Information and Communication Technology (ICT) on Regional Economy in Indonesia 2012-2018. *Asian Journal of Bussiness Environment, 11*, 21-32.

Gharibshah, Z., Zhu, X., Hainline, A., & Conway, M. (2020). Deep Learning for User Interest and Response Prediction in Online Display Advertising. *Data Sci. Eng., 5*, 12-26.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and technique.* San Fransisco: Morgan Kaufman Publisher.

International Maritime Organization. (2021, March 26). *Regulations for carriage of AIS.* Retrieved from https://www.imo.org

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2007). *Introduction to Time Series Analysis and Forecasting.* Canada: John Wiley & Sons, Inc.

National Geospatial-Intelligence Agency. (2021, February 8). *World Port Index: Query Results of Indonesian Port*. Retrieved from Maritime Safety Information: https://msi.nga.mil

Neves, J., & Cortez, P. (1998). Combining Genetic Algorithms, Neural Networks, and Data Filtering for Time Series Forecasting. *IMACS International Conference on Circuits, Systems, and Computers (IMACS-CSC'98)* (pp. 933-939). Piraeus, Greece: IMACS CSC.

Nooraeni, R., Sari, P. N., & Yudho, N. P. (2019). Using Google trend data as an initial signal of Indonesia unemployment rate. *ISI World Statistics Congress.* Kuala Lumpur.

Noyvirt, A. (2019). Faster Indicators of UK Economic Activity: Shipping. *Data Science Campus.*

Rahkmawati, Y., Sumertajaya, I. M., & Aidi, M. N. (2019). Evaluation of Accuracy in Identification of ARIMA Models Based on Model Selection Criteria for Inflation Forecasting with the TSClust Approach. *International Journal of Scientific and Research Publications, 9*(9), 439-443.

Raymond, E. S. (2021, January). *AIVDM/AIVDO protocol decoding.* Retrieved Juni 2021, from https://gpsd.gitlab.io/gpsd/AIVDM.html

Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. Y., & Khan, S. U. (2016). Big Data Reduction Methods: A Survey. *Data Sci. Eng., 1*, 265-284.

Torres, J., Gutiérrez-Avilés, D., Lora, A., & Martínez-Álvarez, F. (2019). Random Hyper-parameter Search-Based Deep Neural Network for Paper Consumption Forecasting. *IWANN*.

UN Global Working Group. (2019). *United Nations Global Platform: Data for the World.* UN Global Working Group.

USCG. (2021, January). *Definition - Vessel Restricted in Her Ability to Maneuver*. (USCG Navigation Center) Retrieved June 2021, from https://www.navcen.uscg.gov

Zissis, D., Xidias, E. K., & Lekkas, D. (2016). Real-time vessel behavior prediction. *Evolving Systems*(7), 29-40.