10-31-2021

Port throughput forecasting using ARIMA and OLS regression: case study : Gwangyang port in Korea

Shin Park

WORLD MARITIME UNIVERSITY
Malmö, Sweden

Port throughput forecasting using ARIMA and OLS regression
(case study: Gwangyang port in Korea)

By

SHIN PARK
The Republic of Korea

A dissertation submitted to the World Maritime University in partial fulfilment of the requirements for the reward of the degree of

MASTER OF SCIENCE
in
MARITIME AFFARS

(PORT MANAGEMENT)

2021

Declaration

I certify that all the material in this dissertation that is not my own work has been identified, and that no material is included for which a degree has previously been conferred on me.

The contents of this dissertation reflect my own personal views, and are not necessarily endorsed by the University.

(Signature):
......Shin Park......

(Date):
......2021.09.21.

Supervised by: Associate Prof. Gang Chen Gang Chen

Supervisor's Affiliation: World Maritime University

Acknowledgements

First of all, I would like to thank the CEO of Yeosu Gwangyang Port Authority, Minsik Cha, for giving me the opportunity to study at the World Maritime University, despite the difficult times due to the Covid-19 corona pandemic.

Also, I would like to thank my supervisor, Professor Gang Chen, for patiently and meticulously guiding my dissertation despite my lack of knowledge. Also, I would like to thank Professor Satya for teaching me a new area of study, which is Matlab program, through Maritime analytics course. Also, I would like to express my sincere gratitude to Professor Dongwook Song, who generously gave me psychological support.

In addition, I want to thank my Port Management and Shipping Management friends who have studied together with me for about a year. It was not easy for each other to study together with different backgrounds from different countries, but I think this experience will be a great help for me to think and judge things with a new perspective in the future. Also, I would like to thank Seongjae Kim, Secretary of the Ministry of Oceans and Fisheries of Korea, who wrote a dissertation in the same space despite different majors.

Lastly, I am also grateful to my family, Yangmi Ryu, Minseo Park, and Taemin Park, who came to far away and stayed with me. Even in Korea, we didn't have a lot of time together with the excuse that I was busy at work, and I feel apologetic that it seems to be the case here. I am very grateful to my wife who silently supported me from behind.

At the end of about a year of studying in Sweden, I think this experience will be of great help and will be remembered for my life in the future.

Abstract

Title of Dissertation: Port throughput forecasting using ARIMA and OLS Regression (case study: Gwangyang port in Korea)

Degree: **Master of Science**

Predicting future cargo volume is the most basic step in establishing mid- to long-term strategies for ports. In this regard, forecasting the volume of cargo is important for Gwangyang Port, which needs to establish master strategy to improve the competitiveness of container terminals, which are suffering from stagnant cargo volume and intensifying competition between container terminals.

Therefore, in this study, the ARIMA model, which is a representative univariate method, and the OLS regression model, which is a multivariate method, were used to confirm which method is suitable for predicting the throughput of Gwangyang port. And, in this process, important variables were identified for the change in Gwangyang port throughput, and the future cargo volume was predicted.

As a result, it was found that the OLS regression model is more suitable for forecasting the Gwangyang port throughput, and in this process, it was confirmed that government consumption, China's imports, and the Korean exchange rate were important variables for the change in cargo volume. In addition, the cargo volume of Gwangyang port was predicted to be stable without significant change.

KEYWORDS: OLS regression model, ARIMA, forecasting, port throughput, port planning, Gwangyang port

Table of Contents

Declaration 1

Acknowledgements 2

Abstract 3

Table of Contents.. 4

List of Tables.. 5

List of Figures 6

List of Abbreviations 7

Chapter 1. Introduction 9

1.1. Background 9
1.2. Gwangyang Port 11
1.3. Research Objective and Questions 13
1.4. Research Contribution 14
1.5. Structure of the Dissertation...14

Chapter 2. Litrature Review... 16

2.1. Port Planning...16
2.2. Forecasting Methods...17
2.3. Variables influencing container throughput 22

Chapter 3. Data and Methodology... 23

3.1. Data and Source..23
3.2. OLS regression modelling process 30
3.3. ARIMA modelling process 44
3.4. Forecasting 47

Chapter 4. Results and findings 51

Chapter 5. Conclusion.. 52

5.1. Conclusion 52
5.2. Limitations 53
5.3. Scope for future research..54

References... 55

List of Tables

Table 01. Trade dependence of 12 countries in the G20 9

Table 02. Descriptive Statistics 31

Table 03. Unit Root Test Result 33

Table 04. Correlation test result of independent variables 34

Table 05. t-test and f-test results 35

Table 06. The correction for the ARCH effect and Serial correlation 39

Table 07. Ramsey's Reset Test result 41

Table 08. Final regression model about GY Import and export 43

Table 09. Unit Root Test Result for ARIMA modelling 45

Table 10. Forecast result using OLS regression model 48

List of Figures

Figure 01. Cargo volume of Korean major port in 2020 12

Figure 02. GY Container port throughput trend (2001~2020) 13

Figure 03. Neural Network Model 18

Figure 04. Method of OLS fitting a line by minimising SSE 21

Figure 05. Container throughput of Korean major port 25

Figure 06. Korea's Export and Import 25

Figure 07. Korea's Export and Import 26

Figure 08. Korea's GDP 27

Figure 09. Korea's Industry Production 28

Figure 10. Korea Government consumption... 28

Figure 11. Korea major import and export countries and their share 29

Figure 12. Flow chart of OLS regression model 30

Figure 13. CUSUM Test result 42

Figure 14. Flow chart of ARIMA model 44

Figure 15. GY port quarterly container throughput 45

Figure 16. ACF and PACF test results 46

Figure 17. Forecast result using ARMA model 49

Figure 18. GY port throughput forecasting result 50

List of Abbreviations

ACF	Autocorrelation function
ADF	Augmented Dickey-Fuller
AIS	Automatic identification system
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroskedasticity
ARIMA	Autoregressive integrated moving average
BLUE	Best Linear Unbiased Estimator
CLRM	Classical Linear Regression Model
CPI	Consumer Price Index
DW	Durbin-Watson
ECT	Error correction term
FTZ	Free Trade Zone
GDP	Gross Domestic Product
GM	Grey Model
GY	Gwnagyang
GWCT	Gwangyang West Container Terminal
HSGT	Hanjin Shipping Gwangyang Terminal
KMI	Korea Maritime Institute
KIT	Korea International Terminal
KPSS	Kwaitowski, Phillips, Schmidt and Shin
MA	Moving Average
MAE	Mean Absolute error
MAPE	Mean Absolute Percentage Error
MOF	Ministry of Oceans and Fisheries
MSE	Mean Squared Error

NN	Neural Network
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Square
O/D	Origin and Destination
PACF	Partial autocorrelation function
PDFC	Port Demand Forecasting Center
Port-MIS	Port Management Information System
PP	Phillip-Perron
P-value	Probability value
RMSE	Root Mean Squared Error
RSS	Residual sum of squares
SARIMA	Seasonal autoregressive integrated moving average
SSE	Sum of the squared errors
TEU	Twenty-foot Equivalent Units
VTS	Vessel traffic service
YGPA	Yeosu Gwangyang Port Authority

1. Introduction

1.1 Background

Korea's dependence upon foreign trade (total imports and exports divided by GDP) is 63.51%, the second highest among the 12 countries with statistics released in G20 countries (Statistics Korea, 2019). This is because Korea, which lacks resources and has a small domestic market, has developed an industry that imports and processes raw materials to produce finished products and export them. In general, trade between countries can be done through land, sea, and air routes. However, since South Korea is virtually cut off from the continent by North Korea and surrounded by sea on three sides, 99.7% of imports and exports are transported by ships (Korea Shipowner's Association, 2020). Therefore, trade through the sea is one of the most crucial parts of the Korean economy, and it will also be important to properly equip port facilities for trade activities.

rank	country	2016	2017	2018	2019
1	Germany	68.93%	72.23%	72.03%	70.82%
2	Korea	60.11%	64.82%	66.08%	63.51%
3	Canada	52.77%	52.54%	53.63%	52.48%
4	Italy	46.48%	49.30%	50.27%	50.28%
5	France	43.43%	44.67%	44.95%	44.96%
6	England	36.65%	39.36%	39.16%	39.53%
7	Australia	32.15%	34.57%	34.35%	35.35%
8	Indo	27.29%	28.24%	30.89%	28.15%
9	Japan	25.44%	28.15%	30.01%	28.08%
10	Argentina	20.32%	19.54%	24.48%	25.40%
11	Brazil	18.30%	18.20%	22.69%	22.27%
12	USA	19.79%	20.30%	20.79%	19.34%

Table 01. Trade dependence of 12 countries in the G20 (2016 ~ 2019)

Source: Korea National Statistical Portal

Port is an essential infrastructure for ships to load or unload cargo, and it takes a lot of time and money from planning to construction and operation to make or expand a port facility. Therefore, in order to properly expand the port facilities to meet the future demand in the rapidly changing maritime logistics environment, a plan should be established based on the demand forecast results. Otherwise, excessive or insufficient facilities can cause various problems such as wasting money and increasing waiting time.

A simple classification of the entities that build port facilities is the government and the private sector. Of course, there are cases where the government receives private investment to build ports. Private companies, especially petrochemical companies, do not need a yard because they handle cargo with pipes and loading arms, so they mainly build pile-type piers that are relatively inexpensive. Although private companies sometimes build gravity-type piers, in particular, container terminals require huge capital, so in Korea, they are mainly built by government or port authorities. According to the Port law of Korea, the government establishes the 'Port Master Plan', which includes the mid- to long-term port development plan, every 10 years, and changes it if necessary every five years from the date on which the Port Master Plan is established. The plan is actually established every five years. In order to establish or modify this plan, it is necessary to predict the demand for cargo throughput. Accordingly, the Ministry of Oceans and Fisheries (MOF) designated the port demand forecasting center (PDFC) of the Korea Maritime Institute (KMI) as an exclusive agency for that (2010). This PDFC predicts future cargo demand every 5 years and recalculates the cargo handling capacity of port to determine the excess or shortage of port facilities. The government establishes a port development plan based on this data.

There are various methods of forecasting port demand. Among them, to briefly explain the method used by the PDFC, first, the total cargo is classified into 32 items and the future cargo volume for each item is predicted. In this process, an appropriate prediction model is used for each item and applied, and a regression model or a time series model is basically used. In particular, containerized cargo is calculated by classifying containerized cargo among the predicted cargo volume by item, and then applying the containerization rate to it. Next, according to the results of container cargo O/D (origin/destination) analysis by region and route, the future container volume for each port is predicted. Although this method has been briefly described here, it is calculated with complex procedures and arithmetic based on extensive qualitative and quantitative data. In addition, since this forecast is basically made every five years for the port master plan and revised plan established by the government, it is difficult to use it to establish an expansion plan at a specific time in each port. In addition, in the rapidly changing maritime logistics environment, it is important to forecast demand using data at that time when demand forecasting is necessary. Therefore, it is necessary for the four port authorities that operate major ports to individually predict future port demand and use it as a basis for establishing master plans for each port. In particular, in Gwangyang port, a new port master plan is needed to solve the situation in which container cargo has been stagnant for a long time, and the competitiveness has been greatly reduced due to excessive price competition and reduced productivity. Therefore, this study intends to contribute to the future development of the Gwangyang container terminal by predicting the cargo demand, which is the basis of the port master plan.

1.2 Gwangyang Port

Korea has 31 international trade ports, of which 4 major ports, namely, Busan port, Gwangyang port, Ulsan port, and Incheon port, are under the jurisdiction of the Port Authority. Each port authority operates and constructs their facilities according to the port master plan established by the government, which the Ministry of Oceans and Fisheries announced the 4th (2021~2030) National Port Master Plan in December 2020. According to this plan, the basic direction of port development for major ports is as follows: Busan port is a logistics hub for trans-pacific intermodal transportation, Incheon port is a comprehensive logistics gateway port for the capital area, Ulsan port is an energy logistics hub, and Gwangyang port is Asia's best high-tech complex logistics port.

As of the end of 2019, Korea's total port cargo volume was about 1,638 million tons. Busan port ranks first with 467 million tons, Gwangyang port second with 310 million tons, Ulsan port third with 202 million tons, and Incheon Port fourth with 157 million tons. However, if we look at only container cargo, Busan port handles 21,910 thousand TEU, Incheon port 3,087 thousand TEU, and Gwangyang port 2,377 thousand TEU. Incheon Port handles more containers than Gwangyang Port. This is because the proportion of container cargo at Gwangyang port is only 16%, while Busan and Incheon account for 94% and 35%, respectively.

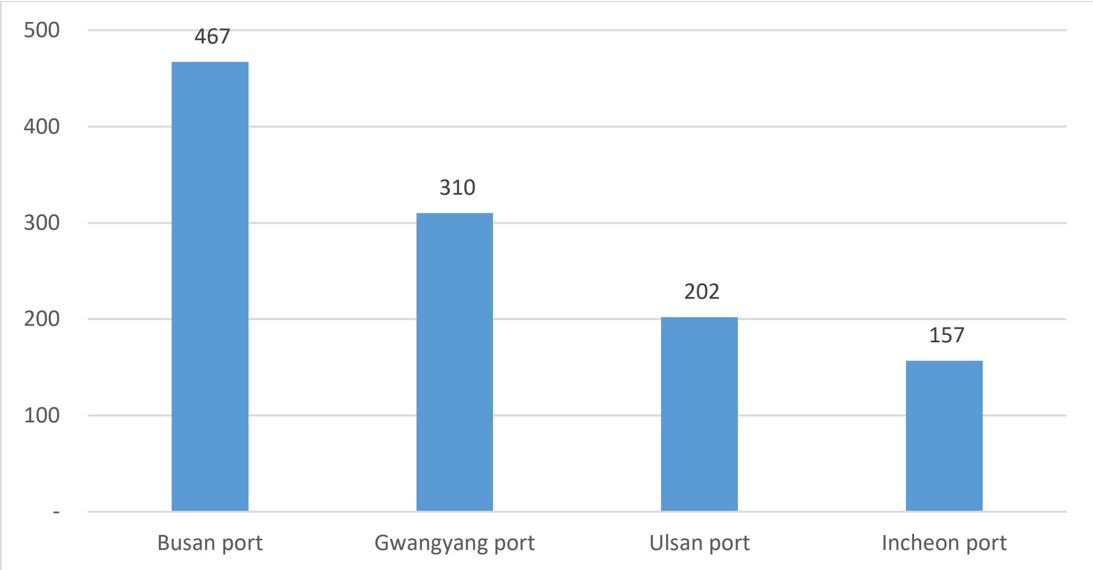


Figure 01. Cargo volume of Korean major port in 2020 (unit: million ton)
Source: Ministry of Oceans and Fisheries

Gwangyang port is a comprehensive port that handles chemical, steel, container, car and general cargo. In terms of cargo composition, chemicals 40%, steel 34%, containers 16%, car and other 10%. Chemicals and steel are creating stable cargo throughput as there are large factories of conglomerates such as GS Caltex, LG Chem, and POSCO near ports. In addition, automobile terminals are operated by Hyundai Glovis, a subsidiary of Hyundai Motors, with fixed cargo, and there is also stable cargo volume. However, unlike Busan and Incheon port, which are adjacent to megacities such as Busan and Seoul, there are insufficient industrial complexes and populations behind Gwangyang port to create or consume containerized cargo. Accordingly, various problems are occurring in the container area.

Of the total 109 berths in Gwangyang port, the container terminals have 12 berths. Yeosu Gwangyang Port Authority (YGPA), as the manager of Gwangyang port, has the ownership of container terminals and leases them to the three terminal operating companies, which are Hanjin Shipping Gwangyang terminal (HSGT), Korea International terminal (KIT), and Gwangyang west container terminal (GWCT). The problem is that the capacity of the container terminals is 4.6 million TEU, but the amount of cargo handled is only 2.4 million TEU per year. This is only the volume that one operator can handle in Busan port. In addition, as three operators compete for a nearly fixed 2 million container throughput every year, container (un)loading freight fee are very low compared to other ports, resulting in a deficit every year. Accordingly, YGPA is in the process of establishing a new master plan to solve the problem of the container terminal area. Therefore, it will be necessary to forecast future container demand in order to establish this plan.

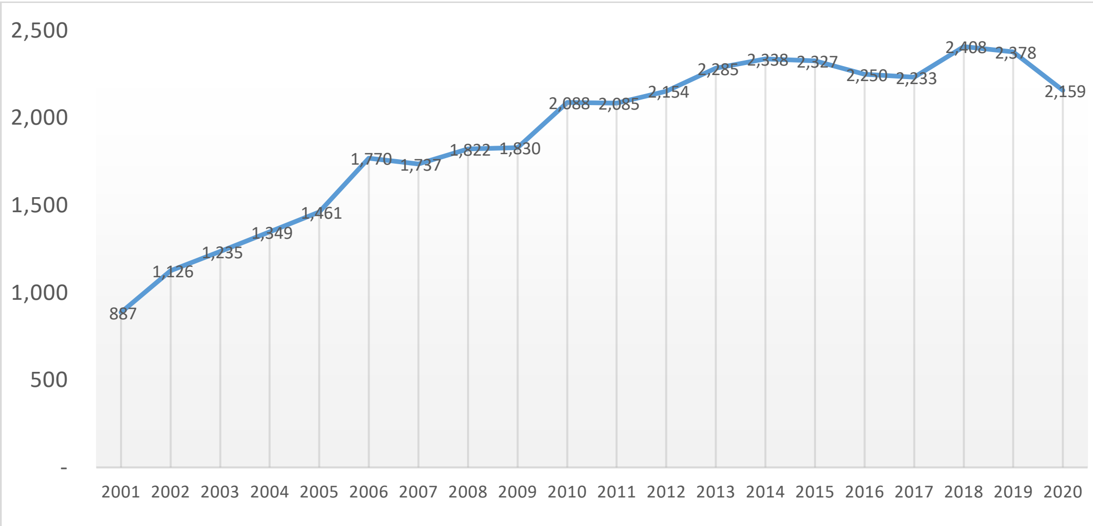


Figure 02. GY Container port throughput trend (2001~2020) (unit: thousand TEU)

1.3 Research Objectives and Questions

In order to overcome the problems of the container terminal in Gwangyang port, Yeosu Gwangyang Port Authority (YGPA) plans to implement various policies such as consolidation of container terminal operating companies, relocation of container terminals, and construction of automated container terminal. This plan is based on the current trend of stagnant container volume. However, in the rapidly changing maritime logistics environment, it is necessary to predict the future container volume through various methods, continuously update demand forecast, and revise policies accordingly. In addition, it will be possible to identify factors affecting the container throughput of Gwangyang port and utilize it for activities to increase container volume.

The research questions of this study are as follows.

- What factors significantly affect Gwangyang port container throughput?
- Which of the ARMA and OLS regression model is suitable for forecasting Gwangyang port throughput?
- How will the demand for container volume in Gwangyang port change in the future?

1.4 Research Contribution

As mentioned above, the PDFC of KMI is in charge of forecasting cargo volume in Korea every five years. Of course, some researchers individually research and publish papers on the forecast of future container demand, but most of them are focused on Busan port, the Korea No. 1 port, and there are few studies on Gwangyang port. Therefore, this study can be seen as almost the first attempt for demand forecasting modelling for Gwangyang port only. Through this study, it will be possible to utilize basic data for the establishment of the Gwangyang Port container terminal master plan. In addition, the constructed demand forecasting model will be able to be used whenever necessary according to the ever-changing maritime logistics situation through continuous revision and supplementation. Moreover, the established modelling will be able to help identify the factors affecting the container cargo throughput of Gwangyang port. YGPA may be able to continuously monitor these factors for increasing the container cargo volume. YGPA is currently trying to increase the stagnant cargo volume by expanding the hinterland near the port (free trade zone). Whether these efforts can contribute to an increase in port throughput in the mid- to long-term can also be confirmed through future demand forecasting modelling.

1.5 Structure of the Dissertation

This dissertation consists of 6 chapters. Chapter 1 briefly describes the background, Gwangyang port, objective, and contribution. Chapter 2 reviews existing studies on port planning and various forecasting techniques. Chapter 3 explains the variables used in forecasting and describes the process of forecasting modelling by applying these variables to Ordinary least square (OLS) regression analysis and Autoregressive integrated moving average (ARIMA) model. Chapter 4 used the forecasting model created in Chapter 3 to forecast the future cargo volume of GY port and evaluated its accuracy. Chapter 5 explains the findings and their meanings through the method applied in this study. Finally, Chapter 6 describes the conclusions and limitations of this study, and the scope for future research.

2. LITERATURE REVIEW

2.1 Port Planning

In the rapidly changing maritime environment, port owners, which can be government, port authorities and private companies, do port planning to expand port facilities in order to respond to future increases in cargo demand. It plays an important role in continuous development of ports. Taneja at el (2010) stated that appropriate investment in port facilities can secure market share and strengthen the competitive position of ports. Memos (2004) mentioned that port planning usually refers to a plan to create a new port or expand an existing port to increase capacity or upgrade port operations.

Port planning can be divided into several stages according to viewpoints and criteria. Prakash Gaur (2005) classified the port planning process into Institutional Framework, demand and supply forecasting, capacity planning and project evaluation. On the other hand, Notteboom et al (2021) divided port planning into mission or strategy establishment, identification of difference between ability and ideal, resource creation to narrow the difference, strategy establishment and implementation, and evaluation of the effectiveness of the selected strategy. As such, the names of each process for port planning may be different, but most port planning processes include demand forecasting, stakeholder involvement, and evaluation of port plans. (Notteboom et al., 2021).

Each step for port planning is important, but demand forecasting is most important for policy makers to establish a plan to secure an appropriate size port facility at the right time (Langen et al., 2012). This is because a port expansion plan established based on an incorrect demand forecast can cause various problems due to the difference between supply and demand of port facilities. If the supply is greater than the demand, it can lead to inefficiency due to overcapacity of the facility. On the other hand, when the supply is greater than the demand, the lack of facilities can cause congestion, waste unnecessary costs and time due to increased waiting time, and reduce the competitiveness of the port (Jarrett, 2015). In addition, since port expansion takes a lot of time for planning and construction, it is most important to establish a port plan based on port demand (Langen et al., 2012).

2.2 Forecasting methods

Cambridge dictionary defines forecasting as “the activity of judging what is likely to happen in the future, based on the information you have now.” In the shipping and port industry, policymakers use forecasting techniques to predict how future container throughput will change before making decisions on port expansion (Gosasang et al, 2011). This is because it is important to accurately forecast the throughput in order to establish an appropriate plan (Shu, Huang & Nguyen, 2013).

There are several methods of forecasting, but they can be broadly divided into qualitative and quantitative methods (Gaur, 2005; Notteboom et al., 2021).

1) Qualitative method

Qualitative method is also called subjective, judgmental and technological methods. This method is used when data is insufficient or ambiguous, and instead of using measurable data, it relies on expert judgment or opinions. Qualitative method includes Delphi method, panel consensus forecast, historical analogy and market research forecasts. Among these, Delphi method and panel consensus forecast mainly use expert opinions, and historical analogy and market research forecasts use expert opinions and economic knowledge. Qualitative method is a viable tool for examining the impact of economic and transportation trends on the future trajectory of cargo volume (Parola et al., 2021). In general, this method is not reliable or highly accurate. Therefore, it is mainly used to make rough predictions in situations of high volatility.

2) Quantitative method

Quantitative method can be used when numerical data on the past are available and the trend of historical data is expected to continue in the future (Hyndman & Athanasopoulos, 2018). There are several types of quantitative forecasting methods. Among them, the most representative ones are Neural Network, Grey forecast, ARIMA model and regression analysis.

① Neural Network

Neural Network (NN) is forecasting models based on simple mathematical structure of the human brain (Lam et al., 2004). NN can be seen as a network configuration of 'neurons' composed of layers. The inputs form the bottom layer, hidden neurons (optional) form intermediate layers, and outputs forms the top layer (Hyndman and Athanasopoulos, 2018). Each layer is interconnected in the form of neurons. When one neuron receives a weight input, the input is sent to another neuron and converted into an output (Jansen, 2014). This model is more accurate in predicting container throughput in the short term than in the long term. In addition, since a lot of data is required for this prediction, it is difficult to use in practice (Lam et al., 2004). Gosasang et al. (2011) used NN and linear regression methods in the research, and concluded that NN method is more suitable for Bangkok port throughput forecasting.

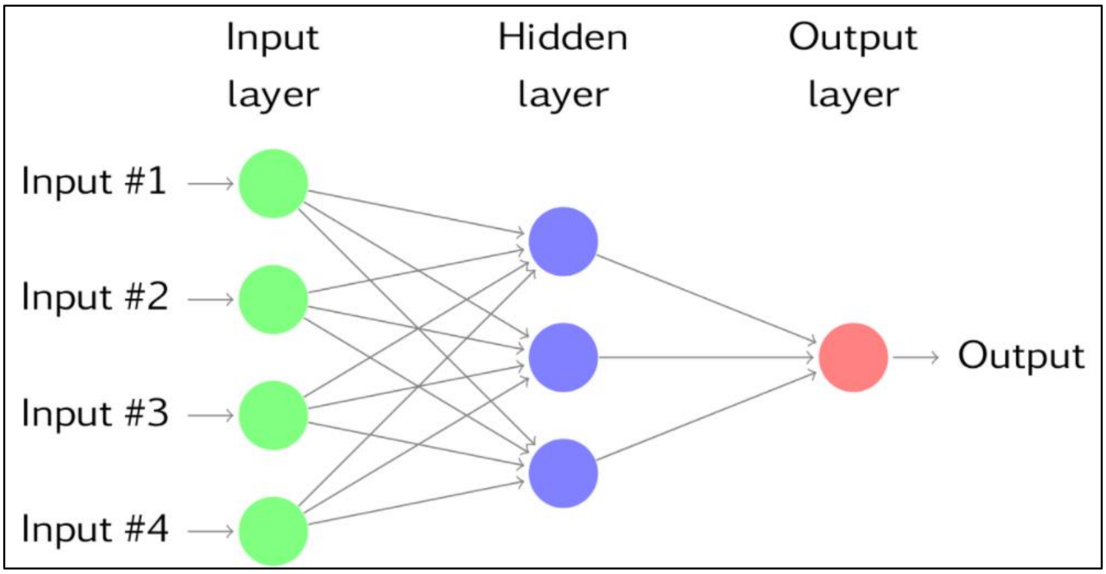


Figure 03. Neural Network Model

Source: Adapted forms "Forecasting: Principles and Practice (Hyndman et al., 2018)

② Grey forecast

Grey forecasting was introduced by Deng (1989). This is a particularly suitable method for forecasting data where incomplete information or uncertain behaviour is a common problem. A feature and advantage of this method is that it requires less data to make predictions. Grey models can be represented by the order of the differential equation and the number of variables included. For example, GM (1,1) means a model with the first order of the differential equation and one variable (Peng & Chu, 2009). Grey models are widely used to predict port throughput, and the main examples are as follows: Qiuhong (2009) predicted the future traffic volume of Qinhuangdao port using the grey model, Du (2013), Jiujiang port, and Hui-yuan (2009), Shenzhen port.

③ ARIMA models

Autoregressive Integrated Moving Average (ARIMA) models are the commonly used prediction models in univariate time series analysis. ARIMA models use historical data of variables to make short-term predictions efficiently (Witt and Witt, 1992). ARIMA model can be classified into five groups, which are autoregressive model (AR), moving average (MA), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA). Autoregressive (AR) predicts future values based on a combination of previous values and AR(p) can be express as follows.

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \dots\ + \theta_p y_{t-p}$$

where y_t represent predicted value at time t and θ is estimated coefficient.

On the other hand, Moving Average (MA) is a method of making predictions based on previous prediction errors and MA (q) can be express as follows.

$$y_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \dots\ + \phi_q u_{t-q}$$

where u_{t-q} the approximate error at time t-q and ϕ is estimated coefficient.

ARMA is the sum of AR and MR. That is, ARMA is a prediction method based on previous values and prediction errors. ARMA model assumes that the data is stationer. However, if non-stationarity element is added to ARMA process, the model becomes ARIMA. The ARIMA model is denoted by Arima (p, d, q), where p means AR process with order p, q, MA with order q, and d, the number differencing required to make the series stationary (Anggraeni et al., 2015; Peng & Chu, 2009). Meanwhile, SARIMA is part of ARIMA, but it is widely used models that have a constant periodic pattern in seasonal time series data (Kotcharat, 2016). For example, in quarterly data, SARIMA model can be used when container volume always tends to be high or low in a particular quarter. ARIMA model has been widely used in the shipping and port field. Dragan and Kramberger (2014) compared exponential smoothing model, classical decomposition model and ARIMA model to find the suitable method for predicting the container traffic volume at the Port of Koper in the North Adriatic Sea. In this study, the authors concluded that among the three methods, ARIMA models gave the most accurate results. Kyung-Chang Min et al (2014) predicted the container traffic volume in Korea using SARIMA model.

On the other hand, unlike the ARIMA model that predicts the future only with the past values of the dependent variable, which is univariate model, ARIMAX is a multivariate model that uses the independent variable. The ARIMAX model is an extension of the ARIMA model with explanatory variables (Stock & Watson, 1999). This model is also referred to as dynamic regression or vector ARIMA model. Several studies of macroeconomic forecasting have found that including external variables improves forecasting performance.

④ Regression analysis

Regression analysis is a statistical technique that uses one or more variables (x_s, independent variables) to explain the movement of one variable (y, dependent variable). This method not only predicts the values of dependent variables, but also find out the strength of the relationship between the dependent variable and the independent variable. The most common of this method is ordinary least squares (OLS) regression analysis. OLS regression analysis is a method of estimating the relationship between the dependent variable and the independent variable by minimizing the sum of the squared errors (SSE), which is the difference between the predicted values and the actual values of the dependent variable formed as one straight line (Hutcheson, 2015; Brooks, 2019).

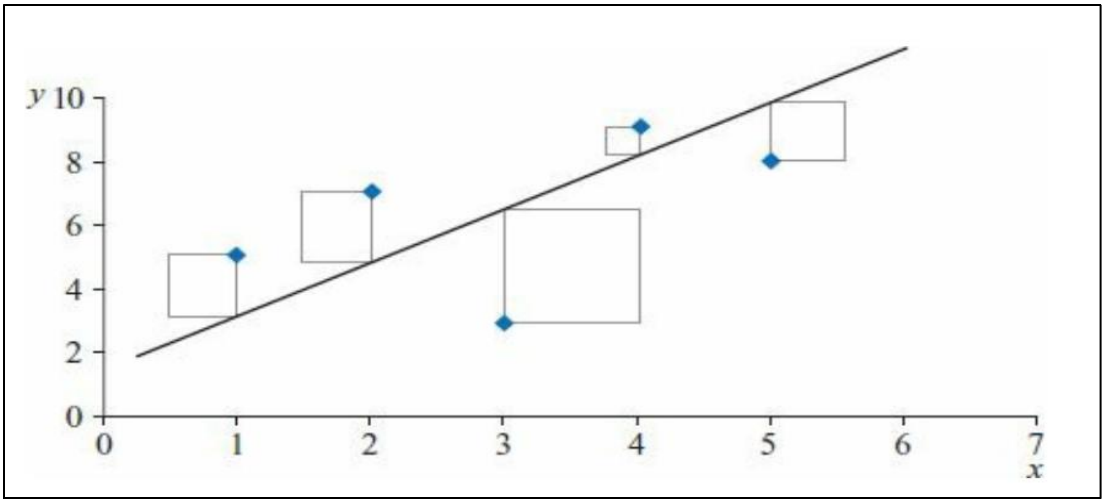


Figure 04. Method of OLS fitting a line by minimising the sum of squared errors
Source: Introduction Econometrics for Finance (Brooks, 2019)

It can be broadly divided into a bivariate and a multivariate model. A bivariate model is a model with only one independent variable, while a multivariate model is a model with two or more independent variables (Lewis-Beck 1980; Vittinghoff et al. 2005). The bivariate model and multivariate model can be expressed as follows, respectively.

Bivariate model: $y = \alpha + \beta x_t + u_t$

where α is the intercept, β is the slope of the line, u represents the error, and the subscript t (= 1, 2, 3, …) denotes the observation number

Multivariate model: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u_t$

where x_k represent the values of k different explanatory variables and β_k are the coefficients.

So far, among various forecasting models, univariate methods such as Neural Network (NN), Grey Model (GM), Autoregressive Integrated Moving Average (ARIMA), and multivariate models, OLS regression analysis, have been examined. Meanwhile, in a comparative study of predictive models conducted by Chan et el. (2019), among the four traditional time series methods, which are Moving Average (MA), ARIMA, GM, and Artificial Neural Network (ANN), ARIMA was found to be the best method for short-term forecasting.

Therefore, in this study, among the many predictive models, the GY port container throughput forecasting modelling was implemented using the ARIMA and OLS regression methods, which are relatively widely used and highly accurate in univariate and multivariate models.

2.3 Variables influencing container throughput

There are many factors, which are explanatory variables, that influence container throughput. In particular, there is a close relationship between port cargo volume and macroeconomic variables since economic development is an important engine of maritime trade (Langen et al., 2012). Among these, many previous studies suggested GDP is the most important variable (Jugovic et al., 2011; Pina and Fei, 2013). In addition, looking at other previous studies, many other variables must also be considered when analyzing port traffic. Gökkusr et al (2017) said that important macroeconomic variables affecting container traffic were GDP, Consumer Price Index (CPI), world GDP, the volume of national import-export trade, and the national population. According to “Empirical analysis Influence factors to container throughput in Korea and China Ports” (Liu and Park, 2011), terminal storage capability, berth length, direct call liner, transhipment, hinterland's GDP, hinterland's import-export volume, port tariff, FTZ area and investment of government was established as an independent variable influencing container volume. As a result, in the case of Korean ports, the transhipment volume and port tariff were important variables, whereas in the case of Chinese ports, hinterland economic level and government’s investment were important factors. Meanwhile, Kotcharat (2016) used government expenditure index, private consumption index, private investment index, industry production index, sale of important products, employment, trade between big trade partners of the country, bunker price Singapore, intra-Asia container freight rate index and exchange rate as independent variables.

As mentioned above, there are various factors that can be used as independent variables. However, not all these factors can be used. Since the reliability of data, the period of presented data, and the possibility of securing the data are different, it is necessary to secure and use the data in consideration of these points.

3. Data and Methodology

3.1 Data and Source

Since this study uses regression models for forecasting, time series data was used. In addition, good modelling is important for good forecasting, but it will be more important to secure various reliable data. Because good data can eventually make good forecasting models. Therefore, only data issued by the government or public institutions was collected.

The time period of data is 20 years from 2001 to 2020, and using quarterly data, the total number of data is 80. Although the dependent variable, GY container port throughput, can be secured on a monthly basis, since major macroeconomic indicators such as GDP are quarterly data, the entire data is adjusted quarterly. This is because, in general, all data used in a model should have the same observation frequency (Chris, 2019).

Meanwhile, the port competitiveness indicators such as (un)loading productivity and dwelling time can affect container throughput. However, like other port-related indicators, these data are difficult to obtain, and reliability may be low even if they are secured, so they were excluded from this study.

3.1.1. Dependent Variable

Gwangyang Port is in the stage of an important long-term policy decision and implementation to break through the currently stagnant container volume, and for this purpose, it is important to predict the future volume. Therefore, the dependent variable in this study is the container throughput of Gwangyang Port. This is an open data that anyone can obtain from the port management information system (Port-MIS) managed by the Ministry of Oceans and Fisheries. Port-MIS is the system that can manage port operation status in real time and charge port fees by connecting ship’s automatic identification system (AIS) and vessel traffic service (VTS) center information.

3.1.2. Independent Variables

There are two major methods for selecting explanatory variables that need to be considered in model-building: 'Forward selection' and 'Backward elimination'. Forward selection is a method that repeats including one by one important variables until no more significant variables appear. On the other hand, forward selection is a method that includes all possible independent variables in the first place, and sequentially removes unimportant variables from the model (Hutcheson, 2015). In this study, all available data that could affect port throughput were included in the starting model. Afterwards, insignificant variables were removed at each step. This method minimizes the possibility that important independent variables are not included in the forecasting model. As a result, 21 independent variables were used at the beginning of the modelling, and they were classified into Korea internal index (11), external index (4), and shipping index (6) according to their characteristics and types.

(1) Internal index

These are and variables related to trade volume and macroeconomic indicators in Korea related to GY container volume. The variables used in this study are Busan and Incheon port throughput, export and import of goods and services, Gross Domestic Product (GDP), Industrial product, Government consumption expenditure, working age population and won-dollar exchange rate.

① **port throughput** (Busan, Incheon, Gwangyang)

Busan, Gwangyang, Ulsan, and Incheon port are the four major ports in Korea and are operated by four port authorities. Among them, Busan, Incheon, and Gwangyang port are handling 27,254 thousand TEU, which is about 94% of Korean total container volume. Busan port is the 6th largest container transshipment hub in the world, and also handling about 75% of Korean container volume. As a gateway port to the capital of Korea, Incheon port had less container cargo than Gwangyang port until 2014, but overtook Gwangyang in 2015 and handled 3,092 thousand TEU in

2020, which is about 1.5 times that of Gwangyang port. Busan and Incheon port are competitive with Gwangyang port for container cargo.

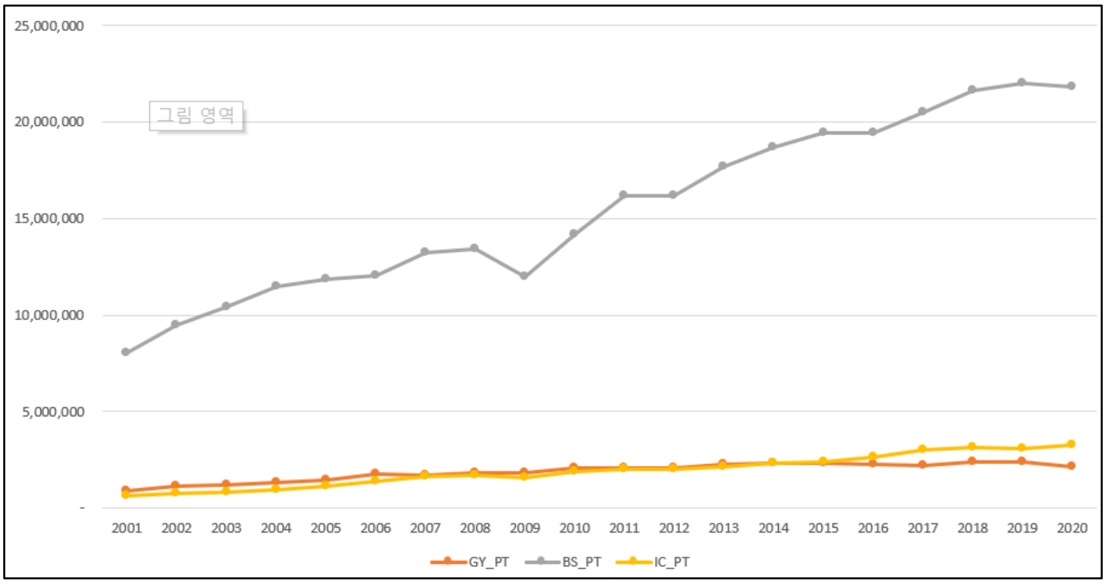


Figure 05. Container throughput of Korean major port

② Import and export of goods and services

The import and export statistics are data on the exchange of cargo between the Korean economy and other countries. The total price of the export and import cargo by year are as follows.

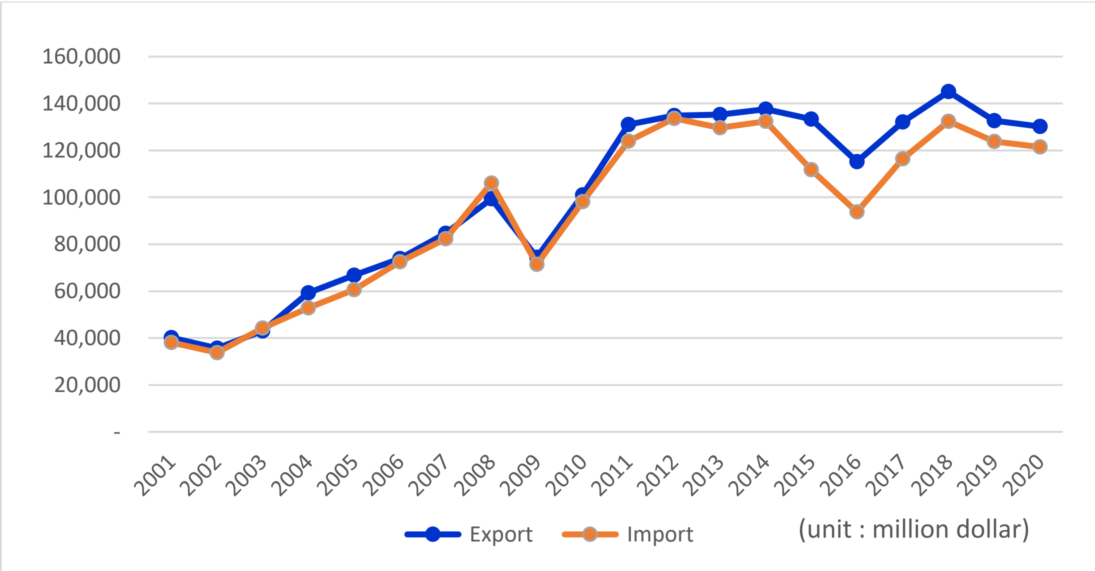


Figure 06. Korea's Export and Import (million dollar)

Korea's total amount of exports and imports has increased almost 3.2 times in 2020 compared to 2001, showing almost the same trend. The total weight of the export and import cargo by year are as follows.

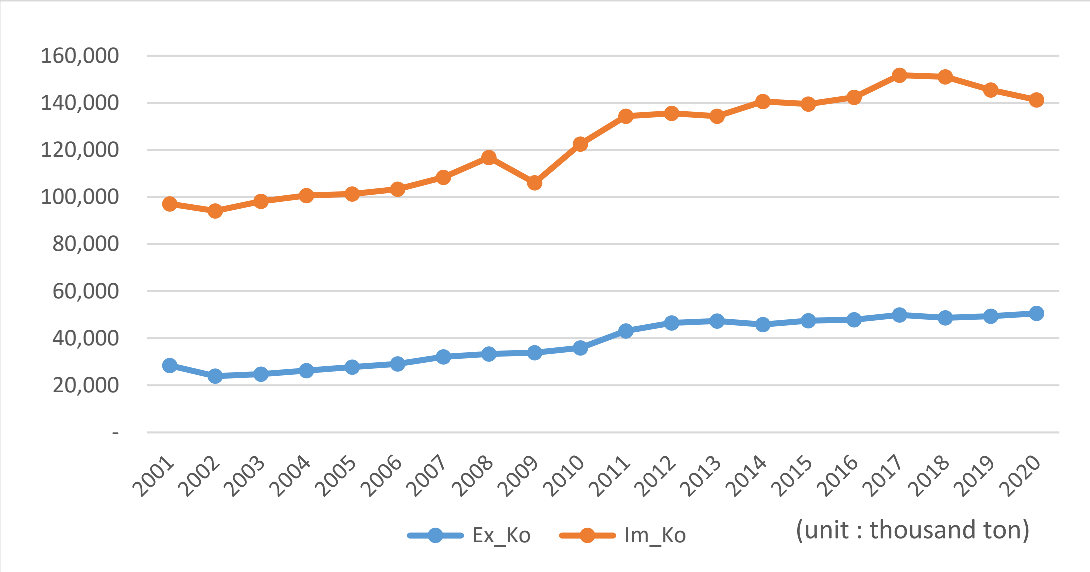


Figure 07. Korea's Export and Import (thousand ton)

Korea's total weight of exports and imports are showing a modest increase, and imports are about three times exports. This seems to be because Korea, which lacks resources, mainly exports heavy-weight raw materials such as coal and iron ore and finished products with relatively low weight. As Korea's total price and weight of import and export increase, the handling cargo volume of ports, which is a major trading hub, will also increase.

③ Gross Domestic Product (GDP)

GDP is calculated by multiplying the quantity of final products produced in Korea by the price at the time, and is an indicator used to determine the size of the economy (Statistics Korea, 2021). GDP data is announced every quarter by Korea Bank. GDP can be expressed as the sum of private consumption spending, investments, government spending, and the differences between exports and imports. According to Michael et al. (2020), container trade is an important determinant of GDP growth.

In addition, Liu & Park (2001) concluded that GDP is an important independent variable for China and Korea port throughput in their research.

Korea's GDP is continuously increasing, and GDP in 2020 is about 2.7 times compared to 2001.

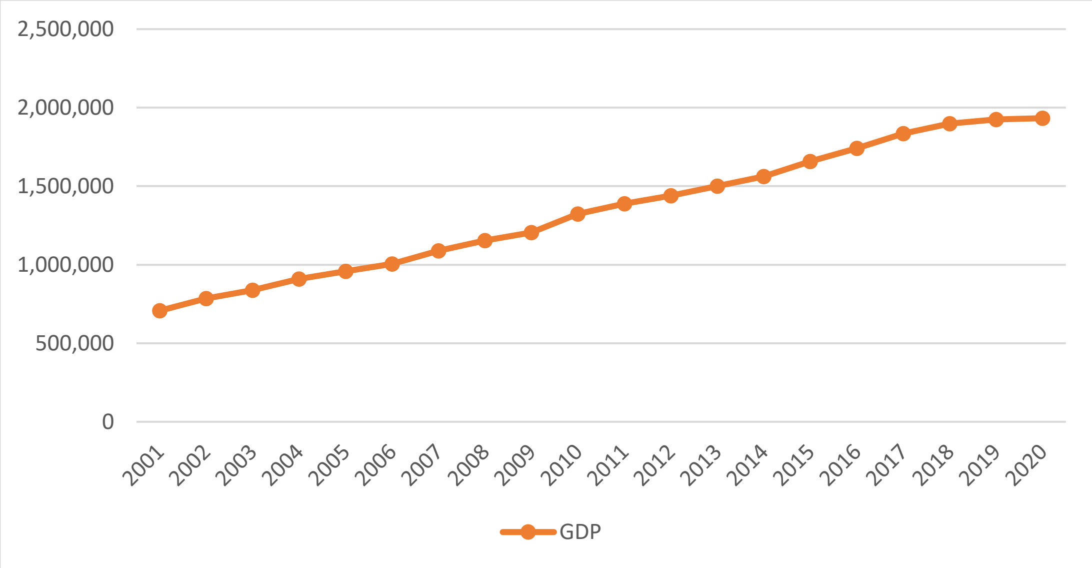


Figure 08. Korea's GDP

④ Other index

In addition to the Korea internal index explained so far, industry production, government consumption, working age population, and exchange rate were used as independent variables affecting port throughput. Industrial production is the output of industrial facilities and includes sectors such as manufacturing, mining etc (OECD). The government final consumption expenditure is the government expenditure on the production of non-market final goods and services and the market goods and services provided in kind through social transfer. The working age population is all people in the aged from 15 to 64 years, which referred to as the potential labour supply. Korea's industry production and government consumption continue to increase. However, Korea's working age population has continued to decline, peaking at 36,871 thousand in the first quarter of 2017 due to low fertility and aging population.

On the other hand, Korea, which is highly dependent on trade, has a large impact on the economy, specially imports and exports, by the exchange rate. In general, a weak domestic currency makes domestic products more price-competitive, stimulating exports, but making imported goods more expensive. Conversely, the strength of the domestic currency lower exports, but the imported goods' price becomes relatively cheap. Korea's exchange rate fluctuated in a complex manner depending on the global economic situation, Korean government's policies, and interest rates.

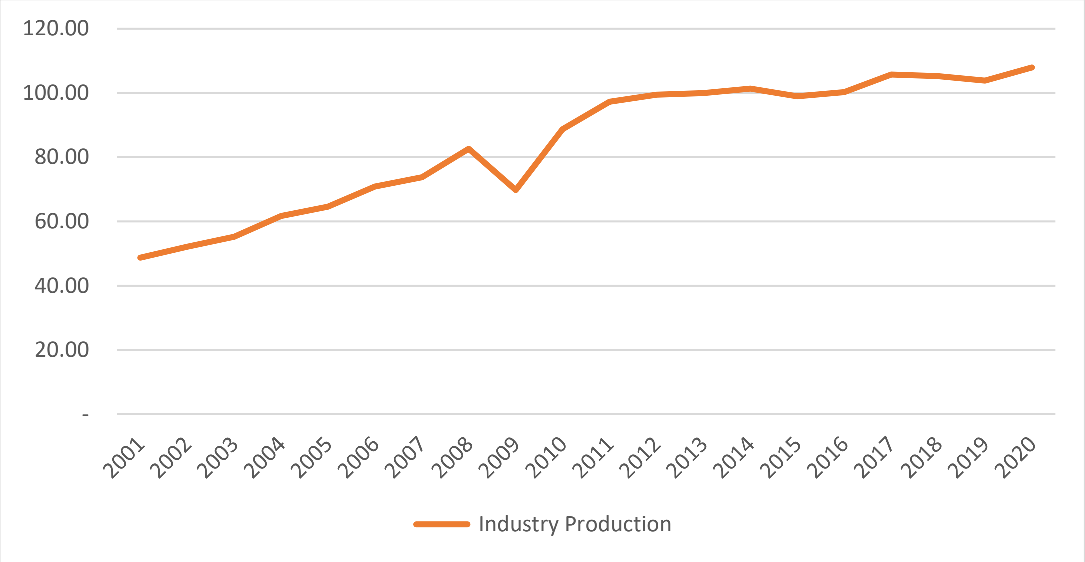


Figure 09. Korea's Industry Production

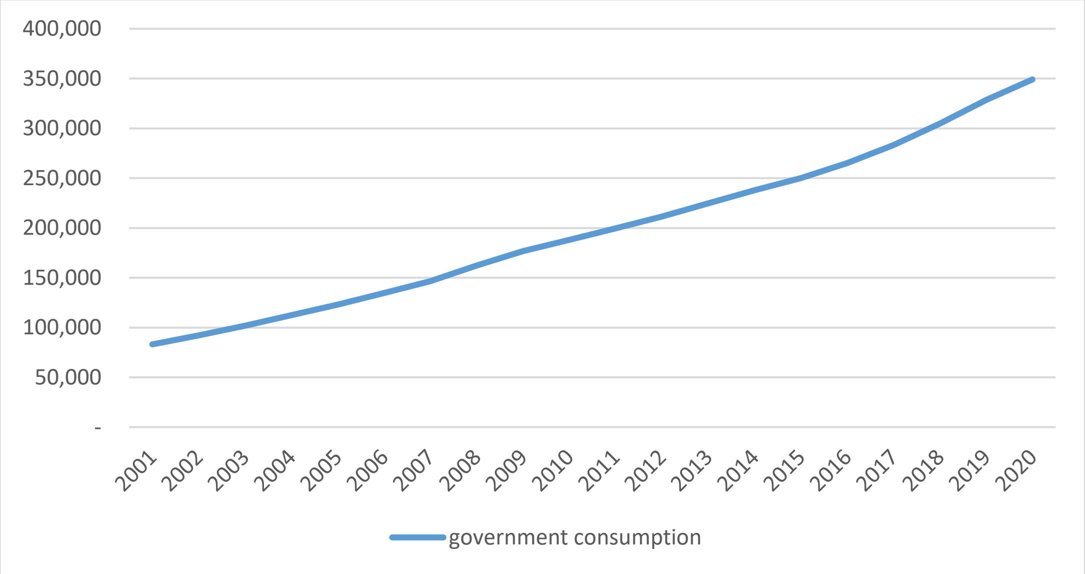


Figure 10. Korea Government consumption

(2) External index

GY port's container volume will also fluctuate depending on the external independent variables such as imports and exports of major trading countries and world trade volume. Therefore, these indicators were also included as independent variables.

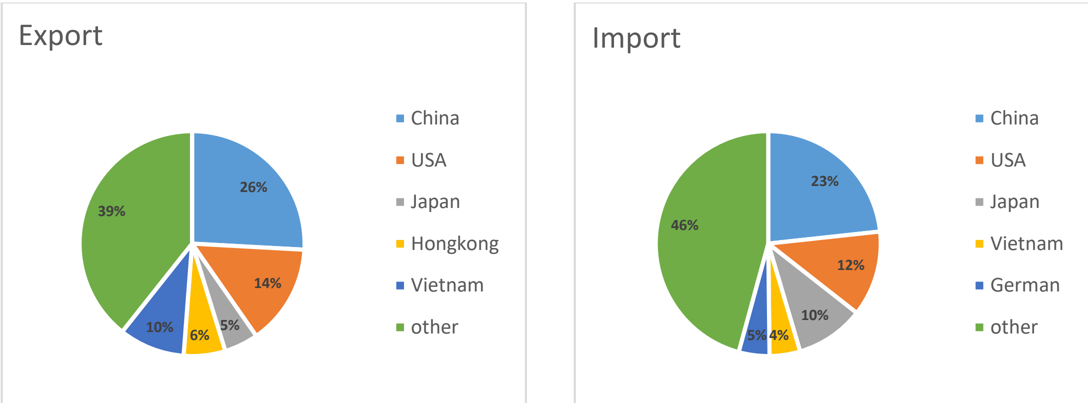


Figure 11. Korea major import and export countries and their share

Korea exported 512,498 million dollars as of 2020, and the main exporting countries are China, the United States, Japan, Hong Kong and Vietnam. These countries account for 60.67% of total exports. Imports amounted to 467,632 million dollars, and the main importing countries are China, the United States, Japan, Vietnam and Taiwan. These countries account for 54.24% of total imports. If the amount of imports and exports with various countries around the world increases, and also world trade volume, container handling volume will increase. Therefore, the amount of import and export by year of major Korea's export and import countries, China, the United States, Japan, and Vietnam, was included as an independent variable.

(3) shipping Index

Shipping related indexes such as containership new building price, second hand price, time charter rate, fleet development, order book, and bunker price are main indicators that change according to the supply and demand of container ships and cargo. Most of these indicators are reflected in shipping freight and can directly or indirectly affect container throughput.

3.2. OLS regression modelling process

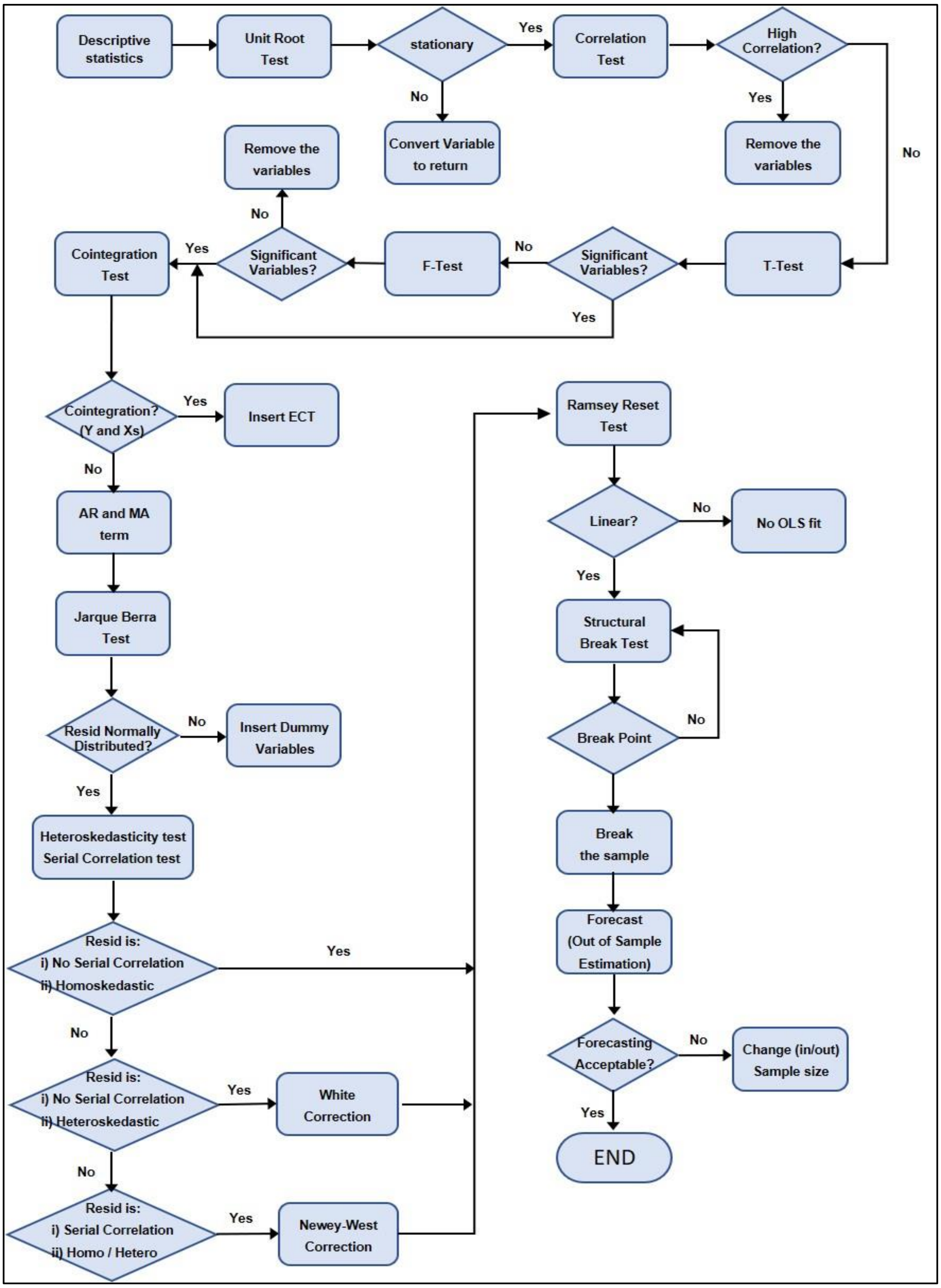


Figure 12. Flow chart of OLS regression model

Source: modified prof. Satya's lecture notes (2021)

3.2.1 Preliminary test

(1) Descriptive statistics

When analyzing a given variable, the characteristics of each data can be described through several important basic measures. ‘Max/Min, Mean, Median, Standard deviation, and Skewness’, which are considered important among these summary statistics, were measured in this paper.

Variables	Maximum	Minimum	Mean	Median	Std_Dev	Skewness
GY_PT	644111	180387	474040	523899	116307	- 0.74
BS_PT	5716983	1923948	3883677	3713146	1099629	0.02
IC_PT	865901	139893	482844	478566	207720	0.13
Ex_Ko_M	154547	35606	106059	118277	37324	- 0.56
Im_Ko_M	139314	33788	96500	106164	33242	- 0.57
Ex_Ko_T	53051	23926	39780	43634	9548	- 0.33
Im_Ko_T	157123	87822	122578	128780	20830	- 0.13
GDP_Ko	492100	172146	335695	340055	99527	0.03
IP_Ko	110	49	86	96	19	- 0.52
Gov_con_Ko	87774500	19,975,100	49689020	48565900	19786186	0.28
Workage_Ko	36,870,976	32,906,740	35,324,346	35,673,224	1,362,495	- 0.34
ER_Ko	1,409	922	1,128	1,131	103	0.04
W_Trade	125	65	101	104	18	- 0.45
China_Im	34	9	23	26	7	- 0.58
China_Ex	76	10	45	46	18	- 0.32
LIBOR	5.47	0.25	1.90	1.41	1.58	1.00
Con_NBP	127	68	88	80	17	1.01
Con_SHP	161	25	68	59	38	0.92
Con_TCR	170	32	70	58	33	1.27
Con_FD	23,461	4,925	14,000	14,249	5,912	- 0.02
Con_OB	6,828,000	1,058,277	3,593,937	3,586,506	1,421,735	0.22
BunkerP	733	118	369	331	170	0.40

Table 02. Descriptive Statistics

(2) Unit Root Test

The (non)stationarity of data can greatly affect its behaviour and characteristics. For example, when an external shock such as the economic recession is given, in the case of a nonstationary series, the effect does not decrease over time and can infinite. Also, when non-stationary data is used, it can show a high R^2 even if the two variables are completely unrelated. Therefore, it is necessary to check the stationarity of the data, and if it is non-stationary, we need to convert it to stationary before using it. Augmented Dickey-Fuller (ADF), Phillip-Perron (PP), Kwaitowski, Phillips, Schmidt and Shin (KPSS) are commonly used for stationary test. After applying the test, if the series are non-stationary, it can be made stationary by differencing. If it has to be differenced d times before it becomes stationary, it can be said to be integrated of order d and written I (d).

Variables	difference	ADF		PP		KPSS	
		P-Value	Stat	P-Value	Stat	P-Value	Stat
GY_PT	I (0)	0.97	1.59	0.97	1.59	0.01	1.32
	I (1)	0.00	- 9.30	0.00	- 9.30	0.10	0.04
BS_PT	I (0)	1.00	2.47	1.00	2.47	0.01	0.42
	I (1)	0.00	- 10.32	0.00	- 10.32	0.10	0.03
IC_PT	I (0)	0.98	1.77	0.98	1.77	0.01	1.00
	I (1)	0.00	- 15.47	0.00	- 15.47	0.10	0.02
Ex_Ko_M	I (0)	0.98	1.67	0.98	1.67	0.01	1.62
	I (1)	0.00	- 10.11	0.00	- 10.11	0.10	0.03
Im_Ko_M	I (0)	0.98	1.71	0.98	1.71	0.01	1.57
	I (1)	0.00	- 6.65	0.00	- 6.65	0.10	0.05
Ex_Ko_T	I (0)	0.92	1.06	0.92	1.06	0.01	1.29
	I (1)	0.00	- 13.09	0.00	- 13.09	0.10	0.07
Im_Ko_T	I (0)	0.90	0.89	0.90	0.89	0.01	0.70
	I (1)	0.00	- 10.32	0.00	- 10.32	0.10	0.04
GDP_Ko	I (0)	1.00	9.54	1.00	9.54	0.01	1.53
	I (1)	0.00	- 4.61	0.00	- 4.61	0.10	0.03
IP_Ko	I (0)	1.00	3.04	1.00	3.04	0.01	1.68
	I (1)	0.00	- 6.45	0.00	- 6.45	0.10	0.03
Gov_con_Ko	I (0)	1.00	14.37	1.00	14.37	0.01	1.81
	I (1)	0.00	- 3.89	0.00	- 3.89	0.10	0.07
Workage_Ko	I (0)	1.00	9.94	1.00	9.94	0.01	1.56
	I (1)	0.02	- 2.34	0.02	- 2.34	0.10	1.19
ER_Ko	I (0)	0.51	- 0.37	0.51	- 0.37	0.01	0.51
	I (1)	0.00	- 6.90	0.00	- 6.90	0.10	0.06
W_Trade	I (0)	1.00	2.46	1.00	2.46	0.01	0.92
	I (1)	0.00	- 6.34	0.00	- 6.34	0.10	0.04
China_Im	I (0)	0.97	1.63	0.97	1.63	0.01	1.59
	I (1)	0.00	- 8.85	0.00	- 8.85	0.10	0.03

China_Ex	I (0)	0.98	1.64	0.98	1.64	0.01	1.32
	I (1)	0.00	- 11.21	0.00	- 11.21	0.10	0.03
LIBOR	I (0)	0.18	- 1.30	0.18	- 1.30	0.01	0.76
	I (1)	0.00	- 4.29	0.00	- 4.29	0.01	0.24
Con_NBP	I (0)	0.58	- 0.20	0.58	- 0.20	0.01	0.80
	I (1)	0.00	- 5.63	0.00	- 5.63	0.10	0.11
Con_SHP	I (0)	0.42	- 0.63	0.42	- 0.63	0.01	0.78
	I (1)	0.00	- 5.07	0.00	- 5.07	0.07	0.14
Con_TCR	I (0)	0.58	- 0.18	0.58	- 0.18	0.01	0.56
	I (1)	0.00	- 4.43	0.00	- 4.43	0.10	0.11
Con_FD	I (0)	1.00	15.67	1.00	15.67	0.01	1.95
	I (1)	0.13	- 1.47	0.13	- 1.47	0.01	0.31
	I (2)	0.00	- 9.66	0.00	- 9.66	0.10	0.02
Con_OB	I (0)	0.78	0.35	0.78	0.35	0.01	1.53
	I (1)	0.00	- 4.29	0.00	- 4.29	0.09	0.13
BunkerP	I (0)	0.79	0.38	0.79	0.38	0.01	1.41
	I (1)	0.00	- 7.60	0.00	- 7.60	0.10	0.04

Table 03. Unit Root Test Result

ADF, PP, and KPSS test result for all given variables, only ‘Container Fleet Development (Con_FD)’ is I (2), and the rest is I (1). Therefore, if they are differenced two times for I (2) and one time for I (1), they will become stationary.

(3) Correlation Test

When independent variables are highly correlated with each other, a problem known as multicollinearity arises. A multicollinearity problem has occurred, but if it is not resolved, the following problems occur. R^2 is high, but individual coefficients will have high standard errors. Also, regression will be very sensitive to small changes, so adding or removing independent variables will have a big impact on the value or significance of coefficients of other variables. Finally, this makes the confidence intervals for the parameters very wide, which may lead to inappropriate conclusions in the significance test. Therefore, it is important to test the correlation between independent variables and to solve it.

To easily test the correlation of variables, Pearson's correlation coefficient is often used. The results of correlation analysis for all independent variables using this method are as follows.

Variables	BS_PT	Ex_Ko_T	Im_Ko_T	GDP_Ko	IP_Ko	......	China_Im	LIBOR	......	BunkerP
BS_PT	100%									
Ex_Ko_T	33%	100%								
Im_Ko_T	5%	30%	100%							
GDP_Ko	21%	35%	36%	100%						
IP_Ko	40%	41%	34%	56%	100%					
......	......	......	......	......	......	......				
China_Ex	15%	24%	- 4%	32%	30%	......	100%			
LIBOR	17%	22%	2%	13%	3%	......	- 8%	100%		
......	......	......	......	......	......	......	......	......	......	
BunkerP	25%	26%	23%	34%	62%	......	35%	12%	......	100%

Table 04. Correlation test result of independent variables

As a result of the test, the correlation between independent variables are below 80%. If these values are above 80%, the multicollinearity problem will happen. There are three methods to solve this problem, which are to ignore it, drop one of the collinear variables or transform the highly correlated variables into a ratio. However, the easiest solution to this problem is to drop one variable considering its importance among the correlated variables. However, no additional measures are taken because multicollinearity problems do not occur between the variables used in this study.

3.2.2 Coefficient Diagnostics test & ARMA terms

In this process, T-test and F-test are used to determine whether independent variables are significant to the dependent variable, and Cointegration test is applied to determine whether there are correlations between variables. In addition, it will be tested whether dependent variables and errors depend on each past value, and include AR and MA terms to solve this problem.

(1) **T-test and F-test**

T-test (t-statistics) is a statistical method for testing whether the independent variable is important in explaining the dependent variable. If the regression equation is 'y=a + β x + ut', a hypothesis can be used to check the relationship between variables. The null hypothesis (H_0) is β =0, which is however the independent variable changes, y value is not affected. This means the independent variable is not important for y value, so this variable can be eliminated from the regression model. The alternative hypothesis (H_1) is β =1, which is the independent variable is important and can affect y value.

However, since T-test is used to test a hypothesis that includes only one coefficient, there is a limitation using only T-test in this study, which use multiple independent variables. Therefore, F-test, which test equations including multiple coefficients together, are used together. The basic equations and hypotheses applied to the F-test are as follows.

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \qquad \text{regression equation}$$

The null hypothesis (H_0) is $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ and the alternative hypothesis (H_1) is $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$.

The results of t-test and f-test in this study are as follows.

Variables	Estimate	SE	t-stat	P-value
Intercept	- 0.0341	0.0135	- 2.5259	0.0137
Gov_con_Ko	2.2611	0.6175	3.6620	0.0005
ER_Ko	- 0.5492	0.1699	- 3.2331	0.0018
China_Im	0.2435	0.0914	2.6637	0.0095

Table 05. t-test and f-test results

The critical values and Probability value (P-value) could be used to evaluate the hypothesis. This study considers the P-value, and it is set at 5 percent of the significant level. If the P-vale is greater than 0.05, the null hypothesis is accepted. This means that the independent variables are not statistically significant for the dependent variable. Conversely, if the P-value is less than 0.05, the null hypothesis is rejected and the alternative hypothesis is accepted. This means that that variable is important to explain the dependent variable. As a result of the test, the P-values of three independent variables, such as Gov_con_Ko, ER_Ko and China_Im were less than 0.05. Therefore, these three are important variables in explaining the dependent variable. The remaining 18 variables were excluded from the regression model because they are not significant to explain the Gwangyang port throughput.

(2) Cointegration Test

Robert Engle and Glive Granger (1987) introduced the problem that linear regression is not the right approach for analyzing time series because of the possibility of spurious correlation. Spurious correlation is when two or more related variables appear to be correlated either by coincidence or by an unknown factor. As a result, misleading statistical relationships between time series variables may appear. The cointegration test is a method used to discover possible correlations between time series variables over a long period of time, and the most famous ones are Engle-Granger test and the Johansen test. If the problem is found as a result of the test, it can be solved by including the error correction term (ECT). ECT is enabled to capture the long-run relationship.

As a result of the Engle-Granger test, it was found that cointegration occurred in the 'China_Im' variable in this regression model. Therefore, ECT was added regarding this variable.

(3) AR and MA

Autoregressive (AR) model is one in which the present value of the dependent variable y depends on the past value of that variable. The AR model of p order is written as AR(p) and can be expressed as follows.

$$y_t = \mu + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + u_t$$

Moving Average (MA) model is one that current errors depend on past errors. The q order MA model is written as MA(p) and can be expressed as follows.

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$

By including AR and MA terms in the regression models, it is possible to improve the accuracy of the forecasting model. However, as a result of checking the appropriate AR and MA terms in this study, there was no need to include AR and MR term. Therefore, AR and MA terms are not included in this model.

3.2.3 Residual Diagnostics Test

The OLS regression model used in this research basically contains five assumptions, which are E (u_t) = 0, Var (u_t) = σ^2 < ∞, Cov (u_i ,u_j) = 0, u_t ~ N (0, σ^2) and Cov (u_t, x_t) = 0. Therefore, it is necessary to test whether this model satisfies these assumptions. If this model satisfies these assumptions, we can call it BLUE (Best Linear Unbiased Estimator), and it means that this model has desirable properties that are consistent, unbiased, and efficient. And since these five assumptions all relate to residuals, we can consider it as the residual diagnostics test.

(1) Assumption: E (u_t) = 0

The first assumption is that the mean value of the residuals is zero. As a result of using the 'mean' function to calculate it, 6.2272e-18 was obtained. Since this value is close to 0, it can be said that this model satisfies the first assumption. In fact, regression models with intercept or constant terms always satisfy this condition, so this does not need to be tested.

(2) Assumption: Var (u_t) = σ^2 < ∞

The second assumption is that the variance of the error is constant and finite over time. If this assumption is satisfied, we can call it Homoskedasticity. Conversely, if the variance of the residuals is not constant, the coefficients are not efficient and standard error estimates could be wrong. This phenomenon can be referred to as the Autoregressive Conditional Heteroskedasticity (ARCH) effect. If the regression model is heteroscedastic, some problems can happen. The most important problems are that some of the significant variables might look insignificant and some of the insignificant variables might look significant. White's general test is generally used to detect of heteroskedasticity. The variance regression for the test is as follows.

$$\hat{u}_t^2 = \alpha_1 + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_4 x_{2t}^2 + \alpha_5 x_{3t}^2 + \alpha_6 x_{2t} x_{3t} + \nu_t$$

The null hypothesis ($\mathrm{H_0}$), which means no ARCH effect, is $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$. The alternative hypothesis ($\mathrm{H_1}$), which means ARCH effect, is $\alpha_2 \neq 0 \; or \; \alpha_3 \neq 0 \; or \; \alpha_4 \neq 0 \; or \; \alpha_5 \neq 0 \; or \; \alpha_6 \neq 0$. As a result of the test, the p-value is 0.8149. Because this value is greater than the five percent significant level, it accepts the null hypothesis and we can say there is no ARCH effect.

(3) Assumption: Cov (u_i ,u_j) = 0

The third assumption is that the covariance between the error terms over time is zero. This means that the error should not repeat. If the errors are correlated with each other, we can say that they are autocorrelated or serially correlated. In the presence of autocorrelation, R^2 could be inflated above its true value. Therefore, if there is autocorrelation in the model, appropriate correction is required. On the other hand, there are three methods for testing autocorrelation. The first method is a graphical test. This method graphs the residuals over time, and if a specific pattern can be found in it, then the model can be estimated to be autocorrelated. Although simple, this method can be inaccurate or difficult to interpret for complex models. The second method is the Dur-bin-Watson (DW) test. However, this method is limited because it can only test for first-order autocorrelation. Therefore, in this study, the Breusch-Godfrey test, which can simultaneously test the autocorrelation up to the r^{th} order, was used. The test statistics are:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \ldots + \rho_r u_{t-r} + \nu_t, \qquad \nu_t \sim N(0, \sigma_\nu^2)$$

The null hypothesis ($\mathrm{H_0}$) is $\rho_1 = \rho_2 = \ldots = \rho_r = 0$, which is No serial correlation while the alternative hypothesis ($\mathrm{H_1}$) is $\rho_1 \neq 0 \; or \; \rho_2 \neq 0 \; or \; \ldots \; or \; \rho_r \neq 0$, which is serial correlation. Since quarterly data were used in this study, the number of the residual (r)'s lags were set to 5 and verified. As a result of the test, the p-value is 0.5133, which is greater than 5 percent. Therefore, it can be confirmed that this model accepts the null hypothesis and there is no serial correlation.

If the regression does not meet assumptions 2 and 3, it will no longer have the characteristics of the BLUE model. Also, if there is heteroscedasticity, some insignificant variables could appear significant, and vice versa. Also, if there is autocorrelation, R^2 could be inflate. Therefore, the following corrections are required for each case.

<table>
<tr><th>No.</th><th>Assumption 2
(ARCH effect)</th><th>Assumption 3
(Serial correlation)</th><th>Correction</th></tr>
<tr><td>1</td><td>×</td><td>×</td><td>Nothing</td></tr>
<tr><td>2</td><td>o</td><td>×</td><td>White correction</td></tr>
<tr><td>3</td><td>×</td><td rowspan="2">o</td><td rowspan="2">Newey-West correction</td></tr>
<tr><td>4</td><td>o</td></tr>
</table>

Table 06. The correction for the ARCH effect and Serial correlation

If there is ARCH effect or Serial correlation in the model, kind of corrections or remedies is needed, which are white correction and Newey-West correction. If there are no ARCH effect and Serial correlation, no action needs to be taken, while if there is ARCH effect and no Serial correlation, white correction is required. In addition, if there is serial correlation regardless of ARCH, Newey-West correction should be performed. As checked in the previous step, there is no ARCH effect and no serial correlation in this regression model, so there is no need to corrections or remedies.

(4) Assumption: u_t ~ N (0, σ^2)

The fourth assumption is that the residuals should be normally distributed. Typically, the Bera-Jarque test is used to determine whether the regression residual follows a normal distribution. The Bera-Jarque test statistic is as follows

$$W = T\left[\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right]$$

where $b_1 = \frac{E[u^3]}{(\sigma^2)^{3/2}}$, $b_2 = \frac{E[u^4]}{(\sigma^2)^2}$ and T is the sample size

The null hypothesis (H_0) is Normally distributed, while the alternative hypothesis (H_1) is No Normally distributed. If this model is not normally distributed, correction should be taken to make it normality. Normality could be improved by using dummy variables for the most extreme residual (outliers).

As a result of the test, the p-value is 0.796, which is greater than the 5 percent significant level. This result value accepts the null hypothesis, so it can be considered that the residual is normally distributed. Meanwhile, if this model does not comply with the normal distribution, it could be improved by adding dummy variables.

(5) Assumption: Cov (u_t, x_t) = 0

The fifth assumption is that x variables are not correlated with the error terms. However, if assumption 1 (E(u_t) = 0) holds and there are enough independent variables, the OLS regression analysis always satisfies this assumption and does not require additional testing.

As a result of the tests for the five assumptions and corrections, this regression can be considered a BLUE model because it satisfies all assumptions.

3.2.4 Stability Diagnostics Test

In this section, the linearity and stability of the regression model will be checked through the Ramsey's RESET test and the CUSUM test.

(1) Ramsey's RESET Test

Since this study uses the OLS regression model, in addition to the five assumptions discussed above, it also includes the assumption that the model should be linear. However, since this assumption is not always met, it is necessary to test whether the model is linear or not. As the test method, Ramsay 1969 RESET is generally used. The test statistic is as follows.

$$\hat{u}_t = \beta_0 + \beta_1 \hat{y}_t^2 + \beta_2 \hat{y}_t^3 + \ldots + \beta_{p-1} \hat{y}_t^p + \nu_t$$

The null hypothesis (H_0) is $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$, which means the model is linear. The alternative hypothesis (H_1) is $\beta_1 \neq 0 \; or \; \beta_2 \neq 0 \; or \; \ldots \; or \; \beta_{p-1} \neq 0$, which means the model is not linear. In this research, the test was conducted with one square term, 'residual = 1 + y^2 ' . The result is as follows.

Variables	Estimate	SE	t-stat	P-value
Intercept	- 0.0023198	0.00629	- 0.36881	0.71328
y_fit_2	0.81967	0.94747	0.86511	0.38967

Table 07. Ramsey's Reset Test result

This model may not be linear at first. However, if all variables are converted into natural logs before proceeding with the test, the function can be changed as linear. However, the OLS regression analysis used in this study cannot be used if the regression is not linear even if all variables is changed to the natural logarithm.

(2) CUSUM test

The regression model also contains the assumption that the parameter is constant over the entire sample. This means that the parameter must be constant not only for the period of data used to estimate the model, but also for any partial period used for forecasting. The test for this assumption is done by dividing the data into two sub-periods and comparing the RSS (residual sum of squares) of three regressions, each of the sub-periods and for whole periods. There are mainly two ways to divide the sample for stability testing: dividing the data according to the obvious structural changes and breaking point or any known significant historical event such as a financial crisis.

For the stability test, the Chow test or the CUSUM test is mainly used. In this study, the CUSUM test, which is simple and can clearly know the result with a graph, was used. The CUSUM test is a parameter stability test method of an estimated model based on the cumulative sum of the residuals (Brooks, 2019). The test results are as follows.

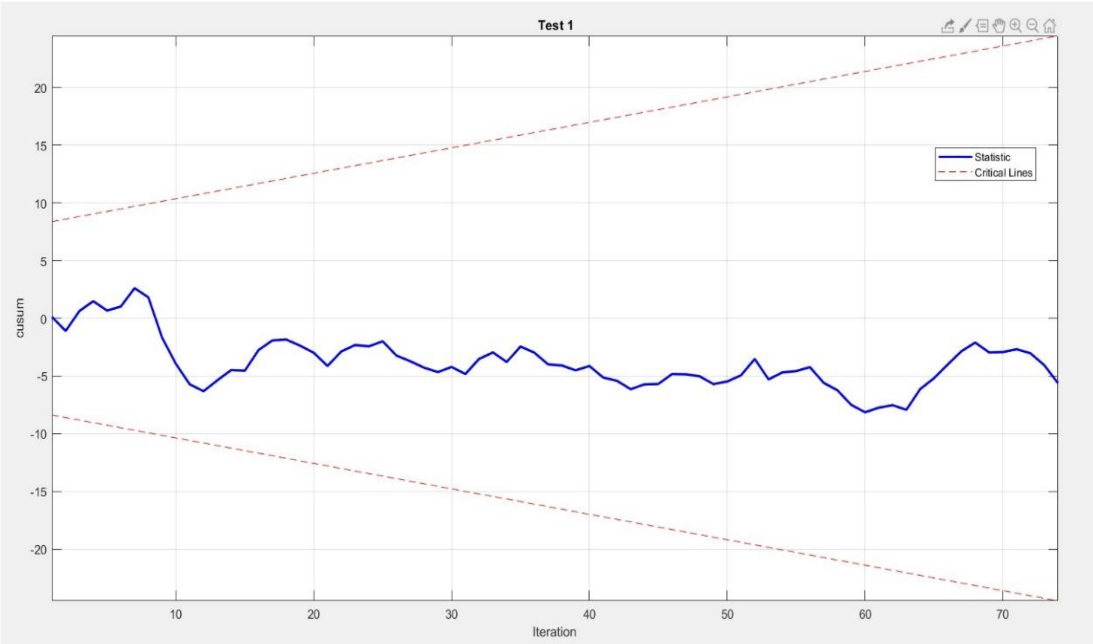


Figure 13. CUSUM Test result

As a result of the CUSUM test, the parameter is stable because the blue line (statistic line) does not exceed the red lines (threshold lines).

Until now, it was confirmed that this regression model is BLUE by the residual diagnostics test, and it was also found that the model is linear and have stability by the stability diagnostics test. Therefore, the following final model could be obtained.

Linear regression model:

GY_PT ~ 1 + Gov_con_Ko + ER_Ko + China_Im + ect_China_Im

And the estimate coefficients are

Variables	Estimate	SE	t-stat	P-value
Intercept	- 0.03188	0.01142	- 2.7917	0.00667
Gov_con_Ko	2.191	0.52419	4.1797	7.9078e-05
ER_Ko	- 0.32958	0.14912	- 2.2102	0.03018
China_Im	0.28287	0.07684	3.6816	0.00043778
ect_China_Im	- 0.40588	0.07614	- 5.3311	1.0179e-06

Table 08. Final regression model about Gwangyang Import and export

Three independent variables, Gov_con_Ko, ER_Ko and China_Im, of the 21 variables and one error correction term for China_Im are estimated to be significant for the dependent variable. This is the OLS regression model for the Gwangyang port throughput.

3.3. ARIMA modelling process

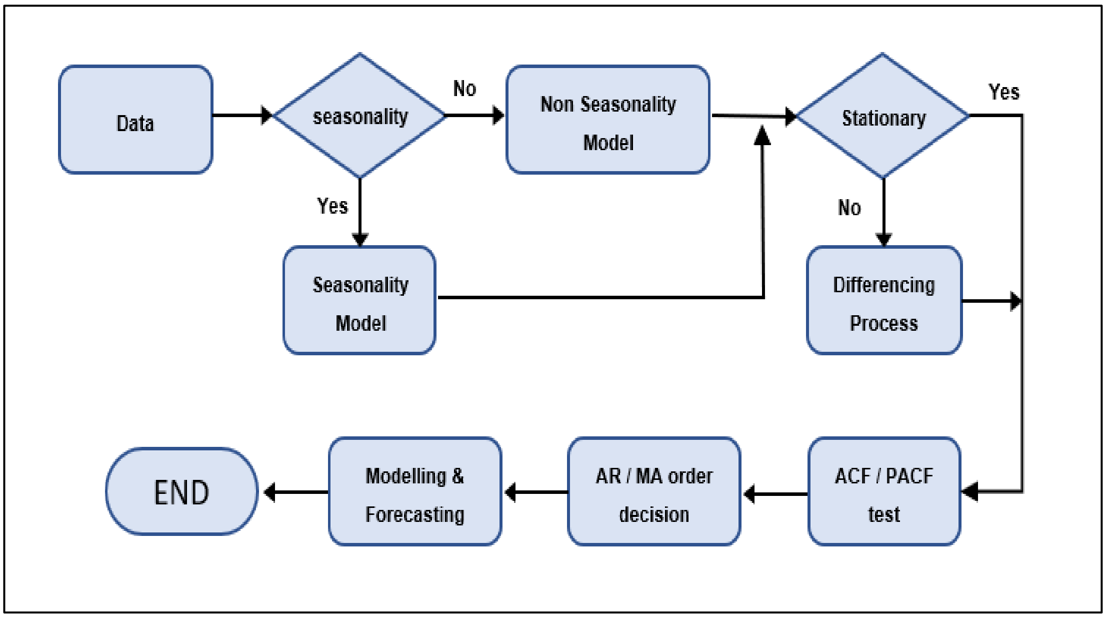


Figure 14. Flow chart of ARIMA model (Source: Author)

(1) Seasonality

Container throughput may have seasonality depending on its demand. In this case, it is possible to increase the accuracy of prediction by using the seasonality models, for example SARIMA, rather than the commonly used non-seasonality ARIMA model. Kyung-Chang Min et al (2014) described that using the SARIMA model for seasonal container volumes in Korea's port can improve the prediction accuracy. Therefore, in order to determine which model to use, it is necessary to first check whether the container volume of GY port has seasonality.

Figure 15 below shows the container volume of GY port divided by quarter. As can be seen from the graph, it cannot be seen that a particular quarter has a specific trend, such as always having a higher or lower volume than another quarters. Therefore, the container volume of GY port does not have seasonality, and it is appropriate to use the non-seasonality model, ARIMA.

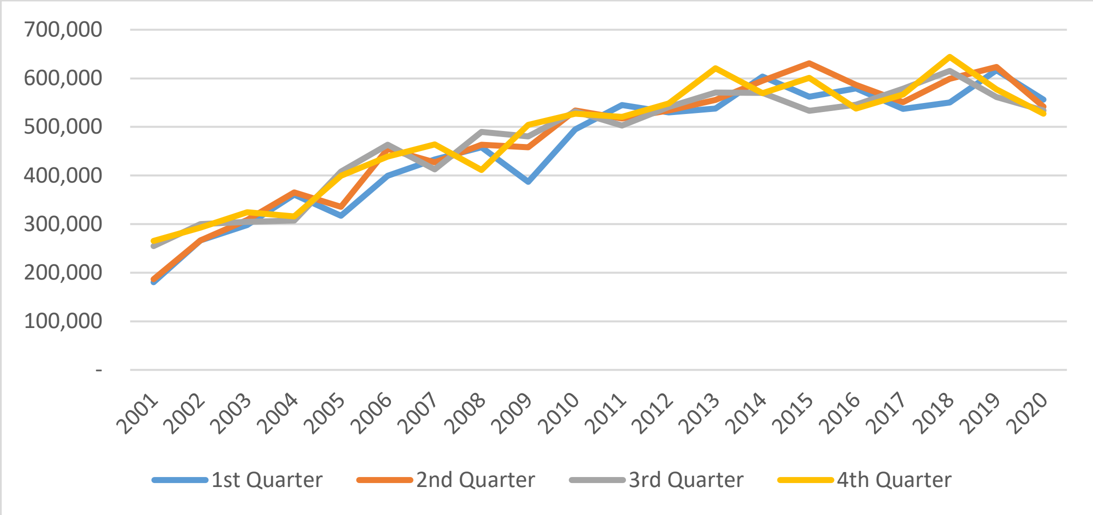


Figure 15. GY port quarterly container throughput (2001~2020)

(2) Stationarity

A non-stationary model may exhibit undesirable characteristics, for example the previous value of the error term is not decrease over time. Therefore, it is necessary to test the stationarity before applying the ARMA model and take measures such as differencing in case of non-stationary model. To test the stationary of the model, the unit root test, which is the same method used in the OLS regression model, was used, and the results are as follows.

Variable	difference	ADF		PP		KPSS	
		P-Value	Stat	P-Value	Stat	P-Value	Stat
GY_PT	I (0)	0.851	0.632	0.851	0.632	0.010	1.213
	I (1)	0.001	- 10.844	0.001	- 10.844	0.101	0.016
	I (2)	0.001	- 18.175	0.001	- 18.175	0.100	0.005

Table 09. Unit Root Test Result for ARIMA modelling

As a result of the test, P-value of ADF and PP is greater than 5% in I (0) and less than 5% after I (1). In addition, P-value of KPSS is less than 5% in I (0) and greater than 5% after I (1). It means that GY_PT is not stable at I (0), and it becomes stable after the 1st differencing. Therefore, I (1) is applied for this variable.

(3) ACF & PACF test

Graphical plots of ACF and PACF are available for ARMA model selection. The autocorrelation function (ACF) is the correlation between the current value and the value at previous time spot. Meanwhile, the partial autocorrelation function (PACF) is the correlation between the current value (y_t) and a k periods ago value (y_{t-k}), after removing the effects of y_{t-k+1}, y_{t-k+2},..., y_{t-1}. We can find order of the AR process using the PACF plot and order of the MA process using the ACF plot.

In general, an AR process has a geometrically decaying ACF and a number of non-zero values of PACF, which is AR order. A MA process has number of non-zero values of ACF, which is MA order and a geometrically decaying PACF. An ARMA process ha: a geometrically decaying ACF and PACF (Brooks, 2019). The results of ACF and PACF test for container volume in GY are as follows.

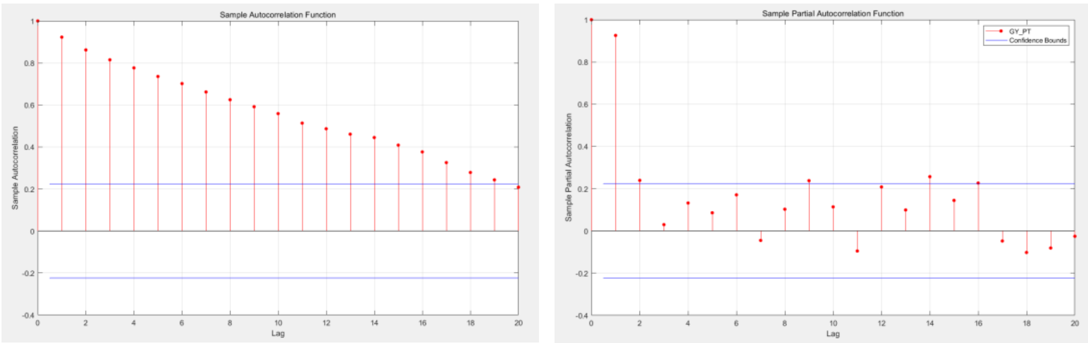


Figure 16. ACF and PACF test results

Looking at the test results, the ACF declines slowly and only the 1st PACF coefficient is significant, while others are not significant. This is the form that appears in the AR (1) model.

Therefore, ARIMA (1,1,0) is a suitable model for GY port container volume.

3.4. Forecasting

Based on the model created in the previous step, the future container throughput of GY port is predicted in the second step. As the first step, in-sample and out-of-sample forecasting are performed to test the model. After that, forecasting the future port throughput of GY port is conducted by using Matlab program. In addition, the mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE), which are the most common measurement methods, are used to determine how accurate the forecast is at each stage.

Forecast error, the difference between the actual values and the predicted values, can be positive or negative, so simply summing them can cancel each other out. Therefore, before summing the forecast errors, square them or take the absolute value. The former is MSE and the latter is MAE and RMSE. Since this is the sum of forecast errors, the smaller this value is, the more accurate the forecasting model is. When the s-step-ahead forecasting values at time t is $f_{t,s}$ and the actual value is y_t, MSE, MAE and RMSE can be expressed as follows, respectively (Brooks, 2019).

$$\text{MSE} = \frac{1}{T-(T_1-1)}\sum_{t=T_1}^{T}(y_{t+s}-f_{t,s})^2$$

$$\text{MAE} = \frac{1}{T-(T_1-1)}\sum_{t=T_1}^{T}\left|y_{t+s}-f_{t,s}\right|$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=T_1}^{T}\left|y_{t+s}-f_{t,s}\right|}{T-(T_1-1)}}$$

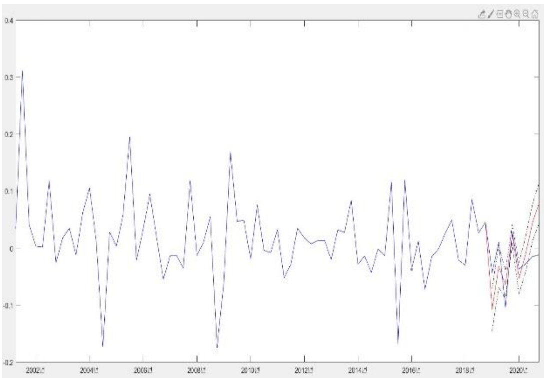
where T is the total sample size, T_1 is first observation of out-of-sample forecast.

On the other hand, MPAE can be defined as follows, and the closer the value is to zero, the more accurate the model is.

$$\mathrm{MAPE} = \frac{100}{T-(T_1-1)}\sum_{t=T_1}^{T}\left|\frac{y_{t+s}-f_{t,s}}{y_{t+s}}\right|$$

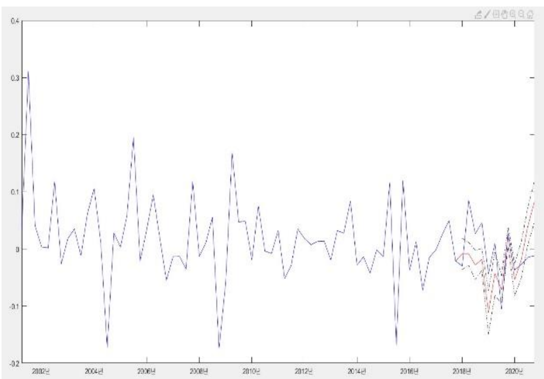
3.4.1 in-sample and out-of-sample forecasting

The in-sample is the data used to estimate the model coefficients, while the out-of-sample is the data outside of in-sample. Out-of-sample forecast is a method of testing forecasting performance by dividing the given data into in-sample and out-of-sample, predicting data for the out-of-sample period, and comparing it with actual data. Splitting the sample is at the discretion of the researcher (Brooks, 2019). In this study, a total of 80 data were divided into 72 in-sample, 8 out-of-sample and 68 in-sample, 12 out-of-sample.



MSE	0.0025
MAE	0.0414
RMSE	0.0496
Bias Proportion	0.0417
Variance Proportion	0.1933
Covariance Proportion	0.9019

Case 1: no. of out-of-sample, 8



MSE	0.0031
MAE	0.0474
RMSE	0.0557
Bias Proportion	0.0476
Variance Proportion	8.9475e-4
Covariance Proportion	1.0381

Case 2: no. of out-of-sample, 12

Table 10. Forecast result using OLS regression model (out-of-sample 4 & 12)

In the forecasting results, MSE, MAE, and RMSE are methods for evaluating whether predictions are accurate, and since each value indicates the degree to which an error appears, the smaller the value, the more accurate the prediction is. In addition, to use this forecast model, the bias and variance proportion should be close to 0, the covariance proportion close to 1, and the covariance proportion should be larger than the variance proportion. Since the MSE, MAE, and RMSE values in Case 1 are smaller than Case 2, the model in Case 1 can be considered more accurate than Case 2. In general, short-term predictions are more accurate than long-term predictions.

Meanwhile, the results of comparing 8 out-of-samples (2019Q1 ~ 2020Q4) using the ARMA model with the actual values are as follows.

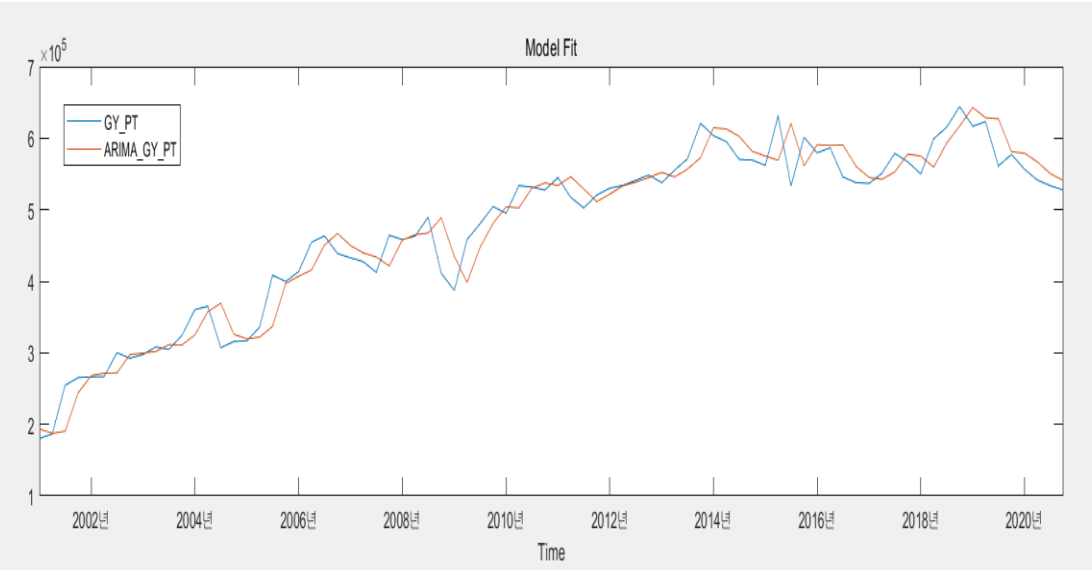


Figure 17. Forecast result using ARIMA model

The MSE of this ARMA model is 0.0028, the MAE value is 0.0419, and the RMSE is 0.0528. This value is larger than the result value of the forecasting model using OLS regression. Therefore, it can be judged that it is more appropriate to use the OLS regression model to forecast the GY port throughput.

The results of forecasting the GY port throughput for the next two years (2021 Q1 ~ 2023 Q4) using the OLS regression model are as follows. As can be seen from the results, the GY port throughput was predicted to be stable for eight months without significant change.

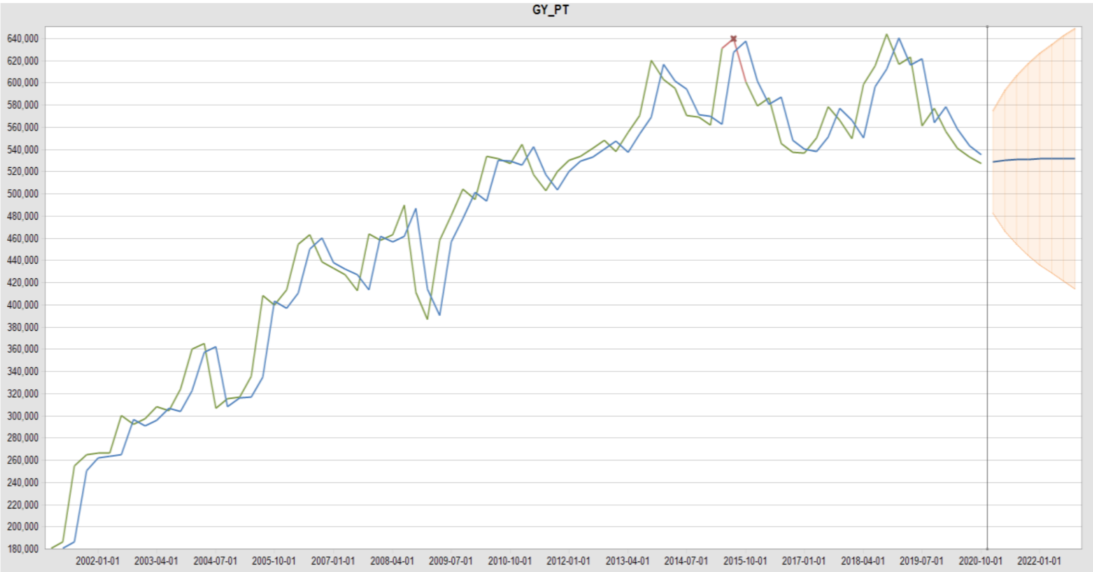


Figure 18. GY port throughput forecasting result

4. Results and findings

The results of the OLS regression modelling using 21 independent variables to forecast container throughput of Gwangyang port are as follows.

$$GY\_PT \sim -0.03188 + 2.191 \times Gov\_con\_Ko - 0.32958 \times ER\_Ko$$
$$+ 0.28287 \times China\_Im - 0.40588 \times ect\_China\_Im$$

Looking at the above results, the Korean government's consumption, the Korean exchange rate, and China's imports are the significant factors in the container throughput of GY port. In addition, it is possible to know the degree of influence and the positive/negative relationship between each term and GY port throughput. When the Korean government's consumption increases by 1, GY port throughput increases by 2.191. When China's imports increase by 1, GY port throughput increases by 0.28287. On the other hand, when Korea's won-dollar exchange rate increases by 1, GY port throughput decreases by 0.32958.

Using the ARMA, which is a representative method of univariate models, and OLS regression, which is multivariate models, the GY port throughput prediction model was made and the MSE, MAE, and RMSE values were compared. As a result, the OLS regression model was more suitable for predicting the GY port throughput. In general, multivariate models using external variables show more accurate results than univariate models.

Meanwhile, as a result of forecasting GY port throughput for the next two years using the OLS regression model, it was found that a stable container volume was secured without significant change. Therefore, this result can be utilized for port master planning.

5. Conclusion

5.1 Conclusion

Through OLS regression modelling, it was found that Government consumption, China import, and Korea exchange rate as independent variables affect GY container port throughput. In particular, as government consumption increases, the GY port throughput increases, and the increase in import of China, the No. one trading partner with Korea, also increases the GY port volume. On the other hand, there is a negative relationship between the exchange rate and the GY port throughput, because as the exchange rate increases, that is, as the value of the domestic currency decreases, the price competitiveness of domestic products increases and exports increase. In this way, by identifying the independent variables that have a major impact on the GY port throughput, it is possible to roughly predict the future prospects of the GY port throughput according to the increase or decrease of this index.

Also, ARMA model and OLS regression model were used for forecasting modelling. This would be a good attempt to start research on the prediction of the GY port container volume in the future, considering that there have been few studies that make prediction modelling for GY port. And, by describing the suitability of the OLS regression model and the detailed modelling process in this study, it will become a standard for improving and supplementing this method in the future.

Finally, as a result of forecasting the GY port throughput, it was forecasted that the container volume would not increase significantly in the future and would be stable. This is a bad situation for container terminal operators' earnings improvement and the competitiveness of the GY port container terminal. Therefore, YGPA should make a master plan in the direction of improving the competitiveness of the current container terminal rather than expanding it. In addition, it can be interpreted that a special plan, different from the previous policies, is required to achieve an improvement in the volume of cargo beyond this prediction. Therefore, the increase in cargo volume through the expansion of FTZ within the port area promoted by YGPA could be a new attempt that has never been done before, and it is necessary to watch with interest whether it will lead to an improvement in container throughput. However, in implementing the plan, another in-depth study on this will be necessary.

The fact that YGPA which operates GY port has the methods to predict the port's cargo volume, means that it can secure an important foundation for port operation and mid- to long-term planning on its own. In a rapidly changing global environment, it is very important to forecast container throughput (Chan et al. 2019). Although this study has not been carried out extensively, if the prediction model suitable for GY port is continuously improved and supplemented in the future based on the OLS regression model established in this study, a more accurate and helpful means for policy planning will be secured and is expected to be easy to use.

5.2 Limitations

In order to make an accurate model and predict future container throughput, the most basic and important thing is to secure a variety of reliable data. However, in this study, data related to port productivity, which is important for container liner shipping companies to choose a port, could not be obtained. In addition, GY port's strategy is to expand the free trade zone near the port to achieve more than 3 million TEU in the future. Because of this, it is important to understand the correlation between the free trade zone and the GY port volume. However, it was difficult to secure the reliability data about FTZ's import and export volume because the authority has only managed data since 2011 and even if it is data after 2011, there is no objective evidence for this. Lack of reliable data sources or inaccuracies in data collection can lead to poor quality forecasting, even with sophisticated forecasting methodologies (Peng and Chu, 2009).

Moreover, only ARIMA and OLS regression methods were used to forecast in this study. It is difficult to build more advanced methods with comprehensive testing of all prediction methods. Therefore, it will be necessary to determine the best method for container throughput forecasting by examining various models applied to the GY port.

5.3 Scope for future research

As of 2020, the import and export volume created in the free trade zone near GY port is 32% of the total volume. In the future, the YGPA plans to create new container cargo by more than doubling the free trade zone. However, in this study, it was found that the FTZ volume is not a significant independent variable for the GY port throughput. There are various factors for this, but it will also include the problem of the period and reliability of collecting data on the amount of volume in the FTZ. Therefore, it is necessary to secure more reliable data in the future and recheck the correlation between the FTZ and total volume. In addition, as mentioned above, in this study, only the ARIMA model and OLS regression analysis were used to predict the GY port throughput. It is necessary to improve these ARIMA and OLS regression model in the future, but it is also necessary to try applying various methods such as grey forecast and neural network. Future researches are needed to make and utilize an optimal forecasting model by improving existing models and applying new methods.

Reference

Anggraeni, W., Vinarti, R. A., & Kurniawati, Y. D. (2015). Performance comparisons between arima and arimax method in moslem kids clothes demand forecasting: Case study. *Procedia Computer Science*, *72*, 630-637.

Brooks, C. (2019). *Introductory econometrics for finance*. Cambridge University Press.

Chan, H. K., Xu, S., & Qi, X. (2019). A comparison of time series methods for forecasting container throughput. *International Journal of Logistics Research and Applications*, *22*(3), 294-303.

De Langen, P. W., Van Meijeren, J., & Tavasszy, L. A. (2012). Combining Models and Commodity Chain Research for Making Long-Term Projections of Port Throughput: an Application to the Hamburg-Le Havre Range. *European Journal of Transport & Infrastructure Research*, *12*(3).

Dragan, D., & Kramberger, T. (2014, June). Forecasting the Container Throughput in the Port of Koper using Time Series ARIMA model. In *Proceedings of the conference ICLST* (Vol. 14, pp. 19-21).

Dragan, D., Kramberger, T., & Intihar, M. (2014). A comparison of methods for forecasting the container throughput in north Adriatic ports. In *Conference IAME*.

Du, Y. (2013). A Prediction of the Container Throughput of Jiujiang Port Based on Grey System Theory. In *The 19th International Conference on Industrial Engineering and Engineering Management* (pp. 51-59). Springer, Berlin, Heidelberg.

Eskafi, M., Kowsari, M., Dastgheib, A., Ulfarsson, G. F., Taneja, P., & Thorarinsdottir, R. I. (2020). Mutual information analysis of the factors influencing port throughput. *Maritime Business Review*.

Gaur, P. (2005). *Port planning as a strategic tool: a typology*. UA, ITMMA.

Gosasang, V., Chandraprakaikul, W., & Kiattisin, S. (2011). A comparison of traditional and neural networks forecasting techniques for container throughput at Bangkok port. *The Asian Journal of Shipping and Logistics*, *27*(3), 463-482.

Hui-yuan, L. L. L. J., & Shuan-zhu, Z. H. A. N. G. (2009). Container throughput forecast for Shenzhen Port based on GM (1, 1) model. *Port & Waterway Engineering*, *2*.

Hutcheson, G., & Sofroniou, N. (2015). *The multivariate social scientist*. SAGE Publications Ltd.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*.

Intihar, M., Kramberger, T., & Dragan, D. (2017). Container throughput forecasting using dynamic factor analysis and ARIMAX model. *Promet-Traffic&Transportation*, *29*(5), 529-542.

Jarrett, J. E. (2015). Throughput Port Demand Forecasting. *Int J Econ Manag Sci*, *4*(293), 2.

Julong, D. (1989). Introduction to grey system theory. *The Journal of grey system*, *1*(1), 1-24.

Kongcharoen, C., & Kruangpradit, T. (2013, June). Autoregressive integrated moving average with explanatory variable (ARIMAX) model for Thailand export. In *33rd International Symposium on Forecasting, South Korea* (pp. 1-8).

Kotcharat, P. (2016). A forecasting model for container throughput: empirical research for Laem Chabang Port, Thailand.

Lewis-Beck, Michael S. 1980. *Applied Regression: An Introduction*. Beverly Hills, CA: Sage.

Liu, L., & Park, G. K. (2011). Empirical analysis of influence factors to container throughput in Korea and China ports. *The Asian Journal of Shipping and Logistics*, *27*(2), 279-303.

Memos, C. D. (2004). Port planning. *Port engineering: planning, construction, maintenance, and security*, 7-64.

Michail, N., Melas, K. D., & Batzilis, D. (2020). The Relationship Between Container Shipping Trade and Real GDP Growth: A Panel Vector Autoregressive Approach. *Available at SSRN 3724480*.

Min, K. C., & Ha, H. K. (2014). Forecasting the Korea's port container volumes with SARIMA model. *Journal of Korean Society of Transportation*, *32*(6), 600-614.

Parola, F., Satta, G., Notteboom, T., & Persico, L. (2021). Revisiting traffic forecasting by port authorities in the context of port planning and development. *Maritime Economics & Logistics*, *23*(3), 444-494.

Peng, W. Y., & Chu, C. W. (2009). A comparison of univariate methods for forecasting container throughput volumes. *Mathematical and computer modelling*, *50*(7-8), 1045-1057.

Qiuhong, Z. (2009, July). Application of grey model in forecasting the port of Qinhuangdao's throughput. In *2009 IITA international conference on services science, management and engineering* (pp. 57-60). IEEE.

Shu, M. H., Huang, Y. F., & Nguyen, T. L. (2013). Forecasting models for the cargo throughput at Hong Kong Port and Kaohsiung Port. *Recent Researches in Applied Economics and Management. WSEAS Press, Chania, Crete Island, Greece*, *507*.

Vittinghoff, Eric, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. 2005. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer.

Witt, S. F. and Witt, C. A. (1992), Modelling and forecasting demand in tourism, Academic Press Ltd.