e-Navi

Available online at http://www.e-navi.kr/
e-Navigation Journal

Original article

A Comparative Study of Multivariate Forecasting Models for Container Throughput

Awah, P. C.,[a] Mabika, C. A.,[b] Hwayoung Kim[c*]

[a] Department of Maritime Transportation, Mokpo National Maritime University, Mokpo, Korea, apennq@g.kmou.ac.kr

[b] Department of International Trade, Korea Maritime and Ocean University, Busan, Korea, mabikaanse@g.kmou.ac.kr

[c] Division of Maritime Transportation, Mokpo National Maritime University, Mokpo, Korea, hwayoung@mmu.ac.kr* Corresponding Author

Abstract

Forecasting port container throughput is crucial due to its impact on economic development. Socio-economic factors, which introduce uncertainty, are increasingly integrated into throughput forecasting. The complexity of common multivariate forecasting models significantly affects accuracy, yet few studies compare their performance on the same time series for throughput modeling. This study implements, evaluates, and compares the performance of eight multivariate forecasting models for port throughput within a proposed multiple-input single-output (MISO) system, chosen for their frequent use in container throughput research. It investigates two data preprocessing approaches: Random Forest Variable Importance Method (RF-VIM) and a Multi Lagged Value approach. The comparison uses six error metrics: normalized root mean squared error, mean absolute error, mean absolute percentage error, mean error, and root mean percentage error. Performances are discussed, and recommendations for adopting a suitable model are provided.

Keywords: Port Throughput, Container Forecast, Machine Learning, Multivariate Forecasting.

https://doi.org/10.52820/j.enavi.2024.22.009

1. Introduction

Ports are becoming increasingly essential to the operations of international trade activities due to the rapid expansion of economic globalization (Notteboom, 2016). The COVID-19 pandemic has had a significant impact on globalization, with serious repercussions for the container shipping industry. The pandemic disrupted supply chains, reduced trade volumes, and created uncertainties that affected container shipping lines, ports, and terminals. Several key effects of the pandemic were observed. These effects included imbalanced container availability in high-demand locations, blank sailings, and schedule disruptions, along with freight rate volatility. All these factors added complexity and uncertainty for businesses relying on container shipping. Consequently, adjustments to shipping routes and vessel deployment became necessary to accommodate changes in trade patterns, especially for industries with higher demand, such as e-commerce and medical supplies. These challenges require resilience, adaptation to the evolving situation, and the optimization of operations to make informed decisions and enhance efficiency. Forecasting port throughput is crucial because it enhances the port's economic development, logistical competitiveness, and operational efficiency. Furthermore, the port logistics industry and other stakeholders can derive significant benefits for port operations from comprehending and analyzing forecasting schemes. It is noteworthy that a poor forecasting scheme would misguide port management decisions, thus negatively impacting development prospects. Thus, as a foundation for port development schemes, accurate throughput forecasting is necessary to map out uncertainty and improve judgment for port development (Chen et al., 2016). Univariate forecasting modes have achieved significant success in throughput forecasting and are commonly used in practice. Their single-input and single-output (SISO) system of making predictions based only on the lag observations of container throughput series has caused a lot of debate due to their limitations under complex and non-linear data patterns (Mishra et al., 2020). Container throughput is often influenced by various socio-economic factors such as gross domestic product (GDP), gross national product (GNP), total imports and exports, and population size, to name a few. Given the primarily non-linear nature and the potentially complex interactions of these variables, representing multiple variables using univariate methods is challenging Univariate methods in container throughput forecasting may be limited because they oversimplify, fail to capture complex connections, and rely only on historical data without considering external factors (Geng, 2015; Huang et al., 2022; Shankar et al., 2021; Zou et al., 2022).

A multivariate forecasting mode within a multiple-input and single-output system is preferred for throughput forecasts that provide accurate insights to assist port management. Although several multivariate forecasting models have been developed and widely applied in the literature, there is currently no derivation of these forecasting schemes available for comparative study. Given the container shipping industry's recent recovery from the profound impacts of the COVID-19 pandemic (Gu et al., 2023; Huang et al., 2022; Tok & Ece, 2022) such a framework would assist future practitioners in determining which scheme is best suited for a given forecasting scenario. This includes considering socio-economic factors that account for the externalities of container throughput. Providing port management, stakeholders, and industry personnel with the required tools to make informed decisions and optimize operations would contribute to a resilient container shipping industry. In this regard, this study seeks to compare a total of eight multivariate forecasting modes within a MISO (multi-input, single-output) forecasting system. Namely, Multiple Linear Regression (MLR), Support Vector Regression (SVR), Long Short-Term Memory Neural Networks (LSTM), The Gray Model GM(1, N), Least Squares Support Vector Regression (LSSVR), Multilayer Perceptron (MLP), Random Forest (RF), Multivariate Adaptive Regression Splines (MARS), and for container throughput forecasting (see literature review for a detailed explanation of the models). As a result, discussions, and suggestions for adopting a suitable mode would be drawn upon.

The rest of the paper is organized as follows: The first section reviews the theoretical background regarding the selected forecasting modes and their potential. Next, a research methodology is developed to provide a succinct description of the data and methods used to derive forecasts based on the models; next, the results and analysis section details the experiment. Finally, the paper presents a discussion of the results, followed by a conclusion that includes directions for further research. Figure 1 depicts the workflow of the study.

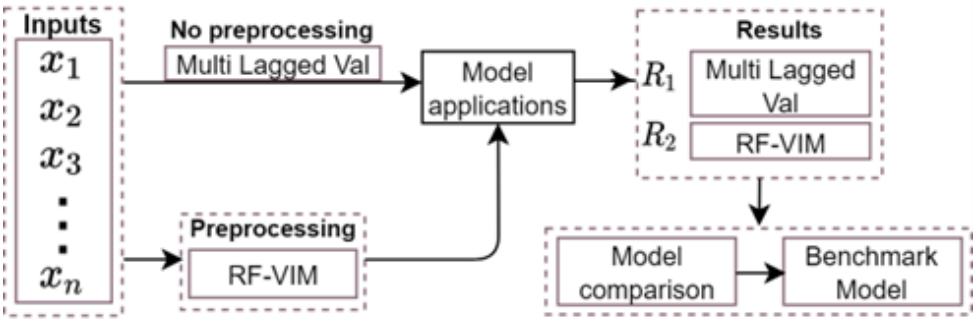

Figure 1: Depiction of the study workflow Source: Author(s)

2. Literature

In recent years, the massive growth of low-dimensional big data has meant that temporally collected data typically has complex and non-linear properties. Traditional time series modes struggle to deal with the complexity and uncertainty inherent in multivariate forecasting modes (Box et al., 2015). When it comes to time series modeling, high-dimensionality, dynamicity, and uncertainty are vital terms determining the methods for feature representation of data and forecasting modes, which are often the prime focus of researchers (Du et al., 2020) Large-scale studies evaluating time series models in container throughput modeling have predominantly focused on the single-input and single-output forecasting systems. These systems typically involve using container throughput data and its lag observations or trends as inputs to predict an expected output, which is also container throughput (Chan et al., 2019; Huang et al., 2021; Xu et al., 2022). This study aims to adopt a multivariate model for forecasting container throughput by drawing conclusions from comparisons of the frequently used models. Numerous models are proposed in the literature for multivariate forecasting systems, and it would be daunting to consider them all in a single study. These prior studies have attested that multivariate models enhance container throughput forecasting by capturing complex relationships between multiple variables, increasing forecast reliability, and supporting informed decision-making in port capacity planning. They achieve this by simultaneously considering economic, environmental, and social factors, thereby improving strategic decision-making (Li & Xu, 2011; Lee et al., 2021; Yang, 2020; Ding, 2019; Gosasang et al., 2010; Tang et al., 2019; Awah et al., 2021; Geng et al., 2015). Multivariate models are used in other fields of study. For instance, in the health science, they are employed in the prediction of Covid-19 occurrences (Afshari Safavi, 2022) and in forecasting animal infectious diseases based on meteorological data (Muñoz-Organero & Queipo-Álvarez, 2022). In life sciences, the utilization of sociodemographic characteristics and medical history allows for accurate prognosis prediction of Covid-19 patience after diagnosis (An et al., 2020) Table 1. Below depicts prior studies on container throughput based on multivariate mode applications.

Table 1: Summary of Prior Studies on Container Throughput Based on Multivariate Models.

Reference	Models	Data type	Sample size	Main Findings	Research case (ports)
(Li & Xu, 2011)	Gray Model (1,1), Multiple Linear Regression, Exponential based on PMVF (Pretreatment, modeling, verification, and feedback forecast)	GDP, Foreign trade volume, total fixed investment, and transportation investment (MISO scheme)	12 data points (1995-2007)	The PMVF-based framework can be an effective mechanism for forecasting throughput within any forecasting scheme.	Shanghai Port, China
(Lee et al., 2021)	TS-decomposition, LSTM	Throughput, Import-Export	212 data point-	Based on results, a prediction mode within	Busan port, South Korea

		price, import-export volume,	monthly (2003-2020)	a variable decomposition scheme has a positive effect on forecasting performance	
(Yang, 2020)	CNN, LSTM, Hybrid	5 port assessment, lag observations of container throughput.	212 data point-monthly (2001-2019)	Making use of a convolutional layer for the feature extraction within the mixed precision scheme based on CNN-LSTM increases performance of predictions.	Five Taiwan ports: Anping, Hualien, Kaohsiung, Taichung, and Suao
(Ding, 2019)	BP neural network and Support Vector Regression	Throughput, Import-Export price, import-export volume, and to GDP	Yearly data	The combined model based on neural network and support vector machines greatly improved forecast accuracy.	Ningbo and Wenzhou ports, China
(Gosasang et al., 2010)	Neural Networks	GDP, word GDP, exchange rate, population, inflation rate, interest rate and fuel prices (MISO scheme)	108 data points based on monthly data (Jan 1999 - Dec 2008)	One of the first of studies to apply the neural network model for container throughput forecasting. The study set the base for future study modifications and adaptations based on its results which were quite decent at the time.	Bangkok port, Thailand
(Tang et al., 2019)	Grey model, triple exponential smoothing model, multiple linear regression model, and BP-neural network model.	Total retail sales of consumer goods, GDP, import & export volume, output of secondary industry and total fixed assets investments	21 data points yearly (1990 - 2011) A comparison between growth and raw datasets are used for prediction	Results depicts the raw datasets to outperform the growth datasets in container throughput forecasting based on all four models.	Shanghai and Lianyungang Ports, China
Awah et al. (2021)	Multilayer Perceptron, Random Forest and	New variable application. – ship turnaround time, container dwell time, average vessel depth, yard storage, berth productivity, custom declaration, throughput	156 observations monthly (Jan 2008 – Dec 2020)	Hybrid MLP-RF model outperformed competing models on throughput forecasting, results assist ports in port development projects that attracts shippers to port.	Port of Douala, Cameroon
Geng et al. (2015)	MARS-RSVR and a chaotic simulated annealing particle swarm optimization algorithm	Port throughput data and socio-economic variables (MISO scheme)	35 data points based on yearly data (1978 - 2013)	They proposed algorithm greatly improves performance accuracy and outperforms other combination modes based on ARIMA and SVR. It is also a valid approach for container throughput prediction.	Shanghai port, China

To begin with, multiple linear regression (MLR) analysis. This is a traditional model that is widely used in numerous research domains to explore the link between any given two or more independent variables and a dependent variable (Eberly, 2007; Peter et al., 2019; Puntanen, 2013). The MLR system allows users to estimate the model's variation and the proportional influence of each explanatory variable on the total variance. The MLR has seen success in its application in the container throughput forecasting literature; (Li & Xu) produced good forecast results through a set of selective hybrid approaches based on the MLR model. The MLR model is often based on strict assumptions when in use, such as that a linear relationship between predictor and predicted variables must be established, predictor variables should not exhibit multicollinearity, and the variance of the residuals is constant. The MLR is formulated as in equation 1.

$$yi = \beta_0 + \beta_1 xi_1 + \beta_2 xi_2 \ldots + \beta_d xi_d + \varepsilon \quad (1)$$

where yi is the predicted variable (container throughput), β_0 is the y-intercept, i.e., the values of y when both xi_1 and xi_2 is 0. β_1 and β_2 are the regression coefficients which represent the change in y relative to a one-unit change in xi_1, xi_2, and xi_d, and $\beta\,\beta_1$, β_2, ...and β_d are slope coefficients for each predictor variable and ε the model residual (error).

Support Vector Machines (SVMs), proposed by Vapnik and colleagues in 1992(Boser et al., 1992) are a class of algorithms used for classification, regression, and other applications. SVMs utilize optimization techniques, statistical learning theory, and kernel functions to analyze data. They work with both linear and non-linear classification by transforming inputs into high-dimensional feature spaces using kernel functions. In cases where data is unlabeled, an unsupervised learning approach is employed, such as the support clustering algorithm (Ben-Hur et al., 2002). Support Vector Regression (SVR) extends SVMs to regression problems, aiming to minimize prediction error by finding a hyperplane that maximizes the margin between the hyperplane and the nearest data points. For non-linear problems, the kernel trick, (Aizerman, 1964) is employed to transform data into higher-dimensional spaces using kernel functions like linear, polynomial, Gaussian radial basis functions, and sigmoid functions. The SVR formulation is commonly expressed as equation (1).

$$f(x) = \left(w \cdot \hat{f}(x)\right) + b \quad (2)$$

where w is the weight vector of the function $\hat{f}(x)$, and x is an input vector. The kernel function $\hat{f}(x)$, transforms the datasets into a high-dimensional space, and depending on the kernel function, the transformation can either be linear or non-linear.

In a SVR model (see Figure 1.), with a threshold ε, as ε increases, the prediction becomes accurate and thus less sensitive to errors, with an objective function and constraint, as shown in equations (3), (4).

$$\text{Minimize: } \frac{1}{2}||w||^2 \quad (3)$$

$$\text{Constraints: } |y_i - w_i x_i| \leq \varepsilon \quad (4)$$

the ε can be tuned to achieve the desired accuracy for a model.

Next is the Long Short-Term Memory (LSTM). The Long Short-Term Memory (LSTM) network is a type of recurrent neural network (RNN) designed to model temporal sequences more effectively than traditional RNNs. Unlike regular RNNs, LSTMs are better equipped to handle long-range dependencies(Hochreiter & Schmidhuber, 1997; Schmidt, 2019). LSTMs are similar to RNNs in that they have a chain-like structure, but each repeating block, or LSTM cell, includes three more fully connected layers than a conventional RNN does. Since they have a different activation (sigmoid) than the regular layer, these extra layers are also known as gate layers (tanh). The cell state $C(t)$can be modified by adding or deleting information via the gate layers as it travels through each LSTM cell. This is how the LSTM model chooses whether to keep or discard data from earlier time steps. A few studies had success with its applications to container throughput series. For instance, (Lee et al., 2021) developed a prediction model for the Busan port's container throughput series, incorporating external variables and outperforming conventional LSTM models. Yang & Chang (2020) proposed a CNN-LSTM neural architecture for forecasting container demand at Taiwan's major ports, while(Shankar et al., 2020) applied LSTM networks to forecast container throughput at the port of Singapore, demonstrating superior performance compared to single conventional models.

Then moving on to the multivariate grey model (1, N). The Gray Model (GM) was first introduced by (Ju-Long,

1982) and has been applied to a wide variety of fields for a few reasons namely, accuracy and effectiveness in small sample modeling, less computational work, and its ability to fit any forecasting system. The Grey Model (GM) can be classified both into univariate and multivariate models, depending on the forecasting system. The GM (1, 1) model and the GM (1, *N*) are distinguished as univariate and multivariate prediction models, respectively. For the purposes of this study, the latter will be best suited for the forecasting system in question (MISO system). Although there have been recent optimizations to the multivariate GM (1, *N*) model (Wang & Qian, 2022; Ye et al., 2022; Zeng, 2019; Zhang et al., 2022), there is barely any study of the multivariate grey prediction model on container throughput series. Refer to (Lao et al., 2021) for a detailed equation and walkthrough of the basic multivariate grey prediction model, GM (1, *N*). A few studies, however, applied the GM (1, 1) to univariate throughput forecasting. Weng (2021) implemented a hybrid forecasting approach combining ARIMA, GM (1, 1), and exponential smoothing for Guangdong Province's throughput forecasting. Similarly, Zhizhen et al. (2016) utilized GM (1, 1) and exponential smoothing in a combined model, focusing on minimum variance for throughput forecasts. He & Wang (2021) developed a throughput model for the Tianjin-Hebei Port, integrating a fractional GM (1, 1) with a back-propagation neural network. These studies demonstrate varied methodologies in forecasting throughput in different geographical contexts. These applications of the univariate GM (1, 1) had good accuracy mostly due to the model's ability to perform accurately on small sample data. Given the inherent uncertainties and the influence of numerous factors in container throughput modeling, as noted by He & Wang (2021), enhancing forecasting accuracy can extend beyond merely combining or refining individual models. It also involves a thorough mapping of uncertainties by incorporating key factors that both directly and indirectly impact throughput. This comprehensive approach allows for a more accurate and reliable forecast in the field of container throughput.

Then, the least squares support vector regression (LSSVR)**.** LSSVR is considered the least-square version of the above-mentioned support vector machines (SVM). It was introduced by (Brabanter et al., 2011) as a reformulation of the SVRs. This version seeks a solution that follows from a linear Karush-Kuhn-Tucker (KKT) system by solving a set of linear equations instead of the more complex quadratic programming task involved in the traditional SVMs. Based on prior research, the LSSVR model is said to have better stability and train efficiently than the SVR (Hasanpour et al., 2010; Mustaffa et al., 2014; Pai et al., 2014; Wang et al., 2011; Wang et al., 2012; Wang & Wang, 2012; Wei et al., 2010). The LSSVR model can be obtained through the following optimization problem: see equations (5), and (6)

$$\text{Minimize: } \frac{1}{2}\|w\|^2 + \gamma \frac{1}{2} \sum_{i=1}^{n} \mathrm{e}_\mathrm{i}^2 \tag{5}$$

$$\text{Subject to: } y_i = w^k \phi(x_i) + b + e_i \tag{6}$$

where $\phi(x_i)$ is the mapping of the high dimensional space as in the traditional SVR, γ is a constant, and e_i are error variables, e_i^2 represents the sum of the squares of the errors (e_i) for all data points i in the dataset. w, is the weight vector to be learned, b, is the bias term. (Yuan & Lee, 2015). The objective of the LS-SVR is to minimize the squared error between the predicted output $w^k \phi(x_i) + b$ and the actual yi. The regularization term $\gamma \frac{1}{2} \sum_{i=1}^{n} \mathrm{e}_\mathrm{i}^2$ helps to prevent overfitting and control the complexity of the model. The Lagrange formulation for the LS-SVR can be formulated as shown in equation (7).

$$\begin{aligned} L(a, b, \varepsilon) = \left(\frac{1}{2}\right) \times \sum [ai^2] + \left(\frac{1}{2c}\right) \times \sum [\epsilon i^2] \\ - \sum [ai \times (y_i - f(xi) - b \\ - \epsilon i)] \end{aligned} \tag{7}$$

Subject to the following constraints: $ai \geq 0$: lagrange multipliers ai are non-negative, $\epsilon i \geq 0$: Epsilon variables ϵi are non-negative. See equation 8.

$$\sum [ai \times (yi - f(xi) - b - \epsilon i)] = 0 \tag{8}$$

: necessary conditions for the solution.

In the above formula, ai are Lagrange multipliers for the training data (xi, yi), C, is the cost parameter that controls the trade-off between maximizing the margin and minimizing the training error. b is the bias term and εi are the epsilon variables representing the deviation of the actual yi from the predicted $f(xi)$. It should be noted that the LSSVR seeks to reduce the squared error directly whereas the traditional SVR minimizes a ε-insensitive loss with a hinge-like loss function. These

discrepancies in loss functions result in unique Lagrange formulas for LSSVR and traditional SVR. Several studies have explored the application of SVR/LSSVR models for container throughput forecasting. For instance, (Mak et al., 2007) compared traditional SVM with an approximate LSSVM, finding that the LSSVM offered faster training and efficient memory usage. (Xie et al., 2013) achieved improved forecasting performance using hybrid approaches based on LSSVR for Shanghai and Shenzhen Ports.(Ding et al., 2019) proposed a hybrid approach combining backpropagation neural networks and SVM for port throughput forecast, which outperformed single models in a MISO system.

A multi-layer perceptron (MLP) is another artificial neural network containing several layers. Linear problems can be well handled in a single perceptron; thus, the MLP was developed to tackle this limitation. It is a neural network that maps linear and non-linear relationships between predictors and predicted variables and belongs to a class of neural networks known as feed-forward neural networks. The structure of an MLP is characterized by an input layer, hidden layers, output layers, weights, bias, and activation functions. Backpropagation is a way to train the MLP model's performance. The weights are fine-tuned by sending errors back into the network from the output layer in a backward direction, which is how the word "backpropagation" comes from about. The MLP can be formulated as in equation 9.

$$T_i = \sum_{i=1}^{n} w_{ij} x_i + d_i \quad (9)$$

where $x_{i\ldots}x_n$ is the input(s), w_{ij} are random weights, d_i is the bias of each node, i is the counter (1 to n). each layer gets a representation as $\theta = f(T_i)$ where f is the activation function. Accordingly, MLP falls under a class of deep neural networks alongside convolutional neural networks (CNN) and recurrent neural networks (RNN). While the latter two are primarily applied to image classification, face authentication, and object detection (Li et al., 2018; Tian, 2020; Tilk & Alumae, 2014). In port throughput forecasting, the use of multi-layer perceptron (MLP) models, a type of artificial neural network, is common. For instance, (Gosasang et al.) (2010) utilized an MLP model to forecast port throughput at the port of Bangkok, favoring it over a multiple linear regression model in a multiple-input and single-output system (MISO). Similarly, (Tang et al., 2019) employed a backpropagation neural network (BPNN) to model container throughput at Lianyungang and Shanghai ports, showing its applicability across ports with different economic situations.

The Random Forest Model is a supervised machine learning system that uses an ensemble technique to solve regression and classification issues. By averaging the output of weak decision trees, an accurate forecast may be obtained. RF is an accuracy-seeking model that relies on two rules: individual tree components must be more accurate than randomness, and their mistakes on fresh datasets must be mutually independent or different. Random forest primarily integrates a large number of decorrelated trees on a set of observations generated using a sample selected from the original data feature, and the model output is the average of forecasts. (Breiman, 2001) discussed the benefits of bootstrap aggregation, which includes reducing variance by averaging unbiased individual trees in the forest.

eliminating correlation through randomization and limiting overfitting. Random forests are formulated as see in equation 10.

$$f(x) = \frac{1}{K}\sum_{k=1}^{K} T_L(O_k)(x) \quad (10)$$

whereby k = 1 to K: (size of the ensemble). K, number of trees in the forest, T_L, grown tree, L, sample data (Auret & Aldrich, 2012).

Due to its strengths in accurate predictions, the random forest model has received widespread recognition and application in diverse fields in the research arena. (Pierdzioch & Risse, 2020) made a comparative study between multivariate and univariate forecasts in a forecasting scheme on precious metal prices. Pierdzioch and Risee's results proved the accuracy of multivariate forecasts over univariate forecasts.

The multivariate adaptive regression splines (MARS) model is an ensemble of linear functions combined with one or more hinge functions. It is more suitable to deal with high-dimensional input or non-linearity and demonstrates exceptional variable selection abilities. The MARS model uses the divide-and-conquer strategy, which finds optimal solutions by generating several regression equations for multiple segments of training data (Chan et al., 2019). Geng et al. suggested a container throughput forecasting scheme that combines SVR, CSAPSO (chaotic simulated annealing particle swarm

optimization), and the MARS model to get accurate results. They used the MARS model to choose which original input variables to use for the final input vectors.

Table 2. Summary of the key concepts of the models under study

Model	Key concept/ Theorem
Multiple Linear Regression (MLR)	The multiple linear regression model operates under the mathematical assumption of a linear association between the independent and dependent variables. Additionally, it presupposes minimal correlation among the independent variables, as their influence is measured solely on the dependent variable, not on each other.
Support vector regression (SVR),	The objective of Support Vector Regression (SVR) is to identify a function that estimates the connection between input variables and a continuous target variable, aiming to reduce the prediction error to a minimum.
Long Short-Term Memory (LSTM).	LSTM networks are designed to model temporal sequences more effectively than traditional RNNs by incorporating gated mechanisms. LSTMs are equipped with specialized units called LSTM cells, which contain memory cells and gate layers (e.g., sigmoid and tanh) that regulate the flow of information. These components enable LSTMs to capture long-range dependencies in sequential data, making them well-suited for time-series forecasting tasks.
multivariate grey model (1, N).	The key theorem of the multivariate grey model (1, N) is the principle of grey prediction, which forms the foundation of the model's forecasting methodology. This principle is based on the idea of reducing uncertainty in forecasting by using limited information and knowledge, particularly when dealing with small sample sizes or incomplete data.
least squares support vector regression (LSSVR).	The LSSVR model aims to minimize the squared error between predicted and actual observations. By minimizing squared errors, the model aims for predictions closely aligned with true values. The Lagrange formulation balances this objective with constraints, incorporating Lagrange multipliers and error variables to ensure adherence to model requirements and objectives.
multi-layer perceptron (MLP)	The key concept involves the backpropagation algorithm for training neural networks, which involves propagating the error backward through the network, adjusting the weights and biases at each layer to minimize the error between the predicted and actual outputs. This iterative process continues until the model's performance reaches a satisfactory level.
Random forests (RF)	The concept of ensemble learning and the benefits of bootstrap aggregation. This concept emphasizes the importance of combining multiple weak decision trees to create a strong predictive model.
Multivariate adaptive regression splines (MARS)	Employs a divide-and-Conquer Strategy to find optimal solutions by generating several regression equations for multiple segments of training data. This strategy allows the MARS model to effectively handle high-dimensional input or non-linearity and demonstrates exceptional variable selection abilities.

Forecasting is crucial for decision-making in the shipping industry as it provides insights into future trends. De Gooijer & Hyndman, (2006) highlighted the underutilization of multivariate models due to limited empirical research. Recent advancements in computing have made time series methods popular for container throughput forecasts, but they often overlook key factors. Munim et al., (2023) stressed the need to consider various variables in port development schemes. Also, the advent of the COVID-19 pandemic and its impact on the shipping industry, along with other external variables, macroeconomic conditions, trade regulations, and geopolitical concerns that affect container throughput, it begs the question as to what multivariate forecasting scheme is suitable to account for the key factors affecting throughput across diverse ports. In this study, we explore various multivariate forecasting models for container throughput at the same port. We employ two preprocessing methods: MULTI LAGGED VAL (basic) and RF-VIM (preprocessed). To optimize data preparation and enhance model accuracy. Additionally, we assess the outputs generated from both preprocessing techniques.

3. Methodology

The study focuses on the port of Shanghai in China. Selected for its significance in global maritime trade, Shanghai port is the busiest port in the world in terms of container TEU (twenty-foot-equivalent units) volume, with a remarkable 49 million TEU recorded in 2023, as reported by the World Shipping Council. Its strategic location along major waterways like the Huangpu and Yangtze Rivers, coupled with its extensive infrastructure and connectivity to global shipping routes, makes it a key entry and exit point for throughput moving in and out of Asia. Below is a depiction of a time series of container throughput at Shanghai port from 1985-2020.

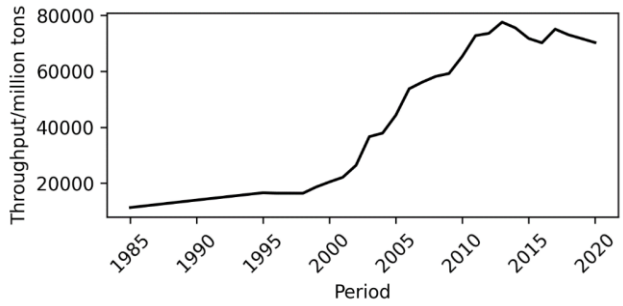

Figure 2: Time series graph of port throughput for the port of Shanghai (million tons) Source: Author(s)

A comparison approach is employed in the research, with multiple forecasting models being assessed based on six assessment criteria (R-squared, NRMSE, ME, MAE, MPE, and MAPE). Previous studies (Gökkuş et al., 2017; Jugović et al., 2011; Langen et al., 2012; Ping & Fei, 2013) have demonstrated a strong link between port throughput and macroeconomic factors, emphasizing the interconnectedness of cargo movement with population, trade, and global economic conditions. In line with these theoretical foundations, we have chosen 14 macroeconomic variables as inputs for our selected port. Table 3. Shows the variables adopted in the study.

Table 3. Container throughput variables for the port of Shanghai

Input	Description	Input	Description
X_1	Gross National Product (GNP)	X_8	Secondary Industry Value (SIV)
X_2	Gross Domestic Product (GDP)	X_9	Tertiary Industry Value (TIV)
X_3	Per Capita Gross Domestic Product (PCGDP)	X_{10}	Population (PP)
X_4	Total Fixed Asset Investments	X_{11}	Total Retail Sales of Consumer goods (TRSCG)
X_5	Imports and Exports (IE)	X_{12}	Freight capacity (FC)
X_6	Industrial Output (IO)	X_{13}	Road Freight Capacity (RFC1)
X_7	Primary Industrial Value (PIV)	X_{14}	Railway Freight Capacity (RFC2)

Source: Author(s)

The datasets are collected from the Shanghai yearbook, covering a period from 1985 – 2021. The selection of these variables and data points was further constrained by the data unavailability. With a total of 36 data points for the study, the original datasets are split into 80:20 datasets. 80% (28 data points) are used for training the models, and 20% (8 data points) are retained as the testing set for gauging the performance of the proposed forecasting schemes.

3.1. Data Preprocessing Methods

Data preprocessing plays a crucial role in enhancing forecast accuracy. By tailoring the data to fit the forecasting model, we ensure its responsiveness to analysis, making it more meaningful and informative. Critical stages in data preparation include data cleansing, data transformation, and feature selection (Chakrabarty et al., 2016). Data cleaning and transformation aim to prepare data for modeling by removing anomalies and standardizing the dataset. Feature selection, widely used in research (Auret & Aldrich, 2012; Awah et al., 2021; Mo et al., 2018), is crucial in this process. In this study, we normalized the original datasets using the Min-Max method due to variations in measurement units, ensuring consistent scales across factors. Normalization enhances forecasting accuracy by minimizing the impact of predictor variables with large values on those with smaller ones.

$$x = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{11}$$

where x is the original value, x_{min} is the minimum value in the column, and x_{max} is the maximum value. We will conduct comparisons using a preprocessing strategy often employed in machine learning forecasting models, as there is a scarcity of papers (if any) that take it into account. There are two approaches under consideration in this study: The first approach is a MULTI LAGGED-VAL (no special preprocessing) the input variables to the models are the lagged values (say

$x_1, x_2, x_3, \ldots, x_{14}.$) and a target output y as $f(x)$ - function of x. The second approach a RF-VIM (random forest variable importance measure). Researchers use the RF-VIM to select important features for prediction tasks. It assesses predictor variable significance by permuting values and observing the impact on model performance. A significant decline indicates a strong predictor-response correlation. Variable importance measure by permutation, $w_j(T_L)$ could be calculated as shown in the steps below were mse is the mean squared error of the model, and $L^j_{OOB^{(0)}}$ is the OOB learning sample with variable X_j permuted:

a. Train the RF Model: The RF model, denoted as $L(0)$, is trained on the original dataset X (excluding X_j) and response vector y. This model consists of K decision trees.
b. Creating a permuted Dataset: For a particular variable X_j, a permuted dataset is created by randomly shuffling the values in the j^{th} column of X while keeping all other variables the same. This new dataset is denoted as $L^{(j)}_{OOB}$, where OOB is the out-of-bag sample.
c. Making predictions: Each of the K trees in the random forest model makes predictions on both the original dataset $L(0)$ and the permuted dataset $L^{(j)}_{OOB}$. The predictions are denoted as $T_{L(0)}(X)$ and $L_{L^{(j)}_{OOB}}(X)$, respectively.
d. Calculating Mean Squared Error (MSE): The MSE is calculated for the predictions made by each tree on both datasets. For tree L, the MSE on the original data is $mse\left(T_{L(0)}(X)\right)$, and the MSE on the permuted data is $mse(T_{L^{(j)}_{OOB}}(X))$.
e. Calculating per-tree VIM: The per-tree variable importance for variable X_j is the difference in MSE due to permuting X_j, calculated for each tree L as:

$$w_j(T_L) = mse\left(T_{L(0)}(X)\right) - mse\left(T_{L^j_{OOB^{(0)}}}(X)\right) \quad (12)$$

Then expanded to an ensemble of trees by averaging the importance measures of individual trees:

$$\delta_j = \frac{1}{K}\sum_{k=1}^{K} w_j(T_{L_K}) \quad (13)$$

T_{L_K} represents the K^{th} tree in the forest. $w_j(T_{L_K})$ is the importance score for variable K^{th} tree, K is the total number of trees in the forest, and δ_j is the average importance score for variable X_j across all trees. The benefit of permuted variable significance measures, such as w_j, is in its consideration of multivariate interactions with other input factors. This is achieved by permuting the variables, which not only eliminates the relationship with the target variable but also disrupts any associations with other input variables. After computing δ_j for all variables, a threshold is applied to determine which variables are significantly important in predicting the response. Variables with a δ_j above this threshold would be considered important. This procedure gives us a quantitative measure of how much the prediction error increases when the association between the variable X_j and the response y is broken. Important variables in the model show a greater increase in prediction error when permuted, resulting in a higher importance score (Nicodemus & Malley, 2009). We employ models mentioned in the literature, using preprocessing methods described earlier. Firstly, models are used in their basic form without alterations (MULTI-LAGGED VAL), and secondly with pre-selected features based on RF-VIM. This approach helps define variable importance thresholds to exclude truly unimportant variables. training many additional forests, where the response variable for each forest is permuted. Below are the results of the RF-VIM score in Fig. 3.

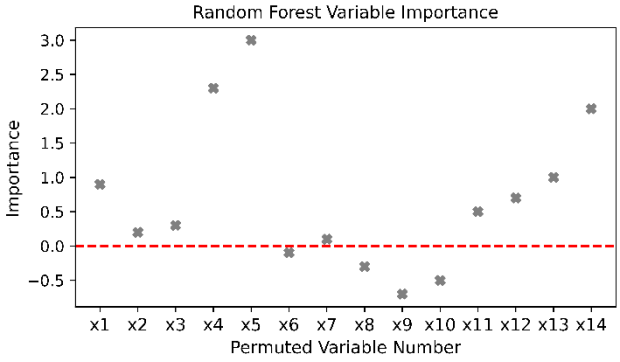

Figure 3: Random Forest variable importance (RF-VIM) score by permutation. Source: Author(s)

Based on the Random Forest Variable Importance Measure (RF-VIM), most of our predictor variables scored above the red dotted threshold line. Variables that fell below this line will not be used in subsequent models because they are considered unimportant. We will compare the performance of models using the RF-VIM with those using multi-lagged values.

The Support Vector Regression (SVR) employs a grid search to find optimal hyperparameters, using the 'radial' kernel function. The function model of the Least Squares Support Vector Regression (LSSVR) is solved using the Lagrange function. For the Multilayer Perceptron (MLP) model, there are twelve input neurons, eight neurons in one hidden layer, and one output neuron. Parameters for the MLP were chosen based on performance evaluation on a validation sample from the original dataset. The hidden layer uses the ReLu activation function, and the learning rate is set to 0.0001. The Long Short-Term Memory (LSTM) model is sequential, with twelve input nodes, six LSTM units in the hidden layer, and one output node. A linear activation function is applied before all unit outputs, followed by a hard sigmoid function for the recurrent step. All model predictions and quantitative analysis in this study were conducted using Python's Scikit-Learn library, version 3.11.2.

Various evaluation measures are commonly used in forecasting schemes to provide a comprehensive assessment. In this study, we considered six criteria to evaluate model performance based on prediction errors, including R-squared, normalized root mean squared error (NRMSE), mean error (ME), mean absolute error (MAE), mean percentage error (MPE), and mean absolute percentage error (MAPE) (Peng & Chu, 2009; Xie et al., 2013), These criteria are calculated using specific equations, as outlined below. R-squared measures the goodness of fit, while NRMSE allows for comparison among models of different scales. ME and MAE assess the average error, with MAE addressing concerns about negative errors. MPE and MAPE account for error percentages, helping to mitigate issues related to sample size. Model comparisons will be based on the results of these assessment criteria.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{14}$$

$$NRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|(y_i - x_i)|^2} \tag{15}$$

$$ME = \frac{1}{n}\sum_{i=1}^{n} y_i - x_i \tag{16}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - x_i)| \tag{17}$$

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - x_i}{y_i} \tag{18}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - x_i}{y_i}| \text{ X } 100 \tag{19}$$

4. Results and Discussion

Following the procedures outlined in the research methodology section, the forecasting models have been constructed. The results of these models are presented in the subsequent table 2. to aid in their performance evaluation, with the best performing model in bold.

Table 4. Forecasting performances of the benchmarked models.

MODELS	MULTI-LAGGED VAL					
	R squared	**ME**	**NRMSE**	**MAE**	**MPE (%)**	**MAPE (%)**
MLR	0.42	-4045.35	0.48	6045.35	7.47	8.47

SVR	0.68	3465.54	0.28	2465.54	3.91	3.01
LSSVR	0.71	1582.37	0.24	1382.60	2.87	2.94
LSTM	0.73	1300.61	0.15	1369.71	2.28	2.33
GM (1, N)	0.77	1298.03	0.13	1298.03	2.43	2.49
MLP	0.84	1204.18	0.06	1204.18	1.73	1.78
RF	0.89	1045.49	0.06	1045.18	1.09	1.09
MARS	**0.97**	**900.45**	**0.05**	**900.43**	**0.89**	**0.92**

Source: Author(s)

Table 5. Forecasting performances of the benchmarked models based on preprocessing.

MODELS	**RF-VIM Method**					
	R-squared	**ME**	**NRMSE**	**MAE**	**MPE (%)**	MAPE (%)
MLR	0.62	-3045.35	0.38	6045.35	3.47	8.47
SVR	0.74	2405.23	0.26	2265.44	1.21	2.01
LSTM	0.74	1462.33	0.34	132.60	1.16	2.84
GM(1, N)	0.70	1700.61	0.36	1609.51	1.28	1.54
LSSVR	0.82	1098.13	0.09	118.03	1.53	1.72
MLP	0.80	1254.08	0.15	1204.26	1.73	1.89
RF	0.95	1005.40	0.05	1045.14	1.01	1.03
MARS	**0.98**	**900.45**	**0.04**	**900.44**	**0.62**	**0.22**

Source: Author(s)

First are the results based on the MULTI-LAGGED VAL. The table displays the performance of various models in our study, ordered from best to worst. The MARS model performed the best, surpassing all others, while the MLR model performed the poorest. This result was expected because the MLR model assumes linearity and cannot capture complex patterns in data. In contrast, the MARS model can capture intricate patterns and relationships without assumptions, making it ideal for multivariate forecasting. Although the SVR model, which is resilient to outliers and captures non-linear patterns, performed slightly better than the MLR model, its performance was still modest. The multivariate GM(1, N) model showed

average accuracy compared to the other models. The Gray Model (1, N) performed decently, but there is potential for improvement by incorporating more variables. In a prior study by Chan et al. (2019), a univariate GM(1, 1) model performed poorly compared to other models in forecasting container throughput under a single-input, single-output scheme. This suggests that a multivariate GM(1, N) model, considering more variables, might yield better result

The RF-VIM technique has enhanced the performance of most models, leading to increased accuracy across various error measures. The MARS model remains the top performer, with its R-squared improving from 0.97 to 0.98. Additionally, the Random Forest model's performance has risen from 0.89 to 0.95. These findings indicate that random forest variable importance effectively identifies and eliminates less significant variables, improving model performance. All data presented in Figures 4, 5, and 6 are based on the results obtained using the RF-VIM method. Figure 4. illustrates the throughput forecasting results, along with error metrics such as R-squared, NRMSE, MPE (%), and

MAPE (%), providing insights into model performance.

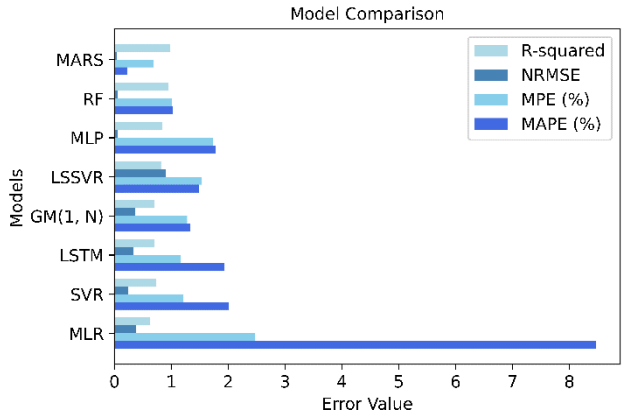

Figure 4. Graphical presentation of model performance (RF-VIM method). Source: Authors

The study found that the LSSVR model performed satisfactorily with MPE and MAPE values of 1.53 and 1.62, respectively. This is significant because the LSSVR model is computationally less complex than the SVR model. It suggests that the LSSVR model could be a viable alternative when computational efficiency is crucial. The highest-performing models in the study were MARS, RF, and multilayer MLP, with MAPE values of 0.22, 1.03, and 1.89, respectively. This could be attributed to the study's use of a multivariate-based forecasting scheme, which required modeling intricate relationships and non-linear data patterns. It's important to analyze each error measure separately due to value variations, including negative values. Figure 5 illustrates the graphical presentation of the models' ME and MAE values.

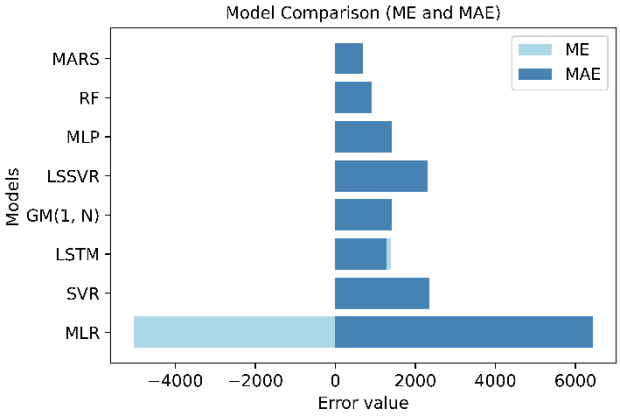

Figure 5: Model performance based on ME and MAE error metrics (RF-VIM method). Source: Author(s)

The MARS model, an enhanced version of the MLR model, demonstrates superior performance compared to other models in the study. It achieves high accuracy with an R2 of 0.98 and minimal errors, including a ME of 900.45 and NRME of 0.03. Analysis of the data presented in Table 3 and Figures 2 and 3 supports this conclusion. The MARS model excels at identifying complex patterns and correlations in the data without assuming causality between variables, making it an ideal benchmark for multivariate prediction. Geng et al. (2015) proposed a combination model using MARS and SVR for container throughput forecasting, further validating the effectiveness of the MARS model. Comparative analysis reveals that their MARS-based approach outperforms other models, as illustrated by mean absolute percentage errors in Figure 3. This solidifies the MARS model's status as a benchmark for throughput prediction. In summary, the MARS model, particularly in its basic form under the MISO forecasting scheme, surpasses other models and serves as an optimal benchmark for container throughput predictive modeling. We recommend its adoption as a benchmark for reliable and robust predictive modeling in similar scenarios.

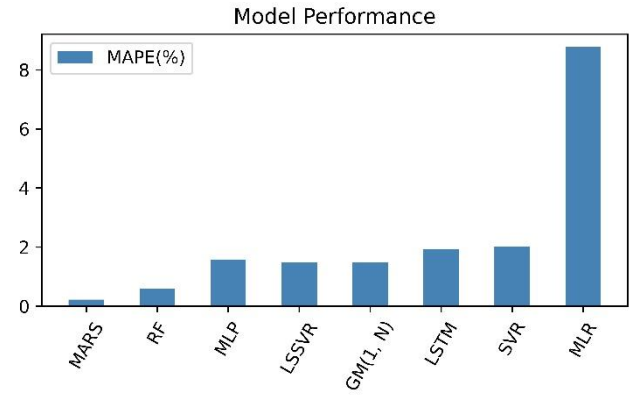

Figure 6: Performance ranking of all models (RF-VIM based method). source: Author(s)

5. Conclusions

Throughout the years, researchers have strived to find the most precise forecasting method to optimize costs and benefits. This study compares eight multivariate models across two scenarios: first, using a RF-VIM preprocessing technique, and then utilizing lagged values of the datasets without preprocessing. The results indicate an overall improvement in model accuracy when truly unimportant variables are excluded through the RF-VIM technique. However, three models—MLP, LSSVR, and LSTM—exhibited poorer performance under the RF-VIM technique compared to the dataset with no preprocessing. Interestingly, variables deemed truly unimportant for some models (SVR, RF, MARS, GM(1, N), MLR) actually contributed to the accuracy of others.

The Random Forest Variable Importance Method (RF-VIM) identifies the MARS model as a benchmark for MISO forecasting, offering opportunities for modifications to enhance accuracy across different scenarios or ports. Incorporating socio-economic factors improves forecast accuracy, but careful variable selection remains crucial. In future studies, researchers should focus on expanding the analysis to include more socio-economic factors, such as the relationship between ports and economic policies, hinterland connectivity, and operational dynamics. It is essential to develop more robust preprocessing techniques for variable selection and enhance the interpretability of machine learning models. Additionally, exploring alternative approaches, such as advanced optimization algorithms and novel hybrid methods, could lead to better parameter combinations and further improvement of the foundational model. Given recent changes in global trade patterns and technical developments in port operations, future study should consider the influence of digitalization and automation on container throughput predictions. These improvements are transforming logistics and supply chain management techniques, impacting port efficiency and capacity use. Integrating real-time data analytics and machine learning algorithms might result in more accurate and flexible forecasting models that can adjust to changing market conditions and operational challenges. Furthermore, investigating the use of sustainability measures and carbon footprint factors into forecasting algorithms might help port operations meet global environmental goals and regulatory frameworks.

References

Afshari Safavi, E. (2022). Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. *Tropical Animal Health and Production*, *54*(1), 55. https://doi.org/10.1007/s11250-022-03073-2

Aizerman, M. A. (1964). Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Automation and Remote Control*, *25*, 821-837.

An, C., Lim, H., Kim, D.-W., Chang, J. H., Choi, Y. J., & Kim, S. W. (2020). Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. *Scientific Reports*, *10*(1), 18716. https://doi.org/10.1038/s41598-020-75767-2

Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *MINERALS ENGINEERING*, *35*, 27-42. https://doi.org/10.1016/j.mineng.2012.05.008

Awah, P. C. N., H. Kim, S. (2021). Short Term Forecast of Container Throughput: New Variables Application for the Port of Douala. *Journal of Marine Science and Engineering*, *9*(7), Article 720. https://doi.org/10.3390/jmse9070720

Ben-Hur, A., Horn, D., Siegelmann, H. T., & Vapnik, V. (2002). Support vector clustering. *J. Mach. Learn. Res.*, *2*, 125–137.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers* Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, USA. https://doi.org/10.1145/130385.130401

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brabanter, K. D., Brabanter, J. D., Suykens, J. A. K., & Moor, B. D. (2011). Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression. *IEEE Transactions on Neural Networks*, *22*(1), 110-120. https://doi.org/10.1109/TNN.2010.2087769

Breiman, L. (2001). Random forests. *MACHINE LEARNING*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chakrabarty, A., Mannan, S., & Cagin, T. (2016). Chapter 8 - Inherently Safer Design. In A. Chakrabarty, S. Mannan, & T. Cagin (Eds.), *Multiscale Modeling for Process Safety Applications* (pp. 339-396). Butterworth-Heinemann. https://doi.org/https://doi.org/10.1016/B978-0-12-396975-0.00008-5

Chan, H. K., Xu, S. J., & Qi, X. G. (2019). A comparison of time series methods for forecasting container throughput. *International Journal of Logistics-Research and Applications*, *22*(3), 294-303. https://doi.org/10.1080/13675567.2018.1525342

Chen, Z. Z., Chen, Y., & Li, T. Y. (2016). Port Cargo Throughput Forecasting Based On Combination Model. *Proceedings of the 2016 Joint International Information Technology, Mechanical and Electronic Engineering*, *59*, 148-154.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*(3), 443-473. https://doi.org/https://doi.org/10.1016/j.ijforecast.2006.01.001

Ding, M. J., Zhang, S. Z., Zhong, H. D., Wu, Y. H., & Zhang, L. B. (2019). A Prediction Model of the Sum of Container Based on Combined BP Neural Network and SVM. *Journal of Information Processing Systems*, *15*(2), 305-319.

https://doi.org/10.3745/jips.04.0107

Du, S. D., Li, T. R., Yang, Y., & Horng, S. J. (2020). Multivariate time series forecasting via attention-based encoder-decoder framework. *Neurocomputing*, *388*, 269-279. https://doi.org/10.1016/j.neucom.2019.12.118

Eberly, L. E. (2007). Multiple Linear Regression. In W. T. Ambrosius (Ed.), *Topics in Biostatistics* (pp. 165-187). Humana Press. https://doi.org/10.1007/978-1-59745-530-5_9

Geng, J. L., M. W. Dong, Z. H. Liao, Y. S. (2015). Port throughput forecasting by MARS-RSVR with chaotic simulated annealing particle swarm optimization algorithm. *Neurocomputing*, *147*, 239-250. https://doi.org/10.1016/j.neucom.2014.06.070

Gosasang, V., Chandraprakaikul, W., Kiattisin, S., & Iaeng. (2010). An Application of Neural Networks for Forecasting Container Throughput at Bangkok Port. *World Congress on Engineering, Wce 2010, Vol I*, 137-141.

Gu, Y., Chen, Y., Wang, X., & Chen, Z. (2023). Impact of COVID-19 epidemic on port operations: Evidence from Asian ports. *Case Studies on Transport Policy*, *12*, 101014. https://doi.org/https://doi.org/10.1016/j.cstp.2023.101014

Gökkuş, Ü., Yıldırım, M. S., & Aydin, M. M. (2017). Estimation of Container Traffic at Seaports by Using Several Soft Computing Methods: A Case of Turkish Seaports. *Discrete Dynamics in Nature and Society*, *2017*, 2984853. https://doi.org/10.1155/2017/2984853

Hasanpour, F., Ensafi, A. A., & Khayamian, T. (2010). Simultaneous chemiluminescence determination of amoxicillin and clavulanic acid using least squares support vector regression. *Analytica chimica acta*, *670*(1-2), 44-50.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, A., Liu, X., Rao, C., Zhang, Y., & He, Y. (2022). A New Container Throughput Forecasting Paradigm under COVID-19. *Sustainability*, *14*(5).

Huang, J., Chu, C.-W., & Hsu, H.-L. (2021). A comparative study of univariate models for container throughput forecasting of major ports in Asia. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, *236*(1), 160-173. https://doi.org/10.1177/14750902211023662

Ju-Long, D. (1982). Control problems of grey systems. *Systems & Control Letters*, *1*(5), 288-294. https://doi.org/https://doi.org/10.1016/S0167-6911(82)80025-X

Jugović, A., Hess, S., & Jugovic, T. P. (2011). Traffic Demand Forecasting for Port Services. *Promet-traffic & Transportation*, *23*, 59-69.

Langen, D., Meijeren, v. J., & Tavasszy, L. A. (2012). Combining Models and Commodity Chain Research for Making Long-Term Projections of Port Throughput: an Application to the HamburgLe Havre Range. *European Journal of Transport and Infrastructure Research*.

Lao, T. F., Chen, X. T., & Zhu, J. N. (2021). The Optimized Multivariate Grey Prediction Model Based on Dynamic Background Value and Its Application. *COMPLEXITY*, *2021*. https://doi.org/10.1155/2021/6663773

Lee, E., Kim, D., & Bae, H. (2021). Container Volume Prediction Using Time-Series Decomposition with a Long Short-Term Memory Models. *Applied Sciences-Basel*, *11*(19), Article 8995. https://doi.org/10.3390/app11198995

Li, X., & Xu, S. (2011). A Study on Port Container Throughput Prediction Based on Optimal Combined Forecasting Model in Shanghai Port. In *ICCTP 2011* (pp. 3894-3905). https://doi.org/doi:10.1061/41186(421)390

Mak, K. L., Yang, D. H., & Int Assoc, E. (2007). Forecasting Hong Kong's container throughput with approximate least squares support vector machines. *World Congress on Engineering 2007, Vols 1 and 2*, 7-+.

Mishra, S., Bordin, C., Taharaguchi, K., & Palu, I. (2020). Comparison of deep learning models for multivariate prediction of time series wind power generation and temperature. *Energy Reports*, *6*, 273-286. https://doi.org/10.1016/j.egyr.2019.11.009

Mo, L. L., Xie, L., Jiang, X. Y., Teng, G., Xu, L. X., & Xiao, J. (2018). GMDH-based hybrid model for container throughput forecasting: Selective combination forecasting in nonlinear subseries. *APPLIED SOFT COMPUTING*, *62*, 478-490. https://doi.org/10.1016/j.asoc.2017.10.033

Munim, Z. H., Fiskin, C. S., Nepal, B., & Chowdhury, M. M. H. (2023). Forecasting container throughput of major Asian ports using the Prophet and hybrid time series models. *The Asian Journal of Shipping and Logistics*, *39*(2), 67-77. https://doi.org/https://doi.org/10.1016/j.ajsl.2023.02.004

Mustaffa, Z., Yusof, Y., & Kamaruddin, S. S. (2014). Enhanced artificial bee colony for training least squares support vector machines in commodity price forecasting. *Journal of Computational Science*, *5*(2), 196-205.

Muñoz-Organero, M., & Queipo-Álvarez, P. (2022). Deep Spatiotemporal Model for COVID-19 Forecasting. *Sensors*, *22*(9).

Notteboom, T. (2016). The adaptive capacity of container ports in an era of mega vessels: The case of upstream seaports Antwerp and Hamburg. *Journal of Transport Geography*, *54*, 295-309. https://doi.org/10.1016/j.jtrangeo.2016.06.002

Pai, P.-F., Hung, K.-C., & Lin, K.-P. (2014). Tourism demand forecasting using novel hybrid system. *Expert Systems with applications*, *41*(8), 3691-3702.

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.

Peter, S. C., Dhanjal, J. K., Malik, V., Radhakrishnan, N., Jayakanthan, M., & Sundar, D. (2019). Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 661-676). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-809633-8.20197-0

Pierdzioch, C., & Risse, M. (2020). Forecasting precious metal returns with multivariate random forests. *EMPIRICAL ECONOMICS*, *58*(3), 1167-1184. https://doi.org/10.1007/s00181-018-1558-9

Ping, F. F., & Fei, F. X. (2013). Multivariant Forecasting Mode of Guangdong Province Port throughput with Genetic Algorithms and Back Propagation Neural Network. *Procedia - Social and Behavioral Sciences*, *96*, 1165-1174. https://doi.org/https://doi.org/10.1016/j.sbspro.2013.08.133

Puntanen, S. (2013). Handbook of Regression Analysis by Samprit Chatterjee, Jeffrey S. Simonoff. *International Statistical Review*, *81*(2), 330-331. https://doi.org/https://doi.org/10.1111/insr.12020_22

Schmidt, R. M. (2019). Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. *ArXiv*, *abs/1912.05911*.

Shankar, S., Ilavarasan, P. V., Punia, S., & Singh, S. P. (2020). Forecasting container throughput with long short-term memory networks. *Industrial Management & Data Systems*, *120*(3), 425-441. https://doi.org/10.1108/imds-07-2019-0370

Shankar, S., Punia, S., & Ilavarasan, P. V. (2021). Deep learning-based container throughput forecasting: a triple bottom line approach. *Industrial Management & Data Systems*, *121*(10), 2100-2117. https://doi.org/10.1108/IMDS-12-2020-0704

Tang, S., Xu, S. D., & Gao, J. W. (2019). An Optimal Model based on Multifactors for Container Throughput Forecasting. *Ksce Journal of Civil Engineering*, *23*(9), 4124-4131. https://doi.org/10.1007/s12205-019-2446-3

Tian, Y. H. (2020). Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm. *IEEE ACCESS*, *8*, 125731-125744. https://doi.org/10.1109/ACCESS.2020.3006097

Tilk, O., & Alumae, T. (2014). *Multi-Domain Recurrent Neural Network Language Model for Medical Speech Recognition* HUMAN LANGUAGE TECHNOLOGIES - THE BALTIC PERSPECTIVE, BALTIC HLT 2014,

Tok, V., & Ece, N. J. (2022). THE IMPACT OF COVID-19 ON MARITIME TRADE AND TRANSPORTATION: AN ESTIMATION OF THE MARITIME TRADE POST-COVID-19. *Mersin University Journal of Maritime Faculty*, *4*(2), 18-30. https://doi.org/10.47512/meujmaf.1200009

Wang, H., Li, E., & Li, G. (2011). Probability-based least square support vector regression metamodeling technique for crashworthiness optimization problems. *Computational Mechanics*, *47*(3), 251-263.

Wang, H., Li, W., & Li, G. (2012). A robust inverse method based on least square support vector regression for johnson-cook material parameters. *Computers Materials and Continua*, *28*(2), 121.

Wang, J., & Qian, W. Y. (2022). An improved discrete grey multivariable model for forecasting the R&D output of China-from the perspective of R&D institutions. *KYBERNETES*, *51*(4), 1365-1387. https://doi.org/10.1108/K-11-2020-0749

Wang, S., & Wang, Q. (2012). Prediction and dispatching of workshop material demand based on least squares support vector regression with genetic algorithm. *International Information Institute (Tokyo). Information*, *15*(1), 213.

Wei, D., Chen, F., & Zhang, T. (2010). Least square-support vector regression-based car-following model with sparse sample selection. 2010 8th World Congress on Intelligent Control and Automation,

Xie, G., Wang, S. Y., Zhao, Y. X., & Lai, K. K. (2013). Hybrid approaches based on LSSVR model for container throughput forecasting: A comparative study. *Applied Soft Computing*, *13*(5), 2232-2241. https://doi.org/10.1016/j.asoc.2013.02.002

Xu, S., Zou, S., Huang, J., Yang, W., & Zeng, F. (2022). Comparison of Different Approaches of Machine Learning Methods with Conventional Approaches on Container Throughput Forecasting. *Applied Sciences*, *12*(19).

Yang, C. H. C., P. Y. (2020). Forecasting the Demand for Container Throughput Using a Mixed-Precision Neural Architecture Based on CNN-LSTM. *Mathematics*, *8*(10), Article 1784. https://doi.org/10.3390/math8101784

Ye, L., Yang, D. L., Dang, Y. G., & Wang, J. J. (2022). An enhanced multivariable dynamic time-delay discrete grey forecasting model for predicting China's carbon emissions. *ENERGY*, *249*, Article 123681. https://doi.org/10.1016/j.energy.2022.123681

Yuan, F.-C., & Lee, C.-H. (2015). Using least square support vector regression with genetic algorithm to forecast beta systematic risk. *Journal of Computational Science*, *11*, 26-33. https://doi.org/https://doi.org/10.1016/j.jocs.2015.08.004

Zeng, L. (2019). Analysing the high-tech industry with a multivariable grey forecasting model based on fractional order accumulation. *KYBERNETES*, *48*(6), 1158-1174. https://doi.org/10.1108/K-02-2018-0078

Zhang, M., Guo, H., Sun, M., Liu, S. F., & Forrest, J. (2022). A novel flexible grey multivariable model and its application in forecasting energy consumption in China. *ENERGY*, *239*, Article 122441. https://doi.org/10.1016/j.energy.2021.122441

Zou, Y., Su, B., & Chen, Y. (2022). Nonparametric Functional Data Analysis for Forecasting Container Throughput: The Case of Shanghai Port [Article]. *Journal of Marine Science and Engineering*, *10*(11), Article 1712. https://doi.org/10.3390/jmse10111712

Meersman, H., Van de Voorde, E. and Vanelslander, T. (2005), Ports as hubs in the logistics chain, In: H. Leggate, J. McConville and A. Morvillo (eds.), *International Maritime Transport: Perspectives*, Routledge: London, pp. 32-45.

Received **29 May 2024**

1st Revised **25 June 2024**

Accepted **25 June 2024**