



Systematic framework for ARIMA and SARIMA model selection using multiple error metrics: An application to export forecasting

Nattavadee Mahanil and Premporn Khemavuk*

Department of Industrial Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand

Received 11 April 2026
Revised 28 September 2026
Accepted 2 October 2026

Abstract

This study proposes a systematic and practical framework for selecting autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) forecasting models based on multiple error evaluation metrics, including root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). In practical forecasting applications, different error metrics often produce conflicting model selection outcomes, resulting in uncertainty in decision-making. Accordingly, this study introduces a percentage-based consistency evaluation approach to identify the most reliable forecasting model across multiple datasets. This study proposes a systematic framework for ARIMA and SARIMA model selection that integrates multiple forecasting error metrics to address conflicting model rankings and improve the consistency of model-selection decisions. This framework is applied to Thailand's export value data, consisting of 69 monthly observations across 12 datasets derived from major industrial product groups (HS-84, HS-85, and HS-87) and key trading partners, including China, Japan, and South Korea. Results indicate that ARIMA(1,2,1) achieves the highest level of consistency across error metrics, with 83.33% under RMSE and 75% under MAPE and MAE. In terms of seasonal models, SARIMA (2,2,2)(2,0,0)₁₂ demonstrates the best overall performance. The findings confirm that the choice of error metrics can significantly influence forecasting outcomes. The proposed consistency-based framework reduces model selection conflicts and enhances the reliability of decision-making. This approach provides practical guidance for forecasting applications in economic planning, industrial systems, and export management.

Keywords: ARIMA, SARIMA, Forecast error metrics, Model selection, Time series forecasting, Export forecasting

1. Introduction

Forecasting is a systematic process of acquiring and analyzing relevant information to support planning and decision-making [1] for anticipating future conditions. This process involves predicting changes driven by technological, economic, and social factors, as well as other variables that may influence plan implementation. Forecasting serves as a critical tool for generating forward-looking information to support business planning, including projections of aggregate demand and market size [2]. Furthermore, forecasting is a critical component of effective management, enabling organizations to make informed decisions based on rigorously analyzed data. This process helps reduce uncertainty, mitigate potential risks and adverse decision outcomes [3], and enhance systematic coordination within organizations by improving the accuracy and reliability of future projections.

Machine learning has emerged as an important approach for time-series forecasting across a wide range of applications [4]. However, traditional statistical methods remain widely used due to their interpretability and effectiveness. The autoregressive integrated moving average (ARIMA) model is one of the most widely utilized time-series forecasting techniques. ARIMA is recognized for its strong predictive performance and its ability to efficiently model temporal data using only historical observations [5]. This feature makes this model a practical and cost-effective solution, particularly in situations where external variables are limited or unavailable. Furthermore, when time-series data exhibit seasonal patterns, the seasonal autoregressive integrated moving average (SARIMA) model extends ARIMA to account for seasonality. SARIMA improves forecasting performance for data with periodic fluctuations by incorporating seasonal components [6]. Accordingly, ARIMA and SARIMA are considered robust and widely adopted tools for forecasting future trends. The application of these models enhances the accuracy and reliability of predictive information, thereby supporting business planning and decision-making and improving organizational performance.

Previous studies have primarily focused on applying ARIMA and SARIMA models, such as Sankalp and Panda [7], and comparing them with machine learning models with ARIMA for time-series forecasting, following the Box-Jenkins methodology, which consists of key steps, such as stationarity testing, differencing, parameter identification using autocorrelation function (ACF) and partial autocorrelation function (PACF), model estimation, and validation [8]. This systematic framework ensures that the model is appropriately specified and capable of producing reliable forecasts. Chyon et al. [9] applied the ARIMA model using Python and Jupyter Notebook for forecasting real-world data. Eiglsperger et al. [10] introduced ForeTiS, an open-source Python framework that extends the use of computational tools across the entire time-series forecasting workflow, enabling efficient model training,

*Corresponding author.

Email address: premporn.k@eng.kmutnb.ac.th
<https://doi.org/10.64960/easr.2026.267189>

comparison, and analysis through a simplified implementation. Ariyanti and Yusnitasari [11] aimed to develop crude oil price forecasting models using ARIMA and SARIMA. The model performance was evaluated using RMSE, where ARIMA(0,1,0) and SARIMA(0,1,0)(0,0,0) achieved the same RMSE value of 0.01905%, indicating high forecasting accuracy, with similar short-term prediction results. Consistent with this approach, Ning et al. [12] compared ARIMA, long short-term memory (LSTM), and Prophet models for oil production forecasting. The model performance was evaluated using RMSE and MAE, where ARIMA achieved an RMSE of 4.97 and MAE of 4.37, with an improved accuracy in the first prediction period (RMSE = 3.63, MAE = 2.87). Suppalakpanya et al. [13] applied exponential time-series methods to forecast oil palm prices, crude palm oil prices, and crude palm oil production in Thailand. Their findings demonstrated the usefulness of time-series forecasting for supporting planning and decision-making. Venkataraman et al. [14] developed time-series models using the Akaike Information Criterion (AIC), emphasizing the importance of systematic model development and model selection for improving forecasting performance.

The forecasting process involves selecting a model that best fits the data characteristics. In this context, accuracy is a key criterion for model selection. Forecast accuracy is commonly evaluated using forecast error, defined as the difference between predicted values and corresponding actual observations. A smaller discrepancy indicates higher accuracy, whereas a larger discrepancy reflects lower accuracy and greater error. Therefore, forecast error provides a fundamental basis for evaluating and comparing the performance of different forecasting models [15].

Previous studies indicate that forecast error, as a measure of accuracy, is often evaluated using multiple metrics rather than a single method. This approach enables a more comprehensive assessment of model performance, as different error measures capture distinct aspects of forecasting accuracy. Consequently, the results from these metrics are compared to identify the most appropriate and reliable model for practical implementation. Demir and Akkaş [16] conducted a comparative study of two forecasting models in a Turkish animal feed company. In their analysis, three commonly used error metrics, namely, mean absolute error (MAE), mean absolute percentage error (MAPE), and mean squared error (MSE), were utilized to evaluate model performance. The results indicate that different error measures produced inconsistent conclusions regarding model accuracy, highlighting the importance of using multiple evaluation criteria in forecasting studies. Zhang and Na [17] investigated agricultural commodity price forecasting using two error metrics: mean squared error (MSE) and the coefficient of determination (R^2). The results indicated that different measures resulted in inconsistent conclusions regarding forecasting accuracy. However, no detailed explanation was provided for these discrepancies. Athipayoon et al. [18] evaluated the performance of the EEMD-ARIMA model across various datasets using MSE and MAPE. The findings indicated that the model accuracy varies depending on the error metric utilized and the characteristics of the dataset.

Recent studies have further emphasized the importance of systematic forecasting model selection. Studies have proposed systematic approaches for forecasting model selection. Poler and Mula [19] used out-of-sample rolling-horizon errors to support model selection. Badulescu et al. [20] applied an MCDM approach to address conflicting forecasting error measures. Koubaa et al. [21] combined ANP and TOPSIS to select forecasting models based on multiple evaluation criteria. Stroilă [22] proposed a data-driven approach for selecting forecasting models according to time-series characteristics.

Recently, forecasting model selection has increasingly been formulated as a multi-criteria decision-making (MCDM) problem because different evaluation metrics often produce conflicting rankings of candidate models. Various MCDM approaches, including the weighted sum method, analytic hierarchy process, technique for order preference by similarity to ideal solution (TOPSIS), VIKOR, and PROMETHEE, have been widely applied to integrate multiple performance criteria into a single decision score. Although these approaches provide systematic mechanisms for model comparison, they generally require predefined criterion weights or pairwise comparisons, which may introduce additional subjectivity into the model selection process.

Recent forecasting studies have also expanded beyond conventional statistical models by incorporating machine learning and deep learning techniques, including LSTM, gated recurrent units, Transformer-based forecasting models, and hybrid forecasting frameworks. Despite the superior predictive capability of these models in many applications, model selection remains highly dependent on the choice of evaluation metrics, and inconsistent conclusions frequently arise when multiple error measures are simultaneously considered.

The proposed framework differs from existing approaches by introducing a consistency-based model selection strategy that integrates multiple forecasting error metrics without requiring subjective weighting schemes. Instead of optimizing a single criterion, the framework identifies forecasting models that consistently efficiently perform across RMSE, MAE, and MAPE, thereby improving the robustness and reliability of model selection. These approaches demonstrate that selection of a forecasting model solely on the basis of a single performance measure may not adequately represent its overall performance. Table 1 summarizes representative studies related to forecasting model evaluation and selection and highlights their relationship to the research focus of this study.

Table 1 Summary of related studies on time-series forecasting and model selection

Authors	Forecasting/selection approach	Evaluation criteria	Main findings	Relevance/gap
Ariyanti and Yusnitasari [11]	ARIMA and SARIMA	RMSE	Compared ARIMA and SARIMA for crude-oil forecasting	Primarily relied on a single error metric
Ning et al. [12]	ARIMA, LSTM, and Prophet	RMSE and MAE	Performance differed among forecasting approaches	Multiple metrics were reported without explicitly resolving metric-based selection conflicts
Demir and Akkas [16]	Comparative forecasting models	MAE, MAPE, and MSE	Different error metrics produced inconsistent conclusions	Demonstrated metric-dependent selection but did not provide a systematic conflict-resolution framework
Zhang and Na [17]	Agricultural price forecasting	MSE, (R^2)	Evaluation measures produced inconsistent assessments	Metric inconsistency was identified but not systematically resolved

Table 1 (continued) Summary of related studies on time-series forecasting and model selection

Authors	Forecasting/selection approach	Evaluation criteria	Main findings	Relevance/gap
Athipayoon et al. [18]	EEMD–ARIMA	MSE and MAPE	Performance varied according to metric and dataset	No integrated model-selection mechanism across multiple criteria
Poler and Mula [19]	Rolling-horizon model selection	Out-of-sample weighted errors	Proposed automatic forecasting-model selection using out-of-sample performance	Emphasized dynamic model selection rather than consistency across multiple error metrics (ScienceDirect)
Badulescu et al. [20]	Two-stage MCDM	Multiple error measures	Integrated several interdependent error measures for forecasting-model selection	Requires an explicit multi-criteria decision procedure (ScienceDirect)
Koubaa et al. [21]	ANP–TOPSIS	MAPE, MAE, RMSE, and other criteria	MCDM resolved conflicting forecasting-performance criteria and sensitivity analysis assessed ranking robustness	Relies on criterion weighting/expert judgments; illustrates an alternative approach to multi-metric selection (MDPI)
Stroilă [22]	Feature-based/tree-based model selection	RMSE and sMAPE	Model performance and ranking varied according to data characteristics and evaluation metric	Demonstrates continuing need for systematic, data-dependent model-selection strategies (Springer)
This study	Consistency-based ARIMA/SARIMA model-selection framework	RMSE, MAPE, and MAE	Evaluates consistency of model-selection decisions across multiple datasets and metrics	Provides an empirical multi-metric consistency framework without relying exclusively on a single error criterion

Table 1 demonstrates that previous studies have addressed forecasting model selection using individual error metrics, multiple performance measures, out-of-sample evaluation, MCDM techniques, and data-driven selection strategies. These studies demonstrate that the preferred forecasting model can depend on the evaluation criterion and the characteristics of the dataset. However, comparatively limited attention has been given to evaluating how consistently the same forecasting model is selected across multiple datasets when several commonly used error metrics are simultaneously considered. This study addresses this issue by developing a consistency-based framework that integrates RMSE, MAPE, and MAE for the systematic evaluation of ARIMA and SARIMA model-selection decisions. Rather than claiming universal superiority over established MCDM approaches, the proposed framework provides a practical alternative for examining metric-dependent selection consistency within the empirical setting considered in this study.

This study is a continuation of the structure of Mahanil and Khemavuk [23] that has defined a conceptual model for developing a selection error measurement method to measure the effectiveness of the time-series forecasting method. The proposed framework provides practical guidelines for decision-makers, thereby enhancing the reliability and robustness of model selection in real-world applications.

The research questions are as follows: (1) How can multiple error metrics be effectively combined to select a forecasting model?; and (2) Is there a possibility of using a single error metric method to evaluate the ARIMA or SARIMA model's performance?

2. Materials and methods

2.1 Data phases

2.1.1 Data collection and extraction

This study utilizes demand forecasting data from the export sector. Specifically, export value data from Thailand were collected and analyzed to develop ARIMA and SARIMA models. The dataset was further categorized into the three product groups with the highest export values, based on the Harmonized System (HS Code) [24]: HS-84 (Machinery and Mechanical Appliances), HS-85 (Electrical Machinery), and HS-87 (Vehicles) [25].

Furthermore, export data from major Asian trading partners with the highest export volumes, specifically China, Japan, and South Korea, were incorporated to validate the performance and effectiveness of the proposed models.

2.1.2 Data preparing and preprocessing

Export value data from Thailand were collected from January 2018 to September 2023, covering a total of 69 months [26]. The dataset was processed and organized using Excel to ensure a structured tabular format. This process facilitated data review, accuracy verification, and consistency of measurement units, with all data obtained from reliable sources.

The study period includes the COVID-19 pandemic, during which international trade and supply chains experienced substantial disruptions. The observations during this period were retained in the time series because this study evaluates forecasting model selection

using the historical series as observed. No intervention variables or exogenous indicators were introduced to explicitly separate crisis-driven effects from a business-as-usual variation. Consequently, the forecasting errors reported in this study reflect model performance over a period containing regular market conditions and extraordinary disruptions.

2.2 Modeling phases

2.2.1 Forecasting models

The ARIMA model, originally developed by Box et al. [27], used a statistical approach for time-series analysis and forecasting. This model is characterized by three key parameters: an autoregressive (AR) component, which represents dependence on past observations; a moving average (MA) component, which captures the influence of past error terms; and an integrated (I) component, which refers to the differencing process applied to achieve stationarity [3]. Thus, ARIMA combines AR and MA components with differencing to form a comprehensive framework for time-series modeling. This model is commonly expressed as ARIMA(p, d, q), where p , d , and q denote the orders of the AR, differencing, and MA components, respectively. The model is provided in Equation (1).

$$\phi(B)(1-B)^d Y_t = c + \theta(B)\varepsilon_t, \quad (1)$$

where ε_t is a white-noise process with zero mean and variance σ^2 ; B is the backshift operator; and $\phi(B)$ and $\theta(B)$ are the autoregressive and moving-average polynomials of orders p and q , respectively. In terms of stationarity and invertibility, the roots of the corresponding AR and MA characteristic polynomials are assumed to lie outside the unit circle [28].

The ARIMA model is commonly used for forecasting time-series data with non-seasonal patterns. However, an extension, known as the SARIMA model, has been developed to address datasets with inherent seasonal variations. SARIMA captures non-seasonal and seasonal characteristics, thereby improving forecasting accuracy for time series with periodic behavior. The SARIMA model is defined by seven parameters and is commonly expressed as SARIMA(p, d, q)(P, D, Q) s . Variables p , d , and q represent the non-seasonal orders, P , D , and Q indicate the seasonal orders, and s denotes the seasonal period. The model is provided in Equation (2).

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = c + \theta(B)\Theta(B^s)\varepsilon_t, \quad (2)$$

where $\phi(B)$ and $\theta(B)$ are the non-seasonal autoregressive and moving-average polynomials of orders p and q , respectively; and $\Phi(B^s)$ and $\Theta(B^s)$ are the seasonal autoregressive and moving-average polynomials of orders P and Q , respectively. Parameter s denotes the seasonal period [29].

The SARIMA model can effectively capture short-term and long-term patterns in time-series data by incorporating seasonal and non-seasonal components (Figure 1) [29].

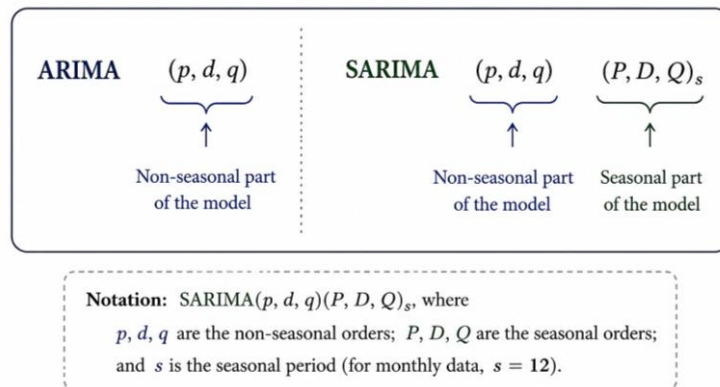


Figure 1 Notation of ARIMA and SARIMA models.

where m is the number of observations per year; the uppercase notation is used for the seasonal parts of the model; and the lowercase notation is utilized for the non-seasonal parts of the model [29].

2.2.2 Modeling

Both models were applied to the prepared dataset, which was processed using the Python programming language with open-source libraries. The experimental procedure for model implementation and evaluation is described as follows.

2.2.2.1 Data importation

Data processing and model implementation were conducted using open-source libraries in Python. The *Pandas* library was used for data ingestion, handling missing values, filtering, and grouping [30]. The *NumPy* library was used for numerical operations and efficient mathematical computations [31]. Meanwhile, the *Statsmodels* library was used to develop ARIMA and SARIMA models for time-series analysis and forecasting [32].

2.2.2.2 Data splitting

The dataset was divided into training and testing sets using a chronological split. The first 59 months (85%) were used for model training, while the remaining 10 months (15%) were utilized for testing [7]. This approach preserves the temporal structure of the data and prevents data leakage [33].

2.2.2.3 Augmented dickey–fuller test

The candidate orders of the ARIMA models were determined following the Box–Jenkins identification procedure rather than by arbitrary manual selection to assess stationarity. The Augmented Dickey–Fuller [34] test was conducted, with a significance level of 0.05. If the p-value was less than 0.05, then the series was considered stationary; otherwise, it was deemed non-stationary, and differencing was applied. The differencing order (d) was initially set to zero and increased as necessary until stationarity was achieved.

Thereafter, the non-seasonal autoregressive and moving-average orders, p and q, were identified from the dominant patterns of the PACF and ACF, respectively. These diagnostic patterns demonstrate that low-order candidate values within the range 0–2 were retained for model estimation. Moreover, this restricted search space was consistent with the principle of model parsimony, avoiding unnecessarily high-order specifications when the observed dependence structure can be represented by lower-order terms.

2.2.2.4 Partial autocorrelation and autocorrelation

The ACF and PACF were examined after the stationarity assessment to identify appropriate candidate orders for the ARIMA and SARIMA models. A slowly decaying ACF may indicate non-stationarity and the need for differencing [35]. In the stationary series, the PACF and ACF patterns were used to guide the selection of the non-seasonal autoregressive and moving-average orders (p and q), respectively. In the SARIMA models, correlations at seasonal lags were additionally examined to guide the selection of the seasonal orders (P and Q), with a seasonal period of s=12 for the monthly data [36]. An example of the ACF and PACF analysis is presented in Figure 2.

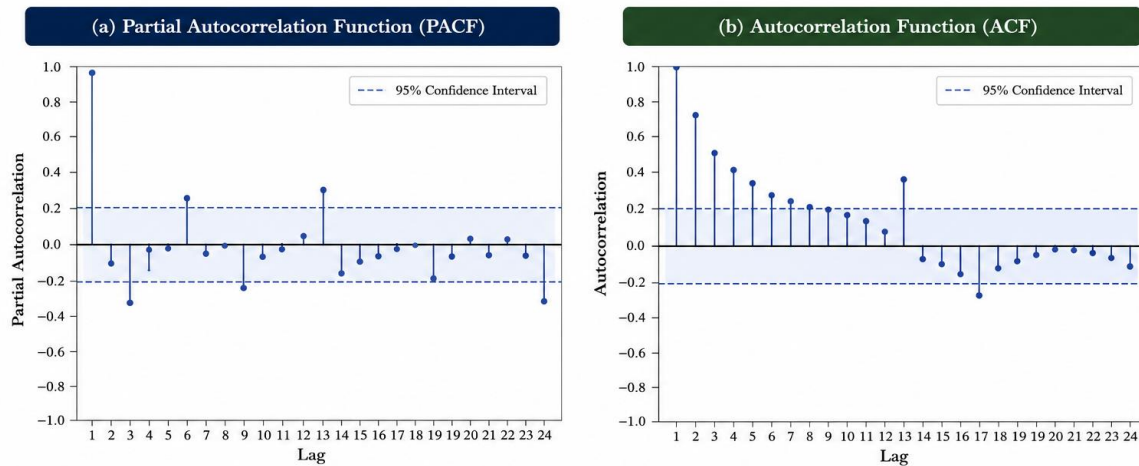


Figure 2 ACF and PACF plots used to guide the identification of candidate ARIMA and SARIMA model orders.

In terms of the seasonal specification, the seasonal period was set to s=12 because the observations are recorded monthly, and the modeling objective was to capture annual seasonal dependence [37]. The seasonal structure was examined using the ACF, with particular attention to correlations at seasonal lags (e.g., lag 12 and its multiples), before seasonal AR and MA candidate terms were specified. Thus, s=12 represents the annual cycle of the monthly series rather than an arbitrarily imposed numerical value.

Accordingly, the parameter ranges were used as a constrained candidate search space informed by the statistical characteristics of the observed series, rather than as manually selected fixed orders.

2.2.2.5 Fit the models

The candidate ARIMA(p,d,q) and SARIMA(p,d,q)(P,D,Q)_s models were fitted to the training data using the parameter combinations determined in the preceding model-identification step. In the monthly series, the seasonal period was specified as s=12. Thereafter, each fitted model was used to generate forecasts over the fixed testing horizon. The resulting forecasts were evaluated using the selected error metrics. The model-fitting and forecasting procedure is illustrated in Figure 3.

2.2.2.6 Forecasting performance test

Forecasts are inherently subject to uncertainty, and the difference between forecasted and actual values is referred to as forecast error [38]. Three error metrics are utilized to evaluate and compare the performance of ARIMA and SARIMA models: MAE, MAPE, and Root Mean Squared Error (RMSE). The formulations of MAE, MAPE, and RMSE are presented in Equations (3)–(5), respectively [39–40].

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|, \quad (3)$$

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right|, \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}, \quad (5)$$

where t denotes the time index, n is the total number of observations, y_t is the observed value at time t , and $\hat{y}_t$ is the corresponding forecast value.

MAE is a widely used metric for evaluating forecasting accuracy, where low MAE values indicate improved model performance. MAE is straightforward to interpret as it is expressed in the same units as the original data [39]. Meanwhile, MAPE is commonly used to measure overall forecast error and provides a scale-independent metric, enabling comparison across different datasets [40]. RMSE is particularly effective for assessing the influence of large errors or outliers, as it assigns great weight to large deviations due to the squaring process. Consequently, RMSE values are typically higher than MAE unless all errors are zero [39].

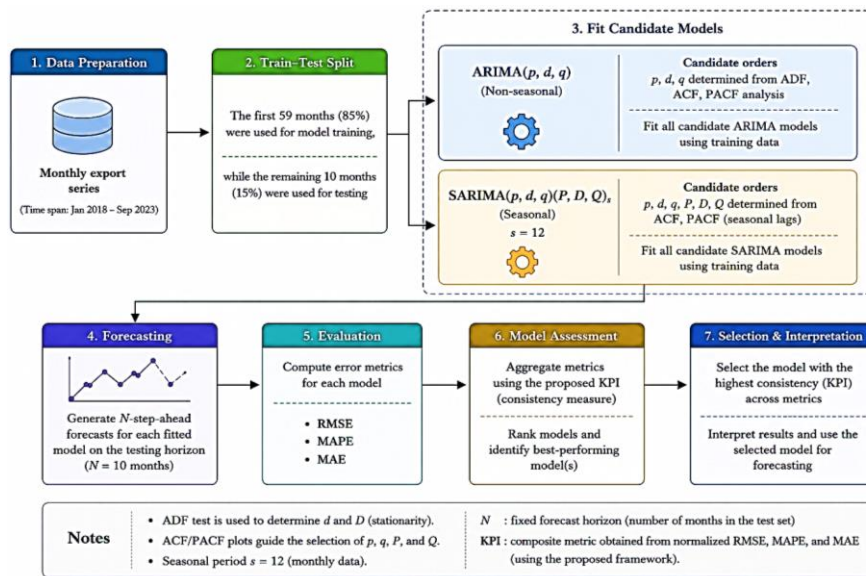


Figure 3 Model-fitting and forecasting procedure for the candidate ARIMA and SARIMA models.

2.2.2.7 Results of ARIMA and SARIMA models

The modeling results based on Thailand's export value dataset are presented in two parts as follows:

1. Performance of ARIMA Models

A total of 12 models were evaluated using Thailand's export value dataset. ARIMA(1,2,1) demonstrated the best overall forecasting performance based on three error metrics, namely, MAE, MAPE, and RMSE. However, when considering the MAPE metric alone, ARIMA(0,2,1) also exhibited the best performance (Table 2).

Table 2 Performance of the ARIMA models in Thailand's export value

ARIMA Model (p,d,q)	Forecasting Error Performance			Residual Diagnostic Results for the ARIMA Candidate Models				Overall
				Ljung-Box		ARCH		
	RMSE	MAPE	MAE	p-value	LB	p-value	ARCH	
(1,2,1)	3942.086	0.063	3090.052	<0.001	Fail	0.244	Pass	Fail
(1,1,0)	3954.283	0.064	3165.413	<0.001	Fail	0.773	Pass	Fail
(0,1,1)	3971.153	0.064	3171.741	<0.001	Fail	0.793	Pass	Fail
(0,2,1)	3971.999	0.063	3116.234	<0.001	Fail	0.322	Pass	Fail
(0,1,2)	3990.796	0.065	3202.989	<0.001	Fail	0.755	Pass	Fail
(2,1,0)	4049.792	0.066	3248.921	<0.001	Fail	0.753	Pass	Fail
(1,1,1)	4133.581	0.067	3321.542	<0.001	Fail	0.771	Pass	Fail
(0,1,0)	4375.003	0.072	3547.539	<0.001	Fail	0.326	Pass	Fail
(2,2,1)	4474.518	0.073	3632.788	<0.001	Fail	0.287	Pass	Fail
(2,1,1)	4549.172	0.075	3724.256	<0.001	Fail	0.77	Pass	Fail
(2,1,2)	4551.079	0.075	3729.401	<0.001	Fail	0.804	Pass	Fail
(1,1,2)	4564.995	0.075	3729.794	<0.001	Fail	0.795	Pass	Fail

2. Performance of SARIMA models

A total of 18 models were evaluated using Thailand's export value dataset. SARIMA(2,2,2)(1,0,1)₁₂ demonstrated the best overall performance, based on MAE and RMSE. By contrast, when considering MAPE alone, SARIMA (2,2,2)(2,0,0)₁₂ SARIMA (2,2,2)(2,0,1)₁₂ and SARIMA(2,2,2)(1,0,0)₁₂ yielded the lowest error values. The selected model results are presented in Table 3.

Table 3 Performance of the SARIMA models in Thailand's export value

SARIMA Model (p,d,q,P,D,Q) _s	Forecasting Error Performance			Residual Diagnostic Results for the SARIMA Candidate Models				Overall
	RMSE	MAPE	MAE	Ljung–Box		ARCH		
				p-value	LB	p-value	ARCH	
(2,2,2)(1,0,1) ₁₂	1758.980	0.057	1397.160	0.029	Fail	0.342	Pass	Fail
(2,2,2)(1,0,2) ₁₂	1772.530	0.057	1395.400	0.004	Fail	0.455	Pass	Fail
(2,2,2)(2,0,0) ₁₂	1796.230	0.056	1397.640	0.008	Fail	0.423	Pass	Fail
(2,1,2)(2,0,2) ₁₂	1812.650	0.061	1496.950	0.002	Fail	0.559	Pass	Fail
(2,1,1)(2,0,2) ₁₂	1814.160	0.062	1501.800	<0.001	Fail	0.908	Pass	Fail
(2,2,2)(2,0,1) ₁₂	1815.530	0.056	1401.230	0.007	Fail	0.45	Pass	Fail
(2,2,2)(0,0,1) ₁₂	1826.610	0.059	1452.050	0.03	Fail	0.258	Pass	Fail
(2,2,2)(0,0,2) ₁₂	1828.900	0.057	1418.450	0.008	Fail	0.377	Pass	Fail
(2,2,2)(1,0,0) ₁₂	1836.010	0.056	1403.910	0.006	Fail	0.400	Pass	Fail
(1,1,2)(1,2,2) ₁₂	2122.730	0.079	1810.580	<0.001	Fail	0.984	Pass	Fail
(2,2,2)(0,0,0) ₁₂	2141.230	0.068	1625.010	0.166	Pass	0.201	Pass	Pass
(0,1,2)(1,0,1) ₁₂	2186.120	0.066	1650.370	0.014	Fail	0.604	Pass	Fail
(0,1,2)(2,0,1) ₁₂	2187.150	0.066	1650.930	0.005	Fail	0.625	Pass	Fail
(0,1,2)(1,0,2) ₁₂	2188.340	0.066	1651.740	<0.001	Fail	0.628	Pass	Fail
(1,1,1)(1,0,1) ₁₂	2200.290	0.067	1664.510	0.117	Pass	0.668	Pass	Pass
(2,1,2)(2,2,0) ₁₂	2416.800	0.095	2159.340	<0.001	Fail	0.914	Pass	Fail
(0,1,1)(1,0,0) ₁₂	2417.860	0.076	1890.940	0.147	Pass	0.740	Pass	Pass
(2,1,0)(1,0,0) ₁₂	2419.490	0.076	1894.440	0.172	Pass	0.520	Pass	Pass

Tables 2–3 present the performance of the ARIMA and SARIMA models based on Thailand's export values. The data show that some models achieve consistent accuracy across all three-error metrics. However, notable discrepancies in model performance are observed when different metrics are applied. These findings indicate that the use of multiple error metrics can result in variations in the evaluation of forecasting accuracy. Such inconsistencies may influence model selection and affect their reliability and effectiveness in practical forecasting applications.

Furthermore, residual diagnostics were performed to assess the adequacy of the fitted models. The Ljung–Box test at lag 12, with adjustment for model degrees of freedom, was used to examine the remaining residual autocorrelation. Meanwhile, the ARCH test was applied to assess conditional heteroskedasticity. Statistical significance was evaluated at $\alpha=0.05$, with $p>0.05$ indicating no significant evidence of the corresponding residual problem.

2.3 Systematic framework for selecting ARIMA and SARIMA forecasting models based on multiple error metrics

The performance results of the models based on MAE, MAPE, and RMSE exhibited variations across different error metrics. Each metric produced distinct error values, reflecting different levels of accuracy in the evaluation of ARIMA and SARIMA models. Consequently, this study proposes a method for selecting ARIMA and SARIMA models by jointly considering all three-error metrics (MAE, MAPE, and RMSE) to identify the most suitable model for accurate and reliable forecasting. The proposed method consists of the following steps:

2.3.1 Input the model

ARIMA and SARIMA models were evaluated under conditions where inconsistencies were observed among the three-error metrics. Models with a MAPE below 10% were also considered, as MAPE provides a scale-independent measure of forecasting accuracy [9]. Accordingly, this study proposes a model selection approach that integrates MAE, MAPE, and RMSE to identify the most appropriate model for reliable forecasting.

2.3.2 Define Key Performance Indicators (KPIs)

The results from each model were used to derive performance indicators based on Step 2.3.1. Since the error metrics MAE, MAPE, and RMSE are expressed in different units, RMSE and MAE are scale-dependent measures, whereas MAPE is scale-independent. Consequently, the raw values of these three metrics were not directly multiplied in the proposed framework. Prior to aggregation, each error metric was independently transformed into a dimensionless score using the normalization procedure in Eq. (6) [41]. This transformation maps the three criteria onto a common interval of [1,2] [42], thereby removing the original measurement units of these metrics and preventing differences in numerical scale from directly dominating the composite score. Consequently, the multiplicative aggregation in Eq. (7) is performed on dimensionless normalized scores rather than on the original RMSE, MAPE, and MAE values.

$$x_{scaled} = 1 + \frac{X_{max} - X_i}{X_{max} - X_{min}} (X_{max} - X_{min}), \quad (6)$$

where: x_{scaled} denotes the normalized value; x_i represents the original observation; and x_{min} and x_{max} denote the minimum and maximum values of the dataset, respectively.

The Min–Max transformation was selected because it preserves the relative ordering of observations within each criterion and provides an interpretable common scale for subsequent aggregation. The normalized interval was shifted from [0,1] to [1,2] to avoid zero-valued components in the multiplicative aggregation. Nevertheless, Min–Max normalization depends on the observed minimum and maximum values and may be sensitive to extreme observations. Accordingly, the present formulation should be interpreted as a practical normalization strategy for the datasets examined in this study rather than as a universally robust transformation against outliers or rank reversal. The shift to [1,2] changes the numerical scale of the resulting KPI while ensuring that all normalized error metrics remain positive and contribute to the multiplicative aggregation.

After normalization, the error values from MAE, MAPE, and RMSE were aggregated using a multiplicative method [43] to compute a composite error score. The proposed KPI adopts a multiplicative aggregation strategy rather than an additive formulation because forecasting models should exhibit consistently good performance across all evaluation metrics. In additive aggregation methods, poor performance in one criterion may be compensated by excellent performance in another, potentially leading to the selection of models with unbalanced forecasting accuracy. By contrast, multiplicative aggregation penalizes models that poorly perform on any individual metric, given that a relatively low normalized score substantially decreases the overall KPI value. Consequently, the proposed KPI favors forecasting models that simultaneously maintain low RMSE, MAE, and MAPE values. This approach integrates information from all three metrics (Equation (7)).

$$KPI = x_{scaled_MAE} \times x_{scaled_MAPE} \times x_{scaled_RMSE}, \quad (7)$$

where: x_{scaled_MAE} , x_{scaled_MAPE} , x_{scaled_RMSE} .

The error values from each measurement method, after normalization to a common scale [44], are presented. This characteristic is consistent with the objective of forecasting model selection, where robustness and consistency across multiple evaluation criteria are typically preferred over optimization with respect to a single error metric. Similar multiplicative aggregation concepts have been widely adopted in MCDM approaches because they effectively minimize compensatory effects among evaluation criteria and provide a balanced assessment of overall model performance.

2.3.3 Classify the models

The normalized KPI values obtained from Section 2.3.2 were clustered using the Elbow Method [45] to determine the optimal number of clusters (k) [46] from the Within-Cluster Sum of Squares (WCSS) curve [47, 48]. were grouped using the K-means clustering algorithm implemented in the Scikit-learn library in Python. Euclidean distance was utilized as the similarity measure to partition models into homogeneous groups according to their overall forecasting performance. The k-means++ initialization strategy was adopted with $n_init=10$, $random_state=42$, and a maximum of 300 iterations to improve clustering stability and reproducibility.

K-means was used to objectively categorize models according to the composite KPI. The composite KPI integrates the normalized RMSE, MAPE, and MAE into a single evaluation score, thereby preserving the combined contribution of the three forecasting error metrics. Subsequently, K-means clustering was applied to these composite scores to provide an objective, data-driven classification of forecasting models with similar overall performance, rather than relying on manually selected threshold values.

2.3.4 Identify the optimal model

The best forecasting model was identified based on the clustering results from Step 2.3.3, demonstrating that it belongs to a group containing other models. KPI values, derived from the three-error metrics—MAE, MAPE, and RMSE—were used to evaluate model performance. This approach provides a highly flexible framework for forecasting and allows KPI values to compare performance across different model parameter sets. The clustering results for the ARIMA and SARIMA models are presented in Tables 4 and 5, respectively.

Table 4 Classification and identification of the optimal ARIMA models in Thailand's export value

ARIMA Model (p,d,q)	Forecasting Error Performance			Forecasting Error Performance (x_{scale})			KPI	Clusters (K=2)
	RMSE	MAPE	MAE	RMSE	MAPE	MAE		
(1,2,1)	3942.086	0.063	3090.052	2.000	2.000	2.000	8.000	A
(1,1,0)	3954.283	0.064	3165.413	1.980	1.916	1.882	7.144	A
(0,1,1)	3971.153	0.064	3171.741	1.953	1.916	1.872	7.009	A
(0,2,1)	3971.999	0.063	3116.234	1.952	2.000	1.959	7.648	A
(0,1,2)	3990.796	0.065	3202.989	1.921	1.833	1.823	6.424	A
(2,1,0)	4049.792	0.066	3248.921	1.827	1.750	1.751	5.600	A
(1,1,1)	4133.581	0.067	3321.542	1.692	1.666	1.638	4.621	A
(0,1,0)	4375.003	0.072	3547.539	1.305	1.250	1.284	2.096	B
(2,2,1)	4474.518	0.073	3632.788	1.145	1.166	1.151	1.538	B
(2,1,1)	4549.172	0.075	3724.256	1.025	1.000	1.008	1.034	B
(2,1,2)	4551.079	0.075	3729.401	1.022	1.000	1.000	1.023	B
(1,1,2)	4564.995	0.075	3729.794	1.000	1.000	1.000	1.000	B

Table 4 indicates that ARIMA(1,2,1) attains the highest KPI value of 8.00, identifying it as the best-performing model within Cluster A. Additionally, several other models—ARIMA(1,1,0), ARIMA(0,1,1), ARIMA(0,2,1), ARIMA(0,1,2), ARIMA(2,1,0), and ARIMA(1,1,1)—are grouped within the same cluster.

Table 5 indicates that SARIMA(2,2,2)(1,0,1)₁₂ attains the highest KPI value of 7.888, identifying it as the best-performing model within Cluster A. Additionally, several other models, SARIMA(2,2,2)(1,0,2)₁₂, SARIMA(2,2,2)(2,0,0)₁₂, SARIMA(2,1,2)(2,0,2)₁₂, SARIMA(2,1,1)(2,0,2)₁₂, SARIMA(2,2,2)(2,0,1)₁₂, SARIMA(2,2,2)(0,0,1)₁₂, SARIMA(2,2,2)(0,0,2)₁₂, and SARIMA(2,2,2)(1,0,0)₁₂ are grouped within the same cluster.

Table 5 Classification and identification of the optimal SARIMA models in Thailand's export value

SARIMA Model (p,d,q)(P,D,Q) ₁₂	Forecasting Error Performance			Forecasting Error Performance (x_{scale})			KPI	Clusters (K=3)
	RMSE	MAPE	MAE	RMSE	MAPE	MAE		
(2,2,2)(1,0,1) ₁₂	1758.980	0.057	1397.160	2.000	1.974	1.998	7.888	A
(2,2,2)(1,0,2) ₁₂	1772.530	0.057	1395.400	1.979	1.974	2.000	7.816	A
(2,2,2)(2,0,0) ₁₂	1796.230	0.056	1397.640	1.944	2.000	1.997	7.763	A
(2,1,2)(2,0,2) ₁₂	1812.650	0.061	1496.950	1.919	1.872	1.867	6.706	A
(2,1,1)(2,0,2) ₁₂	1814.160	0.062	1501.800	1.916	1.846	1.861	6.583	A
(2,2,2)(2,0,1) ₁₂	1815.530	0.056	1401.230	1.914	2.000	1.992	7.628	A
(2,2,2)(0,0,1) ₁₂	1826.610	0.059	1452.050	1.898	1.923	1.926	7.028	A
(2,2,2)(0,0,2) ₁₂	1828.900	0.057	1418.450	1.894	1.974	1.970	7.367	A
(2,2,2)(1,0,0) ₁₂	1836.010	0.056	1403.910	1.883	2.000	1.989	7.492	A
(1,1,2)(1,2,2) ₁₂	2122.730	0.079	1810.580	1.449	1.410	1.457	2.977	B
(2,2,2)(0,0,0) ₁₂	2141.230	0.068	1625.010	1.421	1.692	1.699	4.088	B
(0,1,2)(1,0,1) ₁₂	2186.120	0.066	1650.370	1.353	1.744	1.666	3.932	B
(0,1,2)(2,0,1) ₁₂	2187.150	0.066	1650.930	1.352	1.744	1.666	3.925	B
(0,1,2)(1,0,2) ₁₂	2188.340	0.066	1651.740	1.350	1.744	1.664	3.918	B
(1,1,1)(1,0,1) ₁₂	2200.290	0.067	1664.510	1.332	1.718	1.648	3.770	B
(2,1,2)(2,2,0) ₁₂	2416.800	0.095	2159.340	1.004	1.000	1.000	1.004	C
(0,1,1)(1,0,0) ₁₂	2417.860	0.076	1890.940	1.002	1.487	1.351	2.015	C
(2,1,0)(1,0,0) ₁₂	2419.490	0.076	1894.440	1.000	1.487	1.347	2.003	C

With regard to research question (1)—*How can multiply error metrics be effectively combined to select a forecasting model?*, the findings indicate that KPI-based evaluation provides a flexible mechanism for comparing model performance across different parameter configurations. Specifically, the integration of multiple error metrics into a unified KPI framework enables systematic and consistent assessment. Consequently, this approach offers a robust methodology for identifying the most appropriate forecasting model based on multiple error evaluation criteria.

Models that exhibit conflicting results across different error metrics will be selected and further evaluated. This step aims to test the possibility of relying on a single error metric for assessing the performance of ARIMA or SARIMA models, in accordance with research question (2). The results of this analysis will be presented in the subsequent section.

2.3.5 Model accuracy and validation

A model selection method for ARIMA and SARIMA was developed by integrating three-error metrics (MAE, MAPE, and RMSE) to identify the most suitable model for efficient forecasting, as described in Section 2.3. The models were ranked in a descending order according to their KPI values. The results reveal inconsistencies across error metrics, where certain models exhibit conflicting performance outcomes (Tables 6 and 7).

Table 6 ARIMA model performance based on KPI ranking and conflicting error metrics

ARIMA Model	KPI	RMSE	MAPE	MAE
(1,2,1)	8.000	3942.086	6.300	3090.052
(0,2,1)	7.648	3971.999	6.300	3116.234

Table 7 SARIMA model performance based on KPI ranking and conflicting error metrics

SARIMA Model	KPI	RMSE	MAPE	MAE
(2,2,2)(1,0,1) ₁₂	7.888	1758.980	5.700	1397.160
(2,2,2)(1,0,2) ₁₂	7.816	1772.530	5.700	1395.400
(2,2,2)(2,0,0) ₁₂	7.763	1796.230	5.600	1397.640
(2,2,2)(2,0,1) ₁₂	7.628	1815.530	5.600	1401.230
(2,2,2)(1,0,0) ₁₂	7.492	1836.010	5.600	1403.910

Tables 6 and 7 summarize the performance of ARIMA and SARIMA models using a 69 month export dataset of Thailand. ARIMA(1,2,1) is identified as the optimal model when all three error metrics (RMSE, MAPE, and MAE) are jointly considered. However, this result differs from the MAPE-based evaluation, where ARIMA(0,2,1) also demonstrates superior performance.

When seasonal effects ($s=12$) were incorporated, SARIMA(2,2,2)(1,0,1)₁₂ achieves the lowest RMSE, whereas SARIMA(2,2,2)(1,0,2)₁₂ performs best in terms of MAE. By contrast, multiple models, such as SARIMA(2,2,2)(2,0,0)₁₂, SARIMA(2,2,2)(2,0,1)₁₂, and SARIMA(2,2,2)(1,0,0)₁₂, exhibit the lowest MAPE values.

These findings demonstrate that the choice of error metric significantly influences model selection, often resulting in conflicting conclusions. Accordingly, the proposed method integrates multiple error metrics to provide a robust and reliable framework for selecting the most appropriate forecasting model.

Models with conflicting error values were tested on a new dataset comprising export values from the top three HS-coded product groups (HS-84, HS-85, and HS-87) of Thailand [25] and its major Asian trading partners, namely, China [49], Japan [50], and the Republic of Korea [51], to further validate the proposed approach. This process resulted in 12 datasets, each with 66 monthly observations, representing different combinations of product groups and geographical markets. Although these series provide heterogeneous forecasting patterns within the export domain, they remain economically related datasets. Therefore, the present

validation should be interpreted as evidence of applicability within the export-forecasting context rather than as evidence of universal generalizability across time-series domains.

The analysis included two ARIMA models ARIMA(1,2,1) and ARIMA(0,2,1) and five SARIMA models. The model performance was evaluated using RMSE, MAPE, and MAE to assess the consistency of model selection across multiple error metrics. Non-parametric tests were applied to the RMSE, MAPE, and MAE values obtained from the 12 validation datasets to determine whether differences in forecasting performance were statistically significant. In ARIMA, the Wilcoxon signed-rank test was used because two models were compared using paired observations from the same datasets. In SARIMA, the Friedman test was used because five related models were simultaneously compared across the same datasets. These non-parametric tests were selected due to the small sample size ($n=12$) and their lack of a normality assumption. Statistical significance was evaluated at $\alpha=0.05$.

3. Results and discussion

3.1 Results

3.1.1 Results of the consistency between the ARIMA and SARIMA models

The consistency between the developed ARIMA models and the three-error metrics is presented in Table 8.

Table 8 Validation results of the proposed approach for ARIMA models

ARIMA Model	Error Metric	No. of Datasets	Consistency (%)	Exact Binomial p-Value
(1,2,1)	RMSE	10	83.333	0.019
(0,2,1)	RMSE	2	16.667	0.997
(1,2,1)	MAPE	9	75.000	0.073
(0,2,1)	MAPE	3	25.000	0.981
(1,2,1)	MAE	9	75.000	0.073
(0,2,1)	MAE	3	25.000	0.981

Table 8 shows that the observed percentage of agreement was not due to random variation. ARIMA(1,2,1) and ARIMA(0,2,1) were used. A precise one-sided binomial test was performed under the null hypothesis that each model had a 50% chance of achieving the lowest predictive error among the 12 datasets. The results indicated that ARIMA(1,2,1) achieved significantly higher agreement for RMSE (10/12 datasets, 83.33%; $p=0.019$), confirming that the superior RMSE performance was unlikely to be accidental. Although ARIMA(1,2,1) also showed high agreement for MAPE and MAE (9/12 datasets, 75.00%), these differences did not reach the 5% significance level ($p=0.073$), indicating a tendency toward significant differences. The consistency of the five SARIMA models across the three-error metrics is presented in Table 9.

Table 9 Validation results of the proposed approach for SARIMA models

SARIMA Model	Error Metric	No. of Datasets	Consistency (%)	Exact Binomial p-Value
(2,2,2)(1,0,1) ₁₂	RMSE	1	8.333	0.9313
(2,2,2)(1,0,2) ₁₂	RMSE	2	16.667	0.7251
(2,2,2)(2,0,0) ₁₂	RMSE	4	33.333	0.2054
(2,2,2)(2,0,1) ₁₂	RMSE	2	16.667	0.7251
(2,2,2)(1,0,0) ₁₂	RMSE	3	25.000	0.4417
(2,2,2)(1,0,1) ₁₂	MAPE	1	8.333	0.9313
(2,2,2)(1,0,2) ₁₂	MAPE	3	25.000	0.4417
(2,2,2)(2,0,0) ₁₂	MAPE	5	41.667	0.0726
(2,2,2)(2,0,1) ₁₂	MAPE	2	16.667	0.7251
(2,2,2)(1,0,0) ₁₂	MAPE	1	8.333	0.9313
(2,2,2)(1,0,1) ₁₂	MAE	1	8.333	0.9313
(2,2,2)(1,0,2) ₁₂	MAE	2	16.667	0.7251
(2,2,2)(2,0,0) ₁₂	MAE	5	41.667	0.0726
(2,2,2)(2,0,1) ₁₂	MAE	3	25.000	0.4417
(2,2,2)(1,0,0) ₁₂	MAE	1	8.333	0.9313

Table 9 show that SARIMA(2,2,2)(2,0,0)₁₂ attains the lowest RMSE in 4 out of 12 datasets (33.33%), representing the highest level of model consistency, followed by SARIMA(2,2,2)(1,0,0)₁₂ with three datasets (25%). By contrast, SARIMA(2,2,2)(1,0,2)₁₂ and SARIMA(2,2,2)(2,0,1)₁₂ each achieve optimal performance in two datasets (16.67%), while SARIMA(2,2,2)(1,0,1)₁₂ shows the lowest RMSE in only one dataset (8.33%).

SARIMA(2,2,2)(2,0,0)₁₂ attains the lowest MAPE in 5 out of 12 datasets (41.67%), demonstrating the highest level of model consistency, followed by SARIMA(2,2,2)(1,0,2)₁₂ with three datasets (25%). SARIMA(2,2,2)(2,0,1)₁₂ achieves optimal performance in two datasets (16.67%). Meanwhile, SARIMA(2,2,2)(1,0,1)₁₂ and SARIMA(2,2,2)(1,0,0)₁₂ each perform best in only one dataset (8.333%).

SARIMA(2,2,2)(2,0,0)₁₂ attains the lowest MAE in 5 out of 12 datasets (41.67%), demonstrating the highest level of model consistency, followed by SARIMA(2,2,2)(2,0,1)₁₂ with three datasets (25%). By contrast, SARIMA(2,2,2)(1,0,2)₁₂ achieves optimal performance in two datasets (16.67%), whereas SARIMA(2,2,2)(1,0,1)₁₂ and SARIMA(2,2,2)(1,0,0)₁₂ each perform best in only 1 dataset (8.33%).

An exact one-sided binomial test was conducted to statistically evaluate the consistency percentages of the SARIMA models. Given that five SARIMA models were compared, the null probability of achieving the lowest error by chance was specified as $p=0.20$. The SARIMA(2,2,2)(2,0,0)₁₂ model achieved the highest consistency for MAPE and MAE, with the lowest error in 5 of the 12 datasets

(41.67%). However, the corresponding binomial test result did not reach the conventional 5% significance level ($p=0.073$). In terms of RMSE, the same model achieved the lowest error in 4 of the 12 datasets (33.33%; $p=0.205$). Therefore, the results indicate a favorable performance trend for SARIMA (2,2,2)(2,0,0)₁₂, although the evidence was not statistically significant, possibly due to the limited number of validation datasets.

Furthermore, the results demonstrate why reliance on a single error metric may produce inconsistent model-selection decisions. Although ARIMA(1,2,1) exhibited the highest consistency across all three metrics, the degree of consistency differed among RMSE (83.33%), MAPE (75.00%), and MAE (75.00%). In the SARIMA models, the model rankings substantially varied across the three error measures. These differences support the motivation for integrating multiple error criteria within a unified decision-support framework rather than basing the final model-selection decision on a single error measure.

3.1.2 Statistical comparison of ARIMA and SARIMA model performance

A paired Wilcoxon signed-rank test was conducted to determine whether the performance differences between ARIMA(1,2,1) and ARIMA(0,2,1) were statistically significant across the 12 validation datasets (Table 10).

Table 10 Statistical comparison of ARIMA forecasting performance across the 12 validation datasets

Metric	ARIMA (1,2,1) Lower Error	ARIMA (0,2,1) Lower Error	Wilcoxon Statistic (W)	p-Value	Result
RMSE	10/12 (83.33%)	2/12 (16.67%)	14	0.026	Significant
MAPE	9/12 (75.00%)	3/12 (25.00%)	10	0.01	Significant
MAE	9/12 (75.00%)	3/12 (25.00%)	25	0.151	Insignificant

One-sided paired Wilcoxon signed-rank test; $\alpha=0.05$.

ARIMA(1,2,1) produced a lower RMSE in 10 of the 12 datasets (83.33%), and the difference was statistically significant ($W=14.0$, $p=0.026$). In terms of MAPE, ARIMA(1,2,1) achieved a lower error in 9 of the 12 datasets (75.00%), with a statistically significant difference ($W=10.0$, $p=0.010$). Although ARIMA(1,2,1) produced a lower MAE in 9 of the 12 datasets (75.00%), the difference was not statistically significant ($W=25.0$, $p=0.151$). These results provide inferential evidence that the performance advantage of ARIMA(1,2,1) was statistically supported for RMSE and MAPE, whereas the evidence for MAE remained descriptive rather than statistically significant.

A Friedman test was conducted on the five competing SARIMA models to evaluate whether forecasting performance significantly differed across the 12 validation datasets (Table 11).

Table 11 Statistical comparison of SARIMA forecasting performance across the 12 validation datasets

Metric	Friedman (X2)	p-value	Result ($\alpha=0.05$)
RMSE	6.933	0.139	Insignificant
MAPE	5.8	0.215	Insignificant
MAE	9.133	0.058	Insignificant

Table 12 Mean ranks of SARIMA models from the Friedman test across the 12 validation datasets

SARIMA Model	RMSE Mean Rank	MAPE Mean Rank	MAE Mean Rank
(2,2,2)(1,0,1) ₁₂	...	...	...
(2,2,2)(1,0,2) ₁₂	...	...	...
(2,2,2)(2,0,0) ₁₂	2.083	2.167	2
(2,2,2)(2,0,1) ₁₂	...	...	...
(2,2,2)(1,0,0) ₁₂	...	...	...
Friedman (X2)	6.933	5.8	9.133
p-value	0.139	0.215	0.058

SARIMA (2,2,2)(2,0,0)₁₂ achieved the lowest mean rank for all three error metrics, with mean ranks of 2.083 for RMSE, 2.167 for MAPE, and 2.000 for MAE (Table 12). This notion indicates that the model exhibited the most favorable overall forecasting performance across the 12 validation datasets among the five SARIMA models considered.

However, these rankings should be descriptively interpreted, as the Friedman test did not indicate statistically significant overall differences among the five models at the 0.05 significance level.

3.2 Discussion

These results are consistent with Wu [52], who reported that ARIMA(1,2,1) provided favorable forecasting performance based on RMSE and information criteria. In this study, ARIMA(1,2,1) demonstrated the highest consistency across the evaluated datasets, particularly for RMSE. However, the differences observed among RMSE, MAPE, and MAE indicate that model-selection outcomes remain dependent on the evaluation criterion. Therefore, the results support the use of multiple error metrics rather than exclusively relying on a single metric.

ARIMA(0,2,1) exhibited the lowest error under MAPE and MAE in 3 of the 12 datasets (25%), whereas RMSE seldom identified the optimal performance. This finding is consistent with Dlamini et al. [53], who also reported favorable MAE and MAPE performance for ARIMA(0,2,1). The inconsistency associated with MAPE can be explained by its mathematical formulation. Given that MAPE divides the absolute forecast error by the observed value, observations with relatively small magnitudes can disproportionately influence the resulting percentage error. In ARIMA(0,2,1), second-order differencing ($d=2$) removes level and local trend components before model estimation, which may generate forecast-error patterns that are evaluated differently by scale-dependent metrics, such as

RMSE and MAE and by the relative-error formulation of MAPE. Therefore, the observed disagreement should not be attributed to differencing alone, but rather to the interaction between model specification and the scale sensitivity of the error metric.

The results of the seasonal models are consistent with Yang et al. [54], where SARIMA(2,2,2)(2,0,0)₁₂ showed favorable forecasting performance. In this study, this model achieved the lowest mean rank across RMSE, MAPE, and MAE, although the Friedman test did not indicate statistically significant overall differences among the five SARIMA models.

The multiplicative aggregation improves model balance by minimizing the compensatory effect among individual error metrics. A model efficiently performing on one metric cannot fully compensate for relatively poor performance on another because all normalized components jointly determine the KPI. Consequently, the framework favors models that maintain consistently favorable performance across RMSE, MAPE, and MAE rather than models that perform exceptionally well under only one criterion.

The non-parametric tests further clarify the role of parameter selection. In the ARIMA models, the Wilcoxon signed-rank test provided statistically significant evidence favoring ARIMA(1,2,1) over ARIMA(0,2,1) for RMSE ($p=0.026$) and MAPE ($p=0.010$), whereas the MAE difference was insignificant ($p=0.151$). Accordingly, the statistical test provides inferential support for the (1,2,1) parameter configuration beyond descriptive error comparisons alone. However, the Friedman tests for SARIMA showed no statistically significant overall differences among the five parameter configurations for RMSE, MAPE, or MAE ($p>0.05$). These findings indicate that statistical testing complements, rather than replaces, the proposed KPI-based model-selection framework.

The proposed framework does not replace conventional criteria, such as AIC and BIC, but complements them by providing a systematic approach for selecting forecasting models when multiple forecasting error measures produce conflicting rankings. Unlike traditional model-selection methods based primarily on information criteria, the proposed framework focuses on resolving inconsistencies among multiple forecasting accuracy measures after model estimation. Therefore, the contribution of this framework lies in supporting objective model selection under conflicting evaluation criteria rather than replacing conventional statistical model-identification procedures.

However, the interpretation of the forecasting results should consider the exceptional economic conditions represented in the study period. In particular, the COVID-19 pandemic generated abrupt changes in international trade and supply-chain conditions that may not be fully captured by conventional ARIMA and SARIMA structures. Given that no explicit intervention component was incorporated, part of the observed forecasting errors may reflect crisis-driven structural disturbances rather than regular stochastic variation. Therefore, the model-selection results should be interpreted as comparative performance over the complete observed historical period rather than as forecasting performance under business-as-usual conditions alone.

4. Conclusions and future research

4.1 Conclusions

The findings demonstrate that reliance on a single forecasting error metric may result in different model-selection decisions because RMSE, MAPE, and MAE emphasize varying characteristics of forecast errors. The proposed framework integrates multiple normalized error metrics into a composite KPI rather than selecting one metric as universally preferable to support a more systematic and consistent model-selection process. Therefore, the framework serves as a decision-support mechanism for resolving conflicting rankings among individual error metrics, rather than replacing the individual metrics themselves.

The out-of-sample validation results further demonstrated that ARIMA(1,2,1) achieved the lowest RMSE in 10 of the 12 datasets (83.33%) and the lowest MAPE and MAE in 9 of the 12 datasets (75.00%). An exact one-sided binomial test showed that the RMSE consistency was statistically significant ($p=0.019$), while the MAPE and MAE results showed favorable but insignificant trends ($p=0.073$). These findings provide statistical support for the observed consistency patterns while also indicating that the strength of evidence varies across individual error metrics.

The findings of this study are consistent with previous research reporting that different forecasting error metrics may produce varying model-selection outcomes [14-16]. Similar to recent studies that have adopted systematic model-selection approaches, including out-of-sample evaluation [17], MCDM [18, 19], and data-driven model selection [20], the results confirm that model selection should not rely solely on a single evaluation criterion.

However, unlike these previous studies, the proposed framework focuses on evaluating the consistency of model-selection decisions across multiple datasets using three commonly adopted forecasting error metrics (RMSE, MAPE, and MAE). This consistency-based perspective extends the current literature by providing a practical framework for comparing ARIMA and SARIMA models without requiring subjective weighting schemes or complex optimization procedures.

These findings are consistent with previous studies reporting that forecasting model performance depends on the choice of evaluation metric, while further demonstrating that consistency across multiple datasets can provide an additional perspective for supporting forecasting model selection.

4.2 Future research

The proposed framework has several limitations that should be acknowledged.

1. Validation experiments were exclusively conducted using export forecasting datasets. Although multiple product groups and trading partners were considered, the applicability of the framework to other forecasting domains, such as energy demand, healthcare, transportation, financial markets, and environmental forecasting, remains to be investigated.

2. Proposed framework integrates three commonly used forecasting error metrics (RMSE, MAE, and MAPE). Future studies may investigate additional evaluation measures, including symmetric MAPE, mean absolute scaled error, and probabilistic forecasting metrics.

3. K-means clustering provides an effective mechanism for grouping forecasting models according to their composite KPI values. Future research may compare alternative clustering techniques or multi-criteria optimization approaches to further improve model selection robustness.

4. The proposed framework should be compared with conventional model selection approaches, including AIC-, BIC-, and naïve forecasting-based selection strategies, across broader forecasting applications to further evaluate their relative advantages and limitations.

5. Alternative normalization procedures should be evaluated, MCDM approaches, such as TOPSIS, established, and sensitivity and rank-stability analyses should be conducted to determine whether model-selection decisions remain stable under different scaling and aggregation schemes.

5. Acknowledgments

The authors acknowledge the use of publicly available trade statistics obtained from Thailand Export Data and Top Export Products, the Japan External Trade Organization (JETRO), the Korea Customs Service (KCS), and the General Administration of Customs of the People's Republic of China (GACC), which have supported the analysis presented in this study.

6. References

- [1] Heizer J, Render B, Munson C. Operations management: sustainability and supply chain management. 13th ed. Harlow: Pearson; 2020.
- [2] Makridakis S, Spiliotis E, Assimakopoulos V. The M4 Competition: 100,000 time series and 61 forecasting methods. *Int J Forecast.* 2020;36(1):54-74.
- [3] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 3rd ed. Melbourne: OTexts; 2021.
- [4] Masini RP, Medeiros MC, Mendes EF. Machine learning advances for time series forecasting. *J Econ Surv.* 2023;37(1):76-111.
- [5] Hutahut AH, Anggraeni W, Tyasnurita R. Development of a decision support application for forecasting raw material inventory using the ARIMA (autoregressive integrated moving average) method at CV Asia. *Jurnal Teknik Pomits.* 2014;3(2):169-74. (In Indonesian)
- [6] Silalahi DK. Forecasting of poverty data using seasonal ARIMA modeling in West Java Province. *Jurnal Teori dan Aplikasi Matematika.* 2020;4(1):76-86.
- [7] Sankalp S, Panda PK. A comparative evaluation of machine learning and ARIMA models for forecasting relative humidity over Odisha districts. In: Kasiviswanathan KS, Soundharajan B, Patidar S, He J, Ojha CSP, editors. *Modeling and Mitigation Measures for Managing Extreme Hydrometeorological Events Under a Warming Climate. Developments in Environmental Science.* Vol. 14. Amsterdam: Elsevier; 2023. p. 91-105.
- [8] Mgaya JF. Application of ARIMA models in forecasting livestock products consumption in Tanzania. *Cogent Food Agric.* 2019;5(1):1607430.
- [9] Chyon FA, Suman MNH, Fahim MRI, Ahmmed MS. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. *J Virol Methods.* 2022;301:114433.
- [10] Eiglsperger J, Haselbeck F, Grimm DG. ForeTiS: A comprehensive time series forecasting framework in Python. *Mach Learn Appl.* 2023;12:100467.
- [11] Ariyanti VP, Yusnitasari T. Comparison of ARIMA and SARIMA for forecasting crude oil prices. *Jurnal Resti.* 2023;7(2):405-13.
- [12] Ning Y, Kazemi H, Tahmasebi P. A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. *Comput Geosci.* 2022;164:105126.
- [13] Suppalakpanya K, Nikhom R, Booranawong T, Booranawong A. Forecasting oil palm and crude palm oil data in Thailand using exponential time-series methods. *Eng Appl Sci Res.* 2019;46(1):44-55.
- [14] Venkataraman V, Prasad S, Aswathanarayana B, Barigidad S, Nayak V, Kumar NST. Development of time series models for various pollutants in Bangalore city using the Akaike information criterion. *Eng Appl Sci Res.* 2020;47(3):249-63.
- [15] DeLurgio SA. Forecasting principles and applications. New York: McGraw-Hill; 1998.
- [16] Demir L, Akkaş S. A comparison of sales forecasting methods for a feed company: a case study. *Pamukkale University Journal of Engineering Sciences.* 2018;24(4):705-12.
- [17] Zhang Y, Na S. A novel agricultural commodity price forecasting model based on fuzzy information granulation and MEA-SVM model. *Math Probl Eng.* 2018;2018(1):1-10.
- [18] Athipayoon N, Thongteeraparp A, Chaipitak S. Comparison of forecasting methods for non-linear and non-stationary time series. *Thai Science and Technology Journal (TSTJ).* 2020;28(2):197-207. (In Thai)
- [19] Poler R, Mula J. Forecasting model selection through out-of-sample rolling horizon weighted errors. *Expert Syst Appl.* 2011;38(12):14778-85.
- [20] Badulescu Y, Hameri AP, Cheikhrouhou N. Evaluating demand forecasting models using multi-criteria decision-making approach. *J Adv Manag Res.* 2021;18(5):661-83.
- [21] Koubaa Z, El-Amraoui A, Frikha A, Delmotte F. Multicriteria decision making for selecting forecasting electricity demand models. *Sustainability.* 2024;16(21):9219.
- [22] Stroilă AM. Enhancing univariate time series forecasting: a general tree-based model selection approach. *Comput Stat.* 2026;41:46.
- [23] Mahanil N, Khemavut P. A conceptual model for developing a selection error measurement method to measure the effectiveness of the time series forecasting method. 13th Annual International Conference on Industrial Engineering and Operations Management; 2023 Mar 7-9; Manila, Philippines. Manila: IEOM; 2023. p. 1992-2003.
- [24] European Commission. Harmonised system [Internet]. 2026 [cited 2026 Apr 4]. Available from: https://taxation-customs.ec.europa.eu/customs/customs-tariff/tariff-classification-goods/harmonised-system_en.
- [25] Trading Economics. Thailand exports [Internet]. 2026 [cited 2026 Apr 4]. Available from: <https://tradingeconomics.com/thailand/exports>.
- [26] Tendata. Thailand export data and top export products [Internet]. 2026 [cited 2026 Apr 4]. Available from: https://tradereport.moc.go.th/en/stat/report02?utm_source=chatgpt.com
- [27] Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. 5th ed. Hoboken: Wiley; 2015.
- [28] Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. *J Stat Softw.* 2008;27(3):1-22.
- [29] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 2nd ed. Melbourne: OTexts; 2018.
- [30] McKinney W. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference (SciPy 2010); 2010 Jun 28 – Jul 3; Austin, Texas. p. 56-61.

- [31] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585:357-62.
- [32] Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference (SciPy 2010)*; 2010 Jun 28 – Jul 3; Austin, Texas. p. 92-6.
- [33] James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning: with applications in R*. 2nd ed. New York: Springer; 2021.
- [34] Dickey DA, Fuller WA. Distribution of the estimators for autoregressive time series with a unit root. *J Am Stat Assoc*. 1979;74(366):427-31.
- [35] National Institute of Standards and Technology (NIST). Box-Jenkins model identification [Internet]. 2026 [cited 2026 Apr 4]. Available from: <https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc446.htm>.
- [36] Hamilton JD. *Time series analysis*. Princeton: Princeton University Press; 1994.
- [37] Brockwell PJ, Davis RA. *Introduction to time series and forecasting*. 3rd ed. Cham: Springer; 2016.
- [38] Taylor BW, Russell RS. *Operations management: creating value along the supply chain*. 7th ed. Hoboken: John Wiley & Sons; 2011.
- [39] Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE). *Clim Res*. 2005;30(1):79-82.
- [40] de Myttenaere A, Golden B, Le Grand B, Rossi F. Mean absolute percentage error for regression models. *Neurocomputing*. 2016;192:38-48.
- [41] Shalabi LA, Shaaban Z, Kasasbeh B. Data mining: a preprocessing engine. *J Comput Sci*. 2006;2(9):735-9.
- [42] Han J, Kamber M, Pei J. *Data mining: concepts and techniques*. 3rd ed. Waltham: Morgan Kaufmann; 2012.
- [43] Yoon K. A reconciliation among discrete compromise solutions. *J Oper Res Soc*. 1987;38(3):277-86.
- [44] Çelen A. Comparative analysis of normalization procedures in TOPSIS method: With an application to Turkish deposit banking market. *Informatica*. 2014;25(2):185-208.
- [45] Thorndike RL. Who belongs in the family?. *Psychometrika*. 1953;18:267-76.
- [46] MacQueen J. Some methods for classification and analysis of multivariate observations. *Berkeley Symp Math Statist Prob*. 1967:281-97.
- [47] Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognit Lett*. 2010;31(8):651-66.
- [48] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer; 2009.
- [49] General Administration of Customs of the People's Republic of China (GACC). China customs statistics [Internet]. 2026 [cited 2026 Apr 4]. Available from: <http://english.customs.gov.cn/>.
- [50] Japan External Trade Organization (JETRO). Japan trade statistics [Internet]. 2026 [cited 2026 Apr 4]. Available from: <https://www.jetro.go.jp/en/reports/statistics.html>.
- [51] Korea Customs Service (KCS). Trade statistics [Internet]. 2026 [cited 2026 Apr 4]. Available from: <https://unipass.customs.go.kr>.
- [52] Wu D. A comparative analysis of ARIMA models for forecasting China's GDP. *Adv Oper Res Prod Manag*. 2025;4(2):1-11.
- [53] Dlamini WJ, Melesse SF, Mwambi HG. A comparative study of ARIMA, ARFIMA, and ARIMA-ARFIMA models in predicting under-five mortality rate in four East-African Nations. *Open Public Health J*. 2026;19:1-10.
- [54] Yang W, Su A, Ding L. Application of exponential smoothing method and SARIMA model in predicting the number of admissions in a third-class hospital in Zhejiang Province. *BMC Public Health*. 2023;23(1):2309.