Forecasting the Exchange Rate for the Thai Baht Against the Chinese Yuan by Using a Genetic Algorithm-Based Subset Autoregressive Integrated Moving Average Model

Tassathorn Poonsin, Vayu Thanomsing, Thanakorn Thunjang,
and Worrawate Leela-apiradee(✉)

Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University, Pathum Thani 12121, Thailand
worrawate@mathstat.sci.tu.ac.th

Abstract. Accurate forecasting of foreign exchange rates plays a crucial role in future global financial market investment, international business decision-making, and travel planning. This paper proposes a model for forecasting the daily exchange rate for the Thai baht (THB) against the Chinese yuan (CNY) during the Novel Coronavirus 2019 (COVID-19) pandemic by comparing a genetic algorithm (GA)-based subset autoregressive integrated moving average (ARIMA) model to the classical ARIMA model. Data was gathered from April 1, 2020 to April 14, 2022. Forecast accuracy was measured by mean absolute percentage error (MAPE), root mean squared error (RMSE) and mean absolute error (MAE). A GARI program was developed using non-seasonal time series prediction, with the best model ARIMA$(4, 1, \{1, 5\})$ forecasting daily CYN/THB exchange rate attaining a nadir of MAPE, RMSE and MAE at $1.2180\%, 0.066674$ and 0.064061, respectively. These findings indicate that the GA-based subset ARIMA model via GARI program outperformed the classical ARIMA model in the auto ARIMA with Python. This program may be applicable for predicting other foreign exchange rates and non-seasonal time series data.

Keywords: ARIMA model · Foreign exchange rates · Genetic algorithm · Time series · Chinese yuan currency

1 Introduction

Currently, each country has a different currency to use as a medium of exchange. In a conduct of international financial transactions, foreign exchange is required to be involved. In particular, those who need to study the trends in daily exchange rates for a month ahead including investors in foreign stocks, individuals who buy/sell foreign products, or related companies.

N. Ngoc Thach et al. (Eds.): *Optimal Transport Statistics for Economics and Related Topics*,
SSDC 483, pp. 544–559, 2024.
https://doi.org/10.1007/978-3-031-35763-3_39

Foreign exchange rates are quantitative data that can be collected in many forms. One of the most popular formats is the time series, which is a sequence of discrete-time data collected chronologically in succession. Examples of time series are stock indices daily closing prices, wholesale weekly prices of agricultural products, and the monthly temperature average. We can say that the data correlated or recorded in conjunction with time is very interesting. This information can also reflect a particular event that occurred at that time. For instance,

- The baht was weakened sharply from 25 to 55 per *U.S. dollar (USD)* on July 2, 1997 as a result of speculation on the baht that has continued since the beginning of 1997. This affected lacking confidence of private businesses in the baht. It was deemed necessary to promulgate a floating exchange rate system by the Ministry of Finance and Bank of Thailand.
- The yuan was valued at 6.9999 and weaken to 7.0240 against the USD as of August 1, 2019. It was the first time that the yuan was dropped to 7 in more than a decade. As a result of trade war, China would let the yuan depreciate in order to help export sectors. Therefore, investors in global financial markets were concerned.
- On February 28, 2022, the ruble fell sharply to around 100.96 per USD compared to 83.5 as of February 23, 2022, which is the day before Russia's invasion of Ukraine.

It can be seen from the above three events that the big events happening in a country have impacted on the fluctuation of the country's currency. How could it be if we know the exchange rates in advance? Of course, it gives information to develop data-driven strategies and make decisions for international investors, a person or business which sells/buys goods from abroad, and those who are planning a trip oversea. These are the reason why we are interested in foreign exchange rate forecasting in the form of time series data in this paper.

Many researchers have applied statistical models and machine learning algorithms to time series forecasting, such as *Autoregressive Integrated Moving Average (ARIMA)*, *Nonlinear Autoregressive (NAR)* neural network, *Susceptible-Infectious-Recovered (SIR)*, *Long Short-Term Memory (LSTM)*, *genetic algorithm (GA)*, and *Artificial Neural Network (ANN)*, etc., which can be found in the following publications.

The best ARIMA model presented in [20] was selected by considering the smallest values of the criteria *Akaike Information Criterion (AIC)*, *Schwartz Information Criterion (SIC)*, *Mean Absolute Error (MAE)*, *Root Mean Squared Error (RMSE)*, and *Mean Absolute Percentage Error (MAPE)* in order to predict average daily share price indices of Square Pharmaceuticals Limited with non-stationary data series. Besides ARIMA, a commonly-used statistical model for time series forecasting is *Exponential Smoothing (ETS)*, which has automatic prediction strategies in Python package called auto ETS. Both ARIMA and ETS were applied in [16] to estimate daily exchange rates of the Romanian Leu against other currencies.

Swaraj et al. proposed in [23] a new model ARIMA-NAR that combines the ARIMA model with the NAR algorithm for prediction of COVID-19 cases in India. The ARIMA-NAR model provided better prediction results with low values of RMSE, MAE, and MAPE compared to the single ARIMA model. This article claimed that the ARIMA-NAR model outperforms the SIR and LSTM algorithms for short-term forecasts. Moreover, the study [17] verified unability of the SIR in the long term forecast through the outbreak COVID-19 datasets in Isfahan province of Iran. Recently, the neural network LSTM has been tested its efficacy via streamflow prediction at ten river gauge stations across various climatic regions of the western United States [11], and power consumption in some French cities [15].

A two-level multi-objective GA was established by Al-Douri et al. in [1] to optimize the prediction of time series data on fans used in road tunnels according to the data from the Swedish Transport Administration. The two levels consist of a multi-objective GA implementation and the multi-objective GA utilization to identify an appropriate forecasting. The forecasting data for two life cycle costs obtained from their proposed models was neither realistic nor close to the actual data. A forecasting method integrating GA and *Autoregressive Moving Average (ARMA)*, or briefly called GA-ARMA, was proposed by Ervural et al., [7] to predict the monthly natural gas consumption of İstanbul collected from January 2004 to October 2015, where the fitness function of the GA was specified by MAPE, In [19], the GA was also used to identify ARMA, ARIMA and SARIMA models applied to semiconductor industry in terms of DRAM price forecasting. In addition, [25] has recently utilized a hybrid model of the GA and ARIMA, or briefly called GA-ARMA, to predict drought in synoptic station of Tabriz in northwestern Iran by investigating the Standard Precipitation and Evapotranspiration Index in the short-term, mid-term, and long-term steps of 53 years period. By comparing with the traditional models, the articles [7,19] and [25] summarized that the GA-based models provided more accurate results.

Foreign exchange rate forecasting has been studied in the literature [3,13,24] and [25]. The yearly exchange rates of Kazakhstan currency: USD/KZT, EUR/KZT and SGD/KZT over the period from 2006 to 2014 were analyzed in [25] using MAE, MAPE and RMSE to measure the forecast accuracy. [13] developed three ANN based forecasting models using standard backpropagation, scaled conjugate gradient, and Baysian regression to predict currency exchange rates of Australian dollar with six other currencies: USD, GBP, JPY, SGD, NZD and CHF. All the ANNs outperformed the traditional ARIMA model. Furthermore, forecasting exchange rate between Thai baht and U.S. dollar was investigated using time series analysis found in [3], and using data mining technique found in [24]. In 2014, Bowornchockchai [3] used Box-Jenkins and Holt's methods to forecast the monthly exchange rate. The author reported that the Box-Jenkins is the most suitable one. In 2020, the nine algorithms: Naive Bayes, generalized linear model, logistic regression, fast large margin, deep learning, decision tree, random forest, gradient boosted trees, and support vector machine were applied in [24] to predict the monthly exchange rate. The results of the study showed that logistic regression was reached the highest accuracy together with the three most significant correlated factors: U.S. dollar price index, gold price, and Nasdaq price index. In recent works, the LSTM neural network has been used to forecast foreign exchange rates in [4,21] and [27], or even directional movement of Forex data in [28].

The remainder of the paper is organized as follows. In the next section, we define ARIMA model mathematically together with introducing auto ARIMA at the end of the section. The concept of subset ARIMA model based on GA is explained in Sect. 3. We add user's guide for using our developed program GARI as well as a link for readers who want to download and install the program. The performance of the model presented in Sect. 4 is tested using the criteria MAPE, RMSE and MAE, where the lowest value of them gives the highest accuracy model. The conclusion of the article is addressed in the last section.

2 ARIMA Model

Time series data is a data set collected sequentially in succession over equal periods of time. The period can be considered as a day, a month, a quarter, or a year. However, time series analysis is necessary to take into account the following four components of variation: trend, seasonal, cyclical and irregular components. The goal of the analysis is to identify relationships between observed values in the past through a model and use it to predict future values. The traditional forecasting models ARMA and ARIMA are described as follows.

The ARMA (Autoregressive Moving Average) model was popularized in 1970 by Box and Jenkins [9] developed from the Box-Jenkins method. The model is a combination of two models between *Autoregressive (AR)* and *Moving Average (MA)*, which are introduced in Definitions 2 and 3, respectively. Since ARMA is the most effective linear model for stationary time series forecasting compared with the pure AR and MA models, it is a popular and widely used model nowadays. The concept of the stationary time series can be written as the mathematical definition below.

Definition 1. Let $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ be the set of indices. The time series $\{X_t : t \in \mathbb{Z}\}$ is said to be **stationary** if the following three statements hold:

1. $E(X_t) = \mu$ for all $t \in \mathbb{Z}$.
2. $\mathrm{Var}(X_t) = \sigma^2 < \infty$ for all $t \in \mathbb{Z}$.
3. $\mathrm{Cov}(X_r, X_s) = \mathrm{Cov}(X_{r+t}, X_{s+t})$ for all $r, s, t \in \mathbb{Z}$.

In other words, the time series $\{X_t : t \in \mathbb{Z}\}$ is stationary if the mean and variance values are constants without depending on time t, while the covariance depends on r and s only through their difference $|r - s|$.

Definition 2. Let $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ be the set of indices. The **AR model** of order p, denoted by AR(p), is of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t \tag{1}$$

Here, $\{X_t : t \in \mathbb{Z}\}$ is a stationary time series where $\phi_1, \phi_2, \dots, \phi_p$ are parameters of the model such that $\phi_p \neq 0$ and ϵ_t is a white noise error at time t.

Let $t \in \mathbb{Z}$. When the mean μ of X_t is nonzero, we can replace X_{t-i} of Eq. (1) with $X_{t-i} - \mu$ for each $i \in \{0, 1, \ldots, p\}$ and obtain

$$\begin{aligned} X_t - \mu &= \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \epsilon_t \\ X_t &= (\mu - \mu\phi_1 - \mu\phi_2 - \cdots - \mu\phi_p) + (\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p}) + \epsilon_t \\ X_t &= \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p) + (\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p}) + \epsilon_t \\ X_t &= \beta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t, \end{aligned} \tag{2}$$

where $\beta = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$. It can be inferred that Eq. (2) is similar to the regression model, which has independent variables as its own previous values. Therefore, we call Eq. (1) an autoregression model as a result of "auto" here referring to "self" in this context.

Definition 3. The **MA model** of order q, denoted by MA(q), is of the form

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \tag{3}$$

where $\theta_1, \theta_2, \ldots, \theta_q$ are parameters of the model such that $\theta_q \neq 0$ and $\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q}$ are the current and various past values of a white noise error.

The AR and MA models lead to the concept of ARMA model, which was firstly described in 1951 by Whittle [26] as presented in the definition below.

Definition 4. Let $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ be the set of indices. The **ARMA model** of a given stationary time series $\{X_t : t \in \mathbb{Z}\}$ with p autoregressive terms and q moving average terms, denoted by ARMA(p, q), is of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \tag{4}$$

where $\phi_1, \phi_2, \ldots, \phi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ are parameters of the model such that $\phi_p \neq 0$ and $\theta_q \neq 0$ together with the current and various past values of a white noise error $\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q}$, respectively.

It is easy to see from Eq. (4) that the ARMA model is a generalization of AR and MA models. That is,

- if $q = 0$, then the ARMA model becomes autoregressive model of order p, or AR(p),
- if $p = 0$, then the ARMA model turns into moving average model of order q, or MA(q).

In practice, the ARMA model cannot be applied to "non-stationary" time series data but it is necessary to transform the data to be "stationary" first by performing difference operation. This is the underlying process of **ARIMA model**:

Let $\{X_t : t \in \mathbb{Z}\}$ be the original time series with non-stationary data.

$$\text{If } \Delta^d X_t \text{ is ARMA}(p, q), \text{ then } X_t \text{ is ARIMA}(p, d, q),$$

The term $\Delta^d X_t$ is called **differencing operator** of order d, which comes from "Integrated" concept. Throughout the paper, the notation ARIMA(p, d, q) refers to the non-seasonal ARIMA models with prediction equation:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \quad (5)$$

where $Z_* = \Delta^d X_*$ for all indices $* \in \{t, t-1, \ldots, t-p\}$. The nonnegative integers p, d and q represent as the table below.

Value	Meaning
p	The number of the AR terms
d	The number of differences needed for stationarity
q	The number of lagged forecast errors in the prediction equation

The values p, d and q can be generated in auto ARIMA with python using "pmdarima" package. The steps how it works is explained as seen in Algorithm 1. The model selection criteria AIC and BIC showing up in the algorithm are described in [8].

Algorithm 1 : The auto ARIMA procedure.

```
Step 1: Use augmented Dickey-Fuller (ADF) test to check whether the time series data
        is stationary or not.
        • If the data is stationary, set $d = 0$.
        • If the data is non-stationary, perform the difference to find the order $d$ of the
          model.
Step 2: Identify the suitable orders $p$ and $q$ of the corresponding AR and MA parts
        by using plots of the ACF (Autocorrelation Function) and PACF (Partial
        Autocorrelation Function), respectively.
Step 3: Estimate the parameters $\phi_1, \phi_2, \ldots, \phi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ of the model.
Step 4: Improve the model to be the best fit one by comparing AIC (Akaike Information
        Criterion) and BIC (Bayesian Information Criterion) values.
```

3 GA-Based Subset ARIMA Model

GA (Genetic Algorithm) was invented in early 1970s by Holland and his students at the University of Michigan as an evolutionary optimization method developed from a random search algorithm in conjunction with the survival concept of the strongest individuals according to Darwin's theory of evolution. Due to the ability of GA to solve complex and nonlinear problems, it has been applied to the field of artificial intelligence in recent times.

The GA searches for the globally accepted optimal solution, which is simply called the best solution, to an optimization problem within a reasonable time.

Initially, the algorithm generates randomly a population of candidate solutions, each of which is treated as a chromosome (an individual) of the population. Later, fitness value for each chromosome is evaluated. The chromosomes selected based on their fitness values in selection process become parents in reproduction process in order to produce offspring of the next generation by performing evolutionary strategies including crossover and mutation.

Recalling Algorithm 1, let d be a fixed order of differencing according to Step 1. In Step 3, we can apply a library named "statsmodels" in Python to estimate the parameters. The idea of GA-ARIMA arises from the use of GA to search for the best ARIMA model reached the highest prediction accuracy instead of performing traditional Steps 2 and 4. Since ARIMA with low orders can sufficiently model most time series. To restrict the search space of the GA, the orders p and q focused here do not exceed 5, i.e., $p_{\max} = 5$ and $q_{\max} = 5$, where $p_{\max}$ and $q_{\max}$ denote the maximum possible orders corresponding to MA and AR models. With those orders, the prediction Eq. (5) is expressed as

$$\begin{aligned} Z_t = {} & \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \phi_4 Z_{t-4} + \phi_5 Z_{t-5} + \epsilon_t \\ & + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_3 \epsilon_{t-3} + \theta_4 \epsilon_{t-4} + \theta_5 \epsilon_{t-5}, \end{aligned} \tag{6}$$

where $Z_* = \Delta^d X_*$ for all indices $* \in \{t, t-1, \dots, t-5\}$. However, if some coefficient parameters of Eq. (6) are set to zero, the model is called a **subset ARIMA model**, which is an extended version of the ARIMA model introduced by Lee and Fambro [14]. Throughout the paper, the subset ARIMA model is denoted by ARIMA(P, d, Q), where P and Q are nonempty subset of $\{1, 2, 3, 4, 5\}$. For instance, a model ARIMA$(\{1, 2, 3, 4\}, 1, \{1, 5\})$ implies that the parameters $\phi_1, \phi_2, \phi_3, \phi_4$ exist in the AR part, and θ_1, θ_5 exist in the MA part, and the others are set to zero. This leads to the prediction equation

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + \phi_4 Z_{t-4} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_5 \epsilon_{t-5}, \tag{7}$$

where $Z_* = \Delta X_*$ for all indices $* \in \{t, t-1, t-2, t-3, t-4\}$. Note that if the set P (or Q) is $\{1, 2, \dots, r\}$, we represent it as the number at the end r for convenience. So, the above model is briefly written by ARIMA$(4, 1, \{1, 5\})$. To generate the possible sets P and Q of the subset ARIMA models and search for the best one using the GA, we need to define a chromosome representation used in the algorithm as a string of length $p_{\max} + q_{\max} = 5 + 5 = 10$ depicted in Fig. 1. For each $i, j \in \{1, 2, 3, 4, 5\}$, the u_i and v_j on the chromosome are valued as (8), whose values indicate the existence of the ϕ_i and θ_j in the AR and MA parts of the model, respectively.

$$u_i = \begin{cases} 1, & \text{if } i \in P; \\ 0, & \text{if } i \notin P, \end{cases} \quad \text{and } u_j = \begin{cases} 1, & \text{if } j \in Q; \\ 0, & \text{if } j \notin Q. \end{cases} \tag{8}$$

u_1	u_2	u_3	u_4	u_5	v_1	v_2	v_3	v_4	v_5

Fig. 1. a chromosome representation of the GA in GA-ARIMA model.

With the following chromosome for example,

1	0	0	1	1	1	1	0	0	0

we can convert it to the model ARIMA$(\{1,4,5\}, d, \{1,2\})$, where d is an appropriate differencing order according to the ADF test, whose prediction equation is

$$Z_t = \phi_1 Z_{t-1} + \phi_4 Z_{t-4} + \phi_5 Z_{t-5} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2},$$

where $Z_* = \Delta^d X_*$ for all indices $* \in \{t, t-1, t-4, t-5\}$.

For the improvement step of the model, we try to find out the best chromosome with the smallest value of criteria MAPE, RMSE and MAE as our error measures. Let n be the total number of fitted points. The above criteria can be calculated by

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{X_t - \hat{X}_t}{X_t} \right|, \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{X}_t)^2}, \tag{10}$$

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^{n} \left| X_t - \hat{X}_t \right|, \tag{11}$$

where X_t and $\hat{X}_t$ are the actual and forecast values at time t, respectively.

As the preceding, the proposed procedure in accordance with GA-based subset ARIMA model can be concluded as a flowchart in Fig. 2. In this work, we also develop an application program named "GARI" that comes from GA-ARIMA for Investment in terms of future trends prediction. This program is suitable for not only predicting the exchange rate data but applying to any time series data with non-seasonality and univariate (stationarity) non-stationarity that could be made stationarity by differencing. Note that there are a number of non-stationary time series data, which does not work for taking over-differencing in order to ensure the stationarity according to the paper [10]. Readers can download and install the program from the link: https://drive.google.com/drive/folders/1Z9vEpzfqICDGdL8bjyEV-fUL7ZLsNn5e?usp=sharing. When the installation is complete, the main window will appear displayed as Fig. 3. The parameters of the program defined as Table 1 need to be set at the beginning. Users can learn more about their description in Step 3 of Algorithm 2. In that algorithm, we record the step-by-step implementation of GARI. Moreover, our program provides two additional options:

1. FAST GA: This option enables the users to forecast more comfortably, and the maximum acceptable error is required. Without this option, please choose “disable” mode.
2. Show detail: The chromosomes in each generation will be displayed in details on the lower left-hand side of GARI’s window. With choosing “disable’ mode, the details will be concise.

Normally, we would be able to select “disable” mode for the above options. The program will lead us to the best ARIMA model displayed in “Summary” window. In addition, time series plots of the exchange rate including actual and forecast data are illustrated in “Graph” window. Eventually, we can see the forecast values in “Table” window.

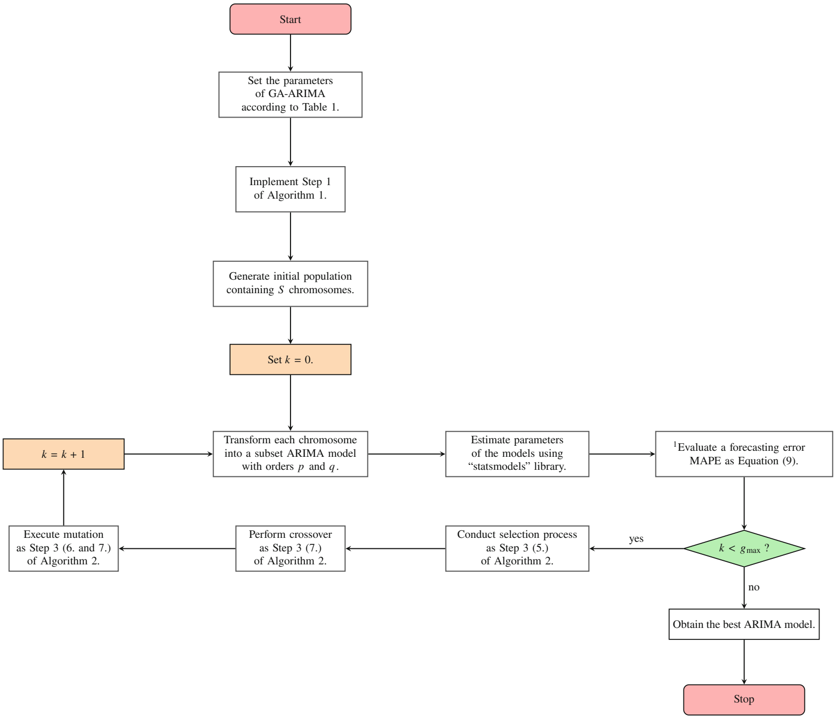

Start
Set the parameters of GA-ARIMA according to Table 1.
Implement Step 1 of Algorithm 1.
Generate initial population containing S chromosomes.
Set $k = 0$.
$k = k + 1$
Transform each chromosome into a subset ARIMA model with orders p and q.
Estimate parameters of the models using “statsmodels” library.
[1]Evaluate a forecasting error MAPE as Equation (9).
Execute mutation as Step 3 (6. and 7.) of Algorithm 2.
Perform crossover as Step 3 (7.) of Algorithm 2.
Conduct selection process as Step 3 (5.) of Algorithm 2.
yes
$k < g_{\max}$?
no
Obtain the best ARIMA model.
Stop

Fig. 2. Flowchart for the GA-ARIMA procedure. ([1] The evaluation could be replaced by RMSE and MAE.)

Table 1. Notation and meaning of parameters used in the algorithm.

Symbol	Meaning	Setting value
C	Confidence level (in percentage).	95%
T	The amount of testing data.	30 days
$g_{\max}$	The maximum number of generations in the algorithm, written herein as "Max generation" for short	200
S	Population size.	10
s	The number of survivors	5
m	The number of mutation points	1
R_m	Mutation rate (in percentage).	30%
R_c	Crossover rate (in percentage).	80%

Algorithm 2 : The GA-ARIMA procedure to predict CNY/THB via our proposed program.

Step 1: Upload a data set accepted only file types '.csv' and '.xlsx'.

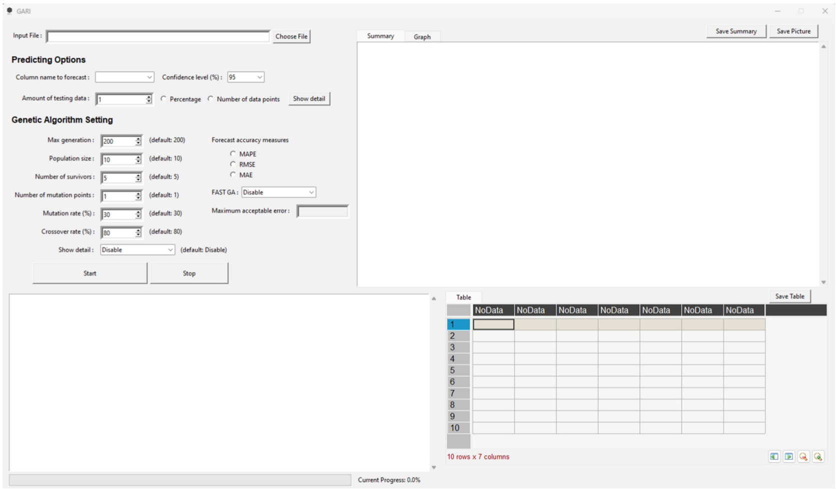
Step 2: Choose a column name addressed the daily exchange rate data.

Step 3: Set inputs related to predicting options and genetic algorithm setting mentioned in Table 1. Explanation of the inputs is presented as below.

1. The accuracy of the model is represented as an interval at $C = (1-\alpha)100\%$ confidence level, whose value is $90\%, 95\%$ and 99% typically corresponding to its significance level α as $0.10, 0.05$ and 0.01, respectively.
2. The T can be specified by either percentage as $T\%$ or the number of data points, which directly means that the last T points become the testing data. The number of data in total, training and testing will be shown when clicking the button "Show detail".
3. The value of $g_{\max}$ is used in this work as a stopping criterion of the GA.
4. The population size S represents the number of chromosomes in each generation.
5. In selection process, the first s chromosomes ranked from the best fitness value to the worst one can survive to the next generation This means a number of chromosomes will be randomly generated to fulfill the population size at that generation.
6. In mutation, since our chromosome representation is a binary string of genes, we randomly select m genes and flip their values from 1s to 0s and vice versa.
7. The rates R_c and R_m are nonnegative integers in the range $[0, 100]$. The survey [2] informed that the crossover rate R_c is typically valued in the range $[60, 100]$. While, the appropriate value of R_m for a given optimization problem is an open research issue.
   - For each parents, we generate a random integer R between 0 and 100 and perform a single point crossover to the parents that produce offspring if $R < R_c$. Otherwise, such execution is none.
   - For each chromosome, we generate R similarly as above. If $R < R_m$, the chromosome is mutated by flipping on its genes according to the number of mutation points m. Otherwise, such execution is none.

Step 4: Choose either one of the forecast accuracy measures MAPE, RMSE or MAE, whose descriptions can be seen in Table 1 of [5] and on pages 5–6 of [12]. These computations are written as Eqs. (9)–(11).

Step 5: Click the button "Start". The forecasting results will be displayed at GARI's windows in a while. The running time depends on the size of input data and the $g_{\max}$ setting. However, the users can stop running anytime by clicking button "Stop".



Input File :

Column name to forecast :

Confidence level (%) : 95

Amount of testing data : 1 ○ Percentage ○ Number of data points

Max generation : 200 (default: 200)

Population size : 10 (default: 10)

Number of survivors : 5 (default: 5)

Number of mutation points : 1 (default: 1)

Mutation rate (%) : 30 (default: 30)

Crossover rate (%) : 80 (default: 80)

Show detail : Disable (default: Disable)

Forecast accuracy measures

○ MAPE

○ RMSE

○ MAE

FAST GA : Disable

Maximum acceptable error :

	NoData	NoData	NoData	NoData	NoData	NoData	NoData
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							

10 rows x 7 columns

Current Progress: 0.0%

Fig. 3. Main window of GARI.

4 Results

The currency exchange rate between *Thai Baht (THB)* and *Chinese Yuan (CNY)* was collected from Bank of Thailand during 2020–2022 in two data sets:

1. The daily exchange rate from April, 1, 2021 to April, 14, 2022 (one year period) without the weekend consists of overall 271 observations.
2. The daily exchange rate from April, 1, 2020 to April, 14, 2022 (two years period) without the weekend consists of overall 532 observations.

The last 30 days of those data sets, i.e., March, 4, 2022 – April, 14, 2022, are used for testing validity of the models. By using auto ARIMA in Python, the best model for the first data set is ARIMA$(1, 1, 2)$. For the second one, it gives the model ARIMA$(1, 1, 1)$. The performance measurement of the ARIMA models obtained from auto ARIMA and the subset ARIMA models obtained from our program GARI can be seen in Table 2. This table reports that the GARI provides less values of MAPE, MAE and RMSE compared to auto ARIMA.

For the two years period, we accomplish the same model ARIMA$(4, 1, \{1, 5\})$ based on those error measures. Its prediction equation as shown in Table 2 could be transformed to the one in terms of the original time series $\{X_t : t \in \mathbb{Z}\}$ as

$$\begin{aligned}
X_t - X_{t-1} &= 0.7826(X_{t-1} - X_{t-2}) + 0.0480(X_{t-2} - X_{t-3}) - 0.0105(X_{t-3} - X_{t-4}) \\
&\quad + 0.1752(X_{t-4} - X_{t-5}) + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5} \\
X_t &= (0.7826 + 1)X_{t-1} + (0.0480 - 0.7826)X_{t-2} + (-0.0105 - 0.0480)X_{t-3} \\
&\quad + (0.1752 + 0.0105)X_{t-4} - 0.1752X_{t-5} + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5},
\end{aligned}$$

Table 2. Forecasting results from auto ARIMA and GARI for the daily CYN/THB exchange rate in two periods of data collection.

Data collection period	Criteria	Auto ARIMA	GARI	The best model from GARI
1 Year	MAPE	1.8381%	1.6937%	ARIMA$(5,1,\{2,4,5\})$ has the prediction equation $Z_t = 0.1055Z_{t-1} - 0.3152Z_{t-2} + 0.0371Z_{t-3}$ $+ 0.2075Z_{t-4} + 0.4990Z_{t-5} + \epsilon_t$ $+ 0.4526\epsilon_{t-2} - 0.0124\epsilon_{t-4} - 0.6550\epsilon_{t-5}.$
	MAE	0.096705	0.089106	
	RMSE	0.099473	0.091623	ARIMA$(3,1,\{2,3\})$ has the prediction equation $Z_t = 0.1312Z_{t-1} + 0.5070Z_{t-2} + 0.3562Z_{t-3} + \epsilon_t$ $- 0.4629\epsilon_{t-2} - 0.5229\epsilon_{t-3}.$
2 Years	MAPE	1.9276%	1.2180%	ARIMA$(4,1,\{1,5\})$ has the prediction equation $Z_t = 0.7826Z_{t-1} + 0.0480Z_{t-2} - 0.0105Z_{t-3}$ $+ 0.1752Z_{t-4} + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5}.$
	MAE	0.101412	0.064061	
	RMSE	0.104062	0.066674	

which implies

$$\hat{X}_t = 1.7826X_{t-1} - 0.7346X_{t-2} - 0.0585X_{t-3} + 0.1857X_{t-4} - 0.1752X_{t-5} + \epsilon_t - 0.5887\epsilon_{t-1} - 0.1073\epsilon_{t-5}. \tag{12}$$

Furthermore, the illustration in Fig. 4 shows the time series plot of the CNY/THB exchange rate data. Within the scope of prediction, the values are clarified in Table 3. We observe that the forecast values lie on the range of 95% confidence interval from March, 9, 2022 onwards.

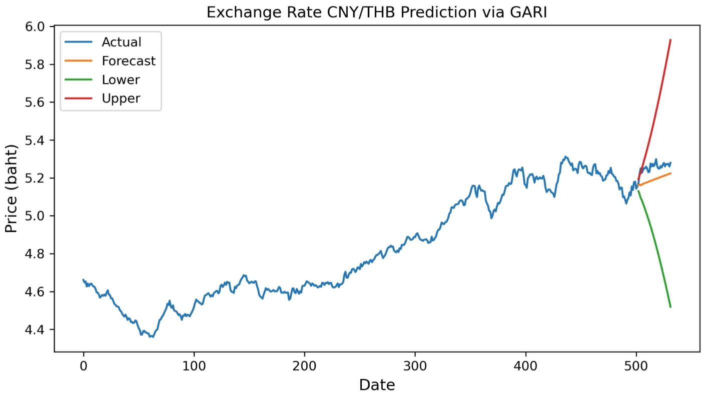


Fig. 4. Actual and forecast values of the daily CYN/THB exchange rate collected in two years.

At the end of this section, we additionally record the errors from ETS model (via auto ETS in Python) and LSTM algorithm in Table 4. The GARI can

Table 3. Actual and forecast values with its lower and upper values at 95% confidence level obtained by the best model ARIMA$(4, 1, \{1, 5\})$ with its forecasting equation as (12) for the daily CYN/THB exchange rate collected in two years.

Date	Actual	Forecast	95% confidence interval	
			Lower	Upper
4-Mar-22	5.1716	5.161101256	5.129606165	5.192596346
7-Mar-22	5.2197	5.164041509	5.114992439	5.21309058
8-Mar-22	5.2503	5.160236278	5.09441576	5.226056796
9-Mar-22	5.2263	5.164947343	5.08302293	5.246871756
10-Mar-22	5.2384	5.167678293	5.06709499	5.268261597
11-Mar-22	5.2518	5.170596575	5.051533326	5.289659825
14-Mar-22	5.2506	5.172295425	5.034476495	5.310114356
15-Mar-22	5.2594	5.174561539	5.017711585	5.331411492
16-Mar-22	5.2475	5.176864204	5.000304475	5.353423932
17-Mar-22	5.2299	5.179268321	4.982419653	5.37611699
18-Mar-22	5.233	5.18153403	4.963841462	5.399226598
21-Mar-22	5.2762	5.183795279	4.944746162	5.422844396
22-Mar-22	5.26	5.18605173	4.925113781	5.446989679
23-Mar-22	5.274	5.188323436	4.904970204	5.471676668
24-Mar-22	5.2618	5.190582652	4.88429728	5.496868023
25-Mar-22	5.2744	5.192832095	4.863113847	5.522550342
28-Mar-22	5.299	5.19507229	4.841431219	5.548713361
29-Mar-22	5.2605	5.197307582	4.81926417	5.575350994
30-Mar-22	5.2502	5.199536508	4.796621393	5.602451623
31-Mar-22	5.2486	5.201758602	4.773513265	5.630003939
1-Apr-22	5.2627	5.203973475	4.749949807	5.657997143
4-Apr-22	5.2548	5.206181578	4.725941451	5.686421705
5-Apr-22	5.2689	5.208382992	4.701497816	5.715268168
6-Apr-22	5.2791	5.210577727	4.676628179	5.744527275
7-Apr-22	5.2605	5.212765718	4.651341437	5.77419
8-Apr-22	5.274	5.214946998	4.62564627	5.804247725
11-Apr-22	5.2714	5.217121598	4.599551037	5.83469216
12-Apr-22	5.2741	5.219289551	4.573063796	5.865515306
13-Apr-22	5.2602	5.22145087	4.546192309	5.896709432
14 Apr-22	5.2796	5.223605574	4.518944071	5.928267077

achieve more accurate than the ETS model but cannot achieve when comparing to LSTM. Since the GARI is developed on the basis of the traditional model ARIMA, it is not possible to reach superior to deep leaning-based algorithm LSTM, which was also confirmed in the articles [6,18] and [22].

Table 4. The MAPEs, MAEs and RMSEs from GARI compared to auto ETS and LSTM for the daily CYN/THB exchange rate in two periods of data collection.

Data collection period	Criteria	GARI	Auto ETS	LSTM
1 Year	MAPE	1.6937%	1.9240%	0.7309%
	MAE	0.089106	0.101218	0.038439
	RMSE	0.091623	0.103869	0.042586
2 Years	MAPE	1.2180%	1.8850%	0.6097%
	MAE	0.064061	0.099165	0.032071
	RMSE	0.066674	0.101783	0.037495

5 Conclusion

This paper utilized GA-based subset ARIMA model with combining best sides of genetic algorithm and ARIMA model to improve prediction accuracy of the single ARIMA model. The ARIMA is a popular model to analyze stationary and non-stationary univariate time series data. The use of the genetic algorithm is to determine which ARIMA model is the best. In addition, we developed a program named GARI to predict time series data on the basis of GA-ARIMA procedure. Our program was applied to forecast the daily exchange rate for the Thai baht against the Chinese yuan reached higher forecast accuracy compared to using auto ARIMA with Python based on the statistical ARIMA model.

References

1. Al-Douri, Y.K., Hamodi, H., Lundberg, J.: Time series forecasting using a two-level multi-objective genetic algorithm: a case study of maintenance cost data for tunnel fans. Algorithms **11**(8), 123 (2018)
2. Boussaïd, I., Lepagnot, J., Siarry, P.: A survey on optimization metaheuristics. Inf. Sci. **237**, 82–117 (2013)
3. Bowornchockchai, K.: Forecasting exchange rate between Thai Baht and the US dollar using time series analysis. Int. J. Math. Comput. Sci. **8**(8), 1186–1191 (2016)
4. Dautel, A.J., Härdle, W.K., Lessmann, S., Seow, H.V.: Forex exchange rate forecasting using deep recurrent neural networks. Digit. Financ. **2**(1), 69–96 (2020)
5. Dzikevičius, A., Šaranda, S.: Smoothing techniques for market fluctuation signals. Bus. Theory Pract. **12**(1), 63–74 (2011)
6. Elsaraiti, M., Merabet, A.: A comparative analysis of the ARIMA and LSTM predictive models and their effectiveness for predicting wind speed. Energies **14**(20), 6782 (2021)
7. Ervural, B.C., Beyca, O.F., Zaim, S.: Model estimation of ARMA using genetic algorithms: a case study of forecasting natural gas consumption. Procedia Soc. Behav. Sci. **235**, 537–545 (2016)
8. Fabozzi, F.J., Focardi, S.M., Rachev, S.T., Arshanapalli, B.G.: The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications. Wiley, Hoboken (2014)

9. George, E., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control. Wiley, Hoboken (1970)
10. Hossain, Z., Rahman, A., Hossain, M., Karami, J.H.: Over-differencing and forecasting with non-stationary time series data. Dhaka Univ. J. Sci. **67**(1), 21–26 (2019)
11. Hunt, K.M., Matthews, G.R., Pappenberger, F., Prudhomme, C.: Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western united states. Hydrol. Earth Syst. Sci. Discuss. **26**, 5449–5472 (2022)
12. Jierula, A., Wang, S., Oh, T.M., Wang, P.: Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. Appl. Sci. **11**(5), 2314 (2021)
13. Kamruzzaman, J., Sarker, R.A.: Forecasting of currency exchange rates using ANN: a case study. In: Proceedings of the 2003 International Conference on Neural Networks and Signal Processing, vol. 1, pp. 793–797. IEEE (2003)
14. Lee, S., Fambro, D.B.: Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transp. Res. Rec. **1678**(1), 179–188 (1999)
15. Mahjoub, S., Chrifi-Alaoui, L., Marhic, B., Delahoche, L.: Predicting energy consumption using LSTM, multi-layer GRU and drop-GRU neural networks. Sensors **22**(11), 4062 (2022)
16. Maria, F.C., Eva, D.: Exchange-rates forecasting: exponential smoothing techniques and ARIMA models. Ann. Faculty Econ. **1**(1), 499–508 (2011)
17. Moein, S., et al.: Inefficiency of sir models in forecasting Covid-19 epidemic: a case study of Isfahan. Sci. Rep. **11**(1), 1–9 (2021)
18. Muncharaz, J.O.: Comparing classic time series models and the LSTM recurrent neural network: an application to s&p 500 stocks. Financ. Markets Valuation **6**(2), 137–148 (2020)
19. Ong, C.S., Huang, J.J., Tzeng, G.H.: Model identification of ARIMA family using genetic algorithms. Appl. Math. Comput. **164**(3), 885–912 (2005)
20. Paul, J.C., Hoque, M.S., Rahman, M.M.: Selection of best ARIMA model for forecasting average daily share price index of pharmaceutical companies in Bangladesh: a case study on Square Pharmaceutical Ltd. Glob. J. Manag. Bus. Res. **13**, 14–25 (2013)
21. Qu, Y., Zhao, X.: Application of LSTM neural network in forecasting foreign exchange price. In: Journal of Physics: Conference Series, vol. 1237, p. 042036. IOP Publishing (2019)
22. Siami-Namini, S., Tavakoli, N., Namin, A.S.: A comparison of ARIMA and LSTM in forecasting time series. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1394–1401. IEEE (2018)
23. Swaraj, A., Verma, K., Kaur, A., Singh, G., Kumar, A., de Sales, L.M.: Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India. J. Biomed. Inform. **121**, 103887 (2021)
24. Tepdang, S., Ponprasert, R.: Forecast of changes in exchange rate between Thai Baht and US dollar using data mining technique. SNRU J. Sci. Technol. **12**(3), 213–221 (2020)
25. Tlegenova, D.: Forecasting exchange rates using time series analysis: the sample of the currency of Kazakhstan. arXiv preprint arXiv:1508.07534 (2015)
26. Whittle, P.: Hypothesis Testing in Time Series Analysis, vol. 4. Almqvist & Wiksells boktr. (1951)

27. Wijesinghe, S.: Time series forecasting: analysis of LSTM neural networks to predict exchange rates of currencies. Instrumentation **7**(4), 25 (2020)
28. Yıldırım, D.C., Toroslu, I.H., Fiore, U.: Forecasting directional movement of forex data using LSTM with technical and macroeconomic indicators. Financ. Innov. **7**(1), 1–36 (2021)