25th June, 2024

TIME SERIES ANALYSIS: FINAL REPORT

Unveiling the Dynamics Underlying the Cocoa Bean Future Markets

Minen Sun,*†‡ Ching Wang,*†‡ and Kuan-Hsun Wu*†‡

†Department of Statistics, National Chengchi University
‡The authors contributed equally to this work and are listed in alphabetical order.
*Corresponding authors. Email: 109304018@g.nccu.edu.tw; 109304015@g.nccu.edu.tw; 110304015@g.nccu.edu.tw

Abstract

Cocoa farmers in the futures market cannot compete with formidable, international financial groups. To mitigate this economic disadvantage, this article aims to demystify the seasonal patterns in cocoa futures trading volume and provide beneficial insights for smallholders in optimizing trading strategies in the agricultural market. Using weekly data from January 2000 to December 2019, we consider both single and multiple seasonalities and analyze them using seasonal autoregressive integrated moving average (SARIMA) methods. Traditional diagnostic approaches are employed to evaluate model performance. To address challenges posed by fitting high-lag ARIMA and SARIMA models, we propose a specific criterion for the Ljung-Box test, which emphasizes the resemblance of residuals to white noise at high lags. Model selection is based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC). The results indicate significant seasonal patterns, which can help smallholders formulate optimal trading strategies. This research provides a robust statistical framework for understanding market dynamics and offers practical support to smallholders and nonprofit organizations in the agricultural sector.

Keywords: Seasonality; cocoa, SARIMA, time series forecast, futures market, Ljung-Box test.

1. Introduction

Cocoa, also known as the "food of the gods," is an important crop that originated in South America (Montagna et al., 2019). The Maya people considered cocoa essential as water, sun, and other necessities. During the age of discovery, the Spanish saw cocoa beans as "black gold," and they eventually introduced it to the world (Ozturk and Young, 2017). Today, cocoa beans still play a crucial role in the economic growth of Africa and South America. However, smallholders, the cocoa beans producers, are lack of participation in the price decision process since it is a privilege of multinational enterprises in the futures market. Nowadays, local farmers have formed agricultural cooperatives to gain more power in the price bargaining of cocoa bean markets. Moreover, some nonprofit organizations (NPOs) like Fair Trade also aim to break the imbalance between smallholders and multinational enterprises by expediting the retrieval of the capability in bargaining for smallholders (Dragusanu et al., 2014). In order to assist the NPOs and the smallholders in their strategies of negotiation with formidable and rich enterprises, it is of great significance to comprehend the dynamics of cocoa bean futures markets. In this work, we aim to achieve the understanding by applying statistical models. Also, some possible difficulties might be complicated seasonalities. Beg et al. (2017) revealed that cocoa tree takes 3 years to grow and 8-9 years for reaching its full-grown period with two production peaks per year. Furthermore, the contracts of cocoa futures will be due

March, May, July, September or December each year, bringing not only seasonalities to our data, but also complexity to our analysis. In fact, Sarfo and Geman (2012) recognized seasonality as a crucial factor in risk management, especially in agricultural commodities. Therefore, it is necessary to search for statistical models capable of accommodating intricate and multiple seasonalities. To conclude, since the cocoa markets reflect the asymmetry between the force of supply (i.e., the smallholders) and demand (i.e, the international tycoons,) it is of abundant benefit for NPOs and smallholders to understand the market operation. By clarifying the motion in cocoa futures market via data, the two major objectives of this work is to (i) provide an explainable, comprehensible statistical modeling for data and (ii) bring insights and assistance to the strategy formation for cocoa farmers.

The remainder is organized as follows. In Section 2, we introduce how the data is generated and collected in futures markets and determine the time scope used in data. Also, some preliminary exploration in data is offered. In Section 3, some models frequently applied to deal with seasonalities are reviewed. In Section 4, we display the details of order selection with autocorrelation function (ACF) and partial autocorrelation function (PACF). In Section 5, diagnostics of models are provided. Since our models are fitted with high lags, most existing criteria in selecting lags in Ljung-Box test are not applicable. Hence, we propose a criterion tailored for our data. In Section 6, we apply information criteria such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) to choose the optimal model. Subsequent to model selection, the results of prediction of selected models are included in Section 7. Moreover, mean squared error (MSE) is utilized to evaluate the performance in prediction. Finally, we conclude the highlights in this article and plan accordingly for smallholders in Section 8. Some plots and diagnostic results are placed in the Appendix in order to not overwhelm or mislead the readers from grasping this article.

2. Data Description and Preliminary Analysis

2.1 Data Source and Structure

2.1.1 Data Source

Intercontinental Exchange (ICE) provides a platform of global exchanges, including trading cocoa futures. Yahoo Finance, which collected data from ICE, offers weekly data of cocoa futures[1], it comprises `Open price` (USD), `Close price` (USD), `Highest price` (USD), `Lowest price` (USD) and `Volume` (metric ton). In this project, we aim to capture the statistical pattern of weekly volume of cocoa futures. The time series data ranges from January 2000 to December 2019. To evaluate the performance of selected models, the last 52 weeks have been removed for the sake of validation. The training data covers from January 3, 2000 to December 3, 2018, and the length of the time series is 989.

2.1.2 Data Processing

Figure (1) is the time series plot of X_t. Apparently, the local variance is growing larger gradually over time t. Hence, X_t performs heteroscedasticity and is a nonstationary process. To achieve the equal variance assumption of a stationary process, we decide to implement log transform. Since the data contains a few zero values, the proper transformation for X_t is $\log(X_t + 1)$.

Figure (2) is the time series plot of $\log(X_t + 1)$ and the unstable variance shown in Figure (1) has been tremendously improved by the logarithmic transformation. Figure (3) displays the autocorrelation function $\gamma(h)$ of the transformed time series for different h. Also, the ACF plot possesses periodicity of 52. More specifically, the ACF exhibits the circular behavior of increasing before $52s$ and decreasing after $52s$, for $s \in \{1, 2, \dots\}$. Also, we observe that the ACF does not decay slowly. Figure (4) is the estimated partial autocorrelation function $\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}$ of the transformed time series

[1] https://finance.yahoo.com/quote/CC%3DF/history/

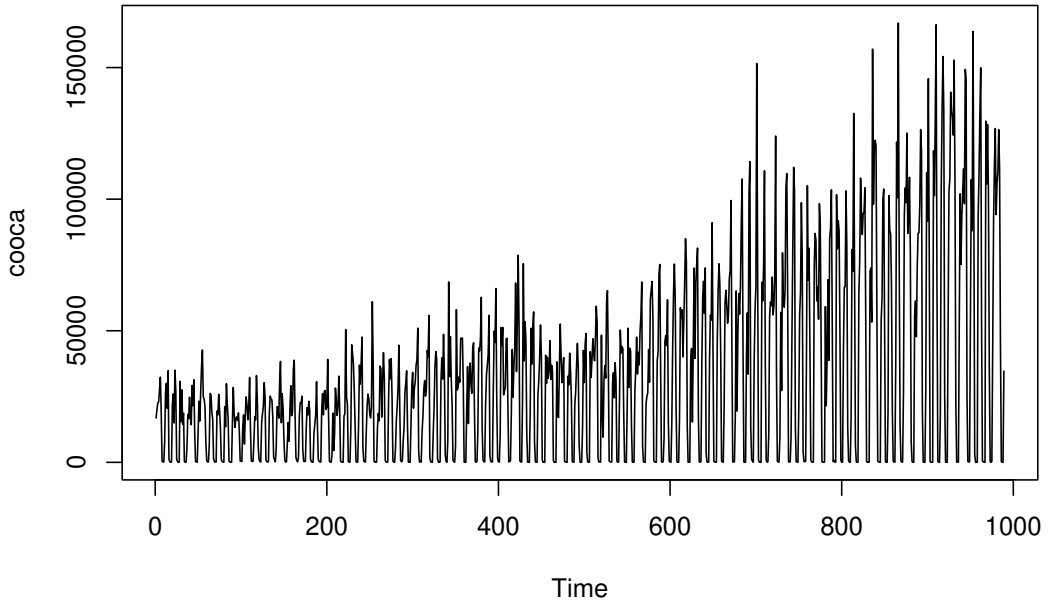

Figure 1. Time Series Plot of X_t

for different h. The PACF after lag 52 has mostly dropped to an insignificant scale. Also, for the sake of readers' convenience, the significant lags of ACF and PACF (Figure 3 and 4) are listed in Table (1).

Combining Figure 2 and ACF plot, we believe the transformed time series is stationary. However, to add credibility and assure the stationarity in data, we conduct augmented Dickley-Fuller (ADF) test (Dickey and Fuller, 1979; Said and Dickey, 1984). The null and alternate hypothesis are stated as follow:

$$H_0 : \text{ There exists an unit root in } \log(X_t + 1);$$
$$H_1 : \text{ There does not exist an unit root in } \log(X_t + 1).$$

Setting the significance level to $\alpha = 0.05$, the test reported p-value far smaller than 0.05. Therefore, there is a significant statistical evidence to support that $\log(X_t + 1)$ is a stationary time series. To conclude, considering the visual proof from Figure 2 and ACF plot (Figure 3) and the ADF test result, we strongly believe that the transformed time series is stationary and does not require first order differencing.

Table 1. Significant Lags of ACF and PACF

	Significant Lags
ACF (See Figure 3)	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 14, 15, 16, 18, 26, 37, 38, 39, 42, 43, 44, 46, 47, 48, 49, 51, 52 . . .
PACF (See Figure 4)	1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 27, 31, 35, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52 . . .

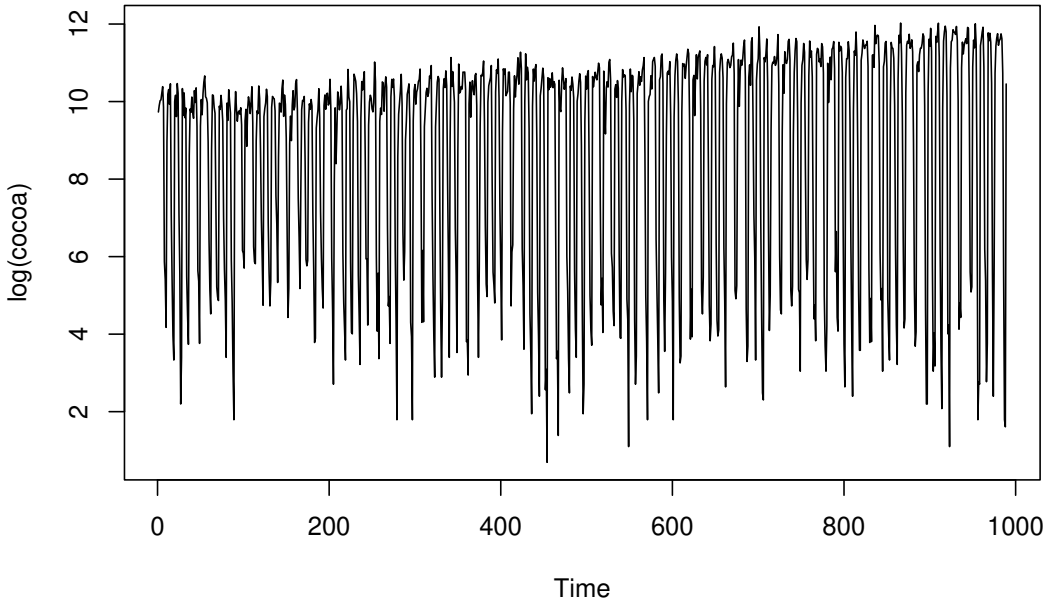

Figure 2. Time Series Plot of $\log(X_t + 1)$.

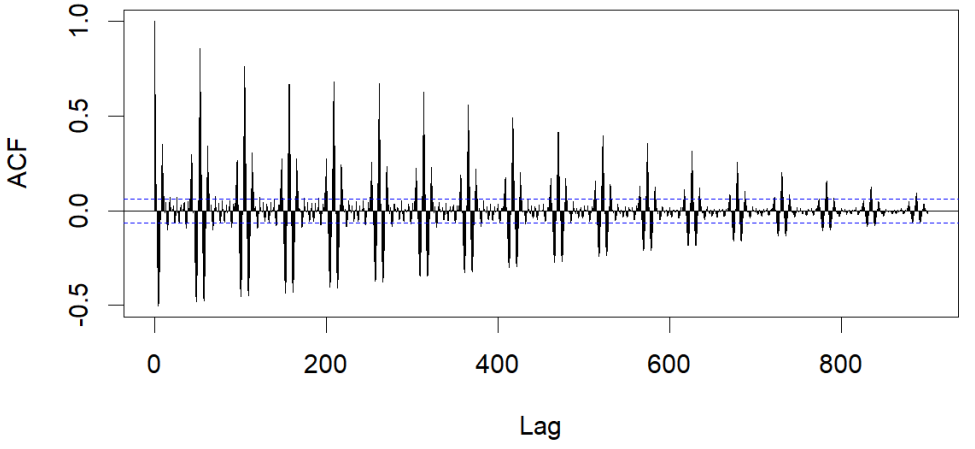

Figure 3. Autocorrelation Plot of $\log(X_t + 1)$.

On the other hand, the behavior of ACF displays periodicity with lag 52. That is, the ACF

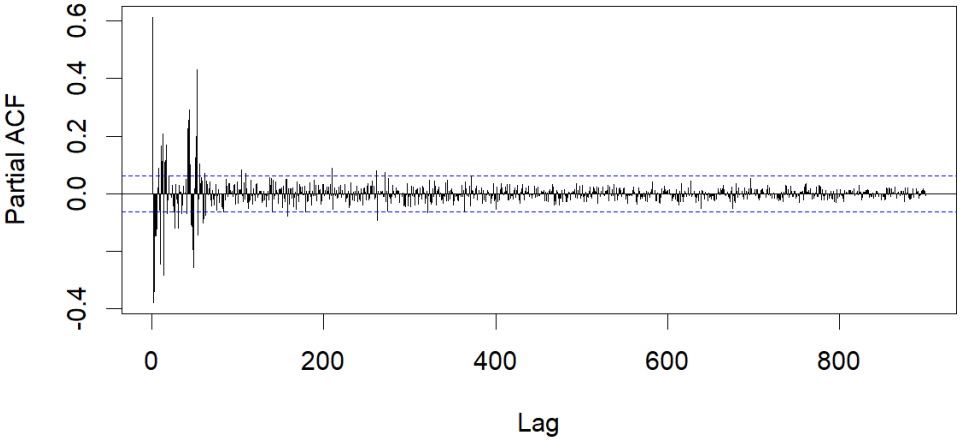

Figure 4. Partial Autocorrelation Plot of $\log(X_t + 1)$.

reaches its peak at the multiples of 52. Since 52 weeks form a year, this echoes with our background knowledge. Although the ACF plot decays fast from a seasonal perspective and the time series plot behave stably, a seasonal unit root test can be applied to enhance our belief in the seasonal stationarity in data. Osborn et al. (1988) developed a test to help validate the seasonal stationarity. The test encompasses hypotheses of the following form:

$$H_0: \text{ There exists a seasonal unit root in } \log(X_t + 1);$$
$$H_1: \text{ There does not exist a seasonal unit root in } \log(X_t + 1),$$

where the seasonal period is set to 52. With significance level $\alpha = 0.05$, we reject H_0 owing to a p-value smaller than 0.05. Hence, with the evidence shown in time series plot, ACF plot and the seasonal unit root test, we confirm that there is no need in conducting first order seasonal differencing in data.

To conclude, we are determined neither implementing first order differencing nor the seasonal one on account of the behavior shown in time series plot, ACF plot and the statistical tests of the existence of unit roots. Consequently, we consider the data to be stationary and ready for further analysis.

3. Methodology

3.1 Strategies for Selecting Models

Cocoa bean volume changes due to external factors. For example, we have mentioned that there are two main production times each year (Beg et al., 2017), and due dates of cocoa future contracts distributed at March, May, July, September and December for each year. The real world consists of interwoven behaviors and causes and therefore beckons complex seasonal patterns in the observed data. Hence, we feel the necessity of time series models capable of contain such patterns.

Several studies have used time series methods to predict financial derivatives or prices in agricultural products. Ouyang et al. (2019) applied both linear and non-linear methods to forecast global agricultural commodity futures prices. Assis et al. (2010) using ARIMA model to forecast the cocoa

bean price from January 1992 to December 2006. Figure (3) suggests the time series may have a seasonal factor, s, of 52. After testing different factors, we also look at seasonal factors of 13 and 26, which represent one quarter and half a year. In this project, we are going to attempt in two different strategies:

- **Strategy 1:** Considering a single seasonality s.
  - **ARMA**(p, q): Strong seasonal factor s can be placed at either p or q and the left one can be used to handle other significant non-seasonal lags.
  - **SARIMA**$(p, d, q) \times (P, D, Q)_S$: With seasonal factor handled by S, the significant seasonal lags can be treated with P and Q and significant non-seasonal lags can be treated with p and q.
- **Strategy 2:** Considering two or multiple seasonalities $(s_1, s_2, \ldots, s_m)$.
  - **SARIMA**$(p, d, q) \times (P, D, Q)_S$: With a more dominant seasonal factor handled by S, the other less significant seasonal factor can be placed in p and q, the non-seasonal part.
  - There are some other models for accommodating multiple seasonalities. For instance, De Livera et al. (2011) proposed a versatile model named *TBATS* which is the abbreviation of **T**rigonometric functions, **B**ox-Cox transformation, **A**RMA errors, **T**rend and **S**easonal components.

3.2 SARIMA and ARIMA

Box and Jenkins (1970) developed seasonal autoregressive integrated average moving model (SARIMA) which incorporates the conventional autoregression and moving average model (ARMA) with seasonal factors. Since SARIMA is capable of capturing potential trends in the data by accommodating both seasonal factors and non-seasonal factors, it is effective for understanding and predicting future values in time series data that exhibit seasonality and other patterns over time. SARIMA is of the form

$$\Phi_S(B^S)\phi(B)\nabla_S^D\nabla^d \log(X_t + 1) = \Theta_S(B^S)\theta(B)Z_t, \tag{1}$$

where

- S is the season period to be determined;
- ∇^d is the non-seasonal difference component;
- ∇_S^D is the seasonal difference component;
- $\Phi_S(B^S) = 1 - \Phi_1 B^S - 1 - \Phi_2 B^{2\cdot S} - \cdots - \Phi_P B^{P\cdot S}$ is the seasonal AR term with order P to be determined;
- $\phi_S(B) = 1 - \phi_1 B - 1 - \phi_2 B^2 - \cdots - \phi_p B^p$ is the non-seasonal AR term with order p to be determined;
- $\Theta_S(B^S) = 1 + \Theta_1 B^S - 1 + \Theta_2 B^{2\cdot S} + \cdots - \Theta_Q B^{Q\cdot S}$ is the seasonal MA term with order Q to be determined;
- $\theta_S(B) = 1 + \theta_1 B - 1 + \theta_2 B^2 - \cdots + \theta_p B^q$ is the non-seasonal MA term with order q to be determined;
- $Z_t \sim N(0, \sigma^2)$ is Gaussian white noise.

A model of form (1) is denoted by **SARIMA**$(p, d, q) \times (P, D, Q)_S$ hereafter. ARIMA is a special case where the seasonal terms are not taken into consideration which we denote it by **ARIMA**(p, d, q).

3.3 Other Models

We encountered some technical problems while fitting TBATS. The only existing R package for fitting TBATS is `forecast` (Hyndman & Khandakar, 2008), which includes function `tbats` for parameter estimation. However, the ARMA part of the model form is not specified from users but suggested by function `auto.arima`. Users of function `tbats` can only determine the significant

seasonal periods $(s_1, s_2, \ldots, s_m)$. Hence, the information extracted from ACF plot and PACF plot cannot be fully utilized. The authors have tried some combinations of seasonal periods; nonetheless, the model performance is not comparable with ARMA models fitted with selected lags. We therefore will not include the model fitting results of TBATS.

4. Model Building

To fit an SARIMA or ARIMA model, we must select (p, q), (P, Q) and S. For a ARIMA model, the behavior of theoretical ACF and PACF function, $\rho(h)$ and ϕ_{hh}, can be classified as *tailing off* or *cutting off at lag h*, where the former describes that a function decreases as h increases in a gradual manner and the latter represents a function dropping to 0 immediately after some specific lag h. As for SARIMA, the behavior of theoretical ACF and PACF function can be specify in the same way with seasonal taken into account.

Hence, it is understandable to study the sample ACF and PACF plot (Figure 3 and 4) and carefully examine the behavior shown in them. Upon successful categorization of sample ACF and PACF into tailing off or cutting off at lag h, we can match them with Table 3.1 and 3.3 in Section 3 of Shumway and Stoffer (2000). The underlying reason is we assume

$$\widehat{\rho}(h) \to \rho(h) \text{ and } \widehat{\phi}_{hh} \to \phi_{hh} \tag{2}$$

as $n \to \infty$, where $\rho(h)$ and ϕ_{hh} is the unknown, true ACF and PACF function of the data. The scheme of matching is to seek an appropriate set of (p, d, q, P, D, Q, S), which implicitly decides $\rho(h)$ and ϕ_{hh}, to allow (2) is held.

Nevertheless, there always exists irreducible noise in data such that the the behaviors of $\widehat{\rho}(h)$ and $\widehat{\phi}_{hh}$ cannot be easily classified as tailing off or cutting off. Consequently, we consider the two following scenarios in order to not miss out some possible cases:

- **Scenario 1**: We can regard the behavior of ACF and PACF (Figure 3,4) as tailing off. Hence, we consider **ARMA**(p, q) models with p and q shown in Table (1).
- **Scenario 2**: We have observed the seasonality of 52 in Figure (3). From a seasonal perspective, the ACF (Figure 3) and the PACF (Figure 4) can be viewed as tailing off and cutting off at lag $h = 52$, respectively. Therefore, we are taking **SARIMA**$(p, 0, q) \times (1, 0, 0)_{52}$ into account.

The computation capability allows us to extend more cases and some successful ones are included. Also, to put an eye on the parsimony principle of statistics, we do not select non-seasonal lags higher than 26. Table (2) displays all the choices made after comprehensive contemplation.

5. Diagnostics

In this section, we perform some basic diagnostics for each fitted model to guarantee the the fitted models have properly captured data patterns. The entire diagnostics procedure aims to examine whether the residuals behave as theory implies. If a model does not pass the diagnostic checks, this reflects that there are some omitted pattern in data which is not yet included in the fitted model. We can thus plan to improve with different modeling strategies suggested from the flaws in residuals.

Before a model is put into use, we require it to pass some of the necessary diagnostic checks, like significance in the coefficients of highest orders (Section 5.1) and Ljung-Box test (Section 5.2). The other diagnostics such as visual checks in plots have function overlapping with Ljung-Box test or can be solved with remedial approaches. For example, the residuals of a model can be spotted with serial correlation via Ljung-Box test or visual inspection in time series plot of residuals. Or, the violation in normality can bring about unstable estimation of confidence or prediction intervals, which can be relieved using bootstrapping. Therefore, a model put into use are expected but not required to pass all of the visual checks.

5.1 Significant Coefficient in Highest Orders

If a model has a coefficient in highest orders insignificantly different from 0, it suggests lowering the selected highest orders since the current ones do not exhibit such importance to be taken into consideration. Hence, we investigate whether or not the models in Table (2) have insignificant coefficient in the highest orders. Take Model 5, for instance, we need to check if $\widehat{\phi}_{13}$, $\widehat{\theta}_1$ and $\widehat{\Phi}_1$ are significantly different from zero. The level of significance α is set to 0.1 since Table (2) is a list of high-order model and the use of such high orders can contribute to volatility in parameter estimation. For a well-fitted model, the all highest orders should be significantly different from zero. Therefore, Model 5 does not pass the check since the estimate of θ_1 is not significantly different from 0. In order to not expose readers to an abundance of plots and thus overwhelm them, we do not include the lengthy procedures of inspecting coefficients for all models. Instead, the results of checking coefficients are summarized with those of Ljung-Box in the column **Passed** of Table (2). Additionally, for detailed model information, please refer to Appendix 1.

5.2 Selection of lags in Ljung-Box Test

The Ljung-Box test (Ljung & Box, 1978) is a portmanteau test for lack of fit which inspects the randomness in residuals. The null and alternate hypothesis of this test are:

$$H_0: \text{ There fitted model is adequate;}$$
$$H_1: \text{ There fitted model is not adequate.}$$

The test statistic is

$$Q = n(n+2)\sum_{h=1}^{H}\frac{\hat{\rho}_e(h)}{n-h} = \frac{n(n+2)}{n-1}\hat{\rho}_e(1) + \frac{n(n+2)}{n-2}\hat{\rho}_e(2) + \cdots + \frac{n(n+2)}{n-H}\hat{\rho}_e(H) \tag{3}$$

Table 2. Possible SARIMA Models and the corresponding result of Ljung-Box test. Model 1-4 are ARMA models. Model 5-12 are SARIMA models with $S = 52$. The columns **p**, **q**, **P** and **Q** represent the AR, MA, seasonal AR and seasonal MA order used in the model, respectively. The column **S** is the period of the model. The column **df** shows the number of used parameters of the model. The column **Lowest Significant Lag** shows the lowest H such that the Ljung-Box test is rejected. If a model shows significance in the coefficient of highest orders and pass the proposed standard of Ljung-Box test, then it receive a ✔ in the column **Passed** and a ✖ otherwise.

Model	p	q	P	Q	S	df	Lowest Significant Lag	Passed	AIC	BIC	MSE
Model 1	13	1	-	-	-	14	5	✖	-	-	-
Model 2	26	1	-	-	-	27	38	✖	-	-	-
Model 3	1	13	-	-	-	14	17	✖	-	-	-
Model 4	1	26	-	-	-	27	39	✖	-	-	-
Model 5	13	1	1	0	52	15	52	✖	-	-	-
Model 6	13	1	0	1	52	15	18	✖	-	-	-
Model 7	13	1	1	1	52	16	61	✔	3111.427	3199.567	2.896954
Model 8	13	0	1	0	52	14	52	✔	3153.561	3231.908	-
Model 9	13	0	0	1	52	14	13	✖	-	-	-
Model 10	13	0	1	1	52	15	61	✔	3113.005	3196.249	2.880642
Model 11	26	0	0	1	52	27	48	✖	-	-	-
Model 12	26	1	0	1	52	28	49	✖	-	-	-
Model 13	26	1	1	0	52	28	61	✖	-	-	-

where n and H denote the length of the time series and the number of lags being tested, respectively. For an fitted $\mathbf{ARMA}(p, q)$ model, $Q \to \chi^2_{H-(p+q)}$ as $n \to \infty$ under H_0 where $p + q$ is the number of used parameters. As a result, we are compelled to always require $H > p + q$ for ensuring that Q is a χ^2 random variable asymptotically while executing the Ljung-Box test. As for SARIMA, the number of used parameters is $p + q + P + Q$ and it is shown that $Q \to \chi^2_{H-(p+q+P+Q)}$ asymptotically (Li, 2003). Therefore, the usage of Ljung-Box test can be naturally extended from ARIMA cases to SARIMA cases.

The optimal choice of H varies from one to another. To name a few, Ljung (1986) proppsed $H = 5$. A far more larger choice of lag is given by Shumway and Stoffer (2000), which recommended the use of $H = 20$; however, the detailed, comprehensive reason for selecting such H is not provided in the context. Some other choices of H incorporate the influence of the length of time series n. For instance, Tsay (2005) advocated $H = \log n$ as the optimal choice. Hyndman and Athanasopoulos (2018) suggested $H = \min(10, \frac{n}{5})$ for non-seasonal data and $H = \min(2S, \frac{n}{5})$ for seasonal data where S is the seasonal period. Also, Hassani and Yeganegi (2020) discovered the optimal H is largely affected by α, the size of test, in addition to length of time series n.

Since $\mathbf{ARMA}(p, q)$ and $\mathbf{SARIMA}(p, d, q) \times (P, D, Q)_S$ are short-memory processes, the criteria discussed above are naturally violated by most high-lags models considered in Table (2). Accordingly, we are in need of inspecting the result of Ljung-Box test for each models to understand whether or not the performance of these models are reasonable.

To find a proper standard of examining the result of Ljung-Box test of models shown in Table (2), we recall that the test statistic Q shown in (3) is constructed based on a selected H and can be viewed as a linear combination of $\widehat{\rho}_e(h)$ with coefficient $\frac{n(n+2)}{n-h}$ for $h \in \{1, 2, \ldots, H\}$. With higher H, Ljung-Box test can take a larger range of lags into consideration. Hence, the test statistic H shows that, implicitly, a Ljung-Box test is based on the users' expectation of how much lags in residuals should be considered. Viewing back the criterion proposed in Shumway and Stoffer (2000), where is applied to test a $\mathbf{MA}(2)$ in the following context, it will be reasonable to explain such rule $H = 20$ implies users' subtle expectation that the residuals of a decent-fitted $\mathbf{MA}(2)$ should bear resemblance to white noise from lag $\mathbf{df}+1$ to at least lag 20. Following such idea, we propose that a valid model should show insignificant result in Ljung-Box test for $H = \mathbf{df} + 20$ where $\mathbf{df}$ is the number of used parameters.

We summarize the diagnostic process done so far. Model 5, Model 7, Model 8, Model10 and Model 13 have passed our proposed criterion of Ljung-Box test. However, Model 5 and Model 13 show insignificance in their coefficients of highest orders and hence are removed from further analysis. The results are conclude in the column **Passed** in Table (2).

5.3 Diagnostic Plots

In this section, we present only the models that successfully passed the essential diagnostic checks. For ones that failed to meet our diagnostic criteria, we invite the interested readers to refer to Appendix 2 for further inspection. Our primary focus is on models that demonstrated satisfactory performance in necessary diagnostics (see the column **Pass** in Table 2). The diagnostic plots consist of four main components. Firstly, the time series plot of residuals is used to check there is no abnormal behavior shown in residuals. For example, serial correlation, heteroscedasticity or extreme outliers can be detected visually via time series plot of residuals. Secondly, the ACF plot of residuals should exhibit no significant lags since residual should approximate white noise. Thirdly, normal probability plots are applied to check normality of residuals. The quantile of residuals is computed and plotted against its expected theoretical quantile under normality. Hence, it is also named quantile-quantile plot (Q-Q plot). A Q-Q plot near to a line suggests normality in residuals while one far from a line indicates the underlying distribution is non-normal. Lastly, the p-value of Ljung Box test for each h

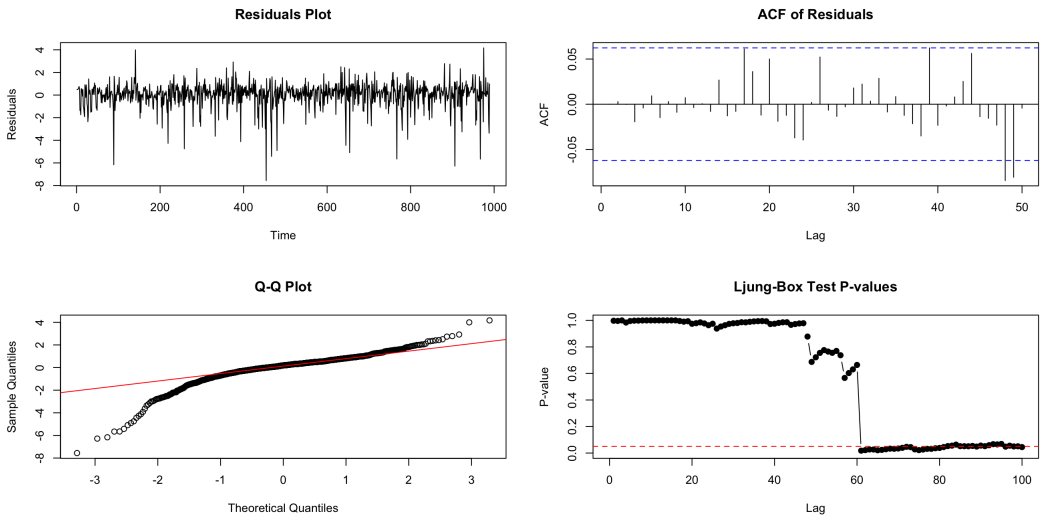

Figure 5. Diagnosis of Model 7, **SARIMA**$(13,0,1)\times(1,0,1)_{52}$

is also plotted for supplementary information to Section (5.2). While we have included the results of normal probability plots for reference, these were not treated as decisive factors in our model selection process since the impact of non-normality can be soothed using bootstrapping.

Since Model 7, Model 8 and Model 10 exhibit similar performance, we select only Model 7 to interpret for the sake of convenience. As shown in Table (2) or Figure (5), the Ljung-box test remains insignificant up to lag 61. This indicates that the model maintains good performance even at high lags. Upon examination of the residuals plot, we observe that the residuals is nearly equal variance, except for some drastic outliers. The ACF plot display only a few significant lags, which can be attributed to the effect of white noise. However, the normal Q-Q plot indicates a departure from normality in the residuals. This observation is expected, since any quantity in financial market of argicultural product is complicated and often observed as non-Gaussian. Also, the Model 7 passed the check in Section 5.1, which indicates the importance to sustain such high orders and can be involved in the next stage of model selection. With the same diagnostic procedure repeated for Model 8 and Model 10, we come to the conclusion that they are held for model selection.

We include more other cases to establish a more comprehensive understanding for readers. Model 5, model 13 both passed our standard in Ljung-Box test, which also indicates their resemblance to white noise even in high lags. Since these two models exhibit similarity, we select Model 5 as our focus of further interpretation. The residuals plot and ACF plot display traits akin to white noise, which echoes with the remarkable performance shown in Ljung-Box test. However, the normal probability plot indicates the residuals are far from normal distribution, especially in the extreme quantiles. Since both of them have insignificant coefficients in their highest orders, these two models are discarded.

6. Model Selection

We have performed some basic diagnostic check in the previous section. For the passed models (see the column **Pass** in Table 2), we utilized Akaike Information Criterion (AIC; Akaike, 1998) and Bayesian Information Criterion (BIC; Schwarz, 1978) to find the optimal models among them. The calculation of AIC is

$$2\mathbf{df} - 2\ln\widehat{L}$$

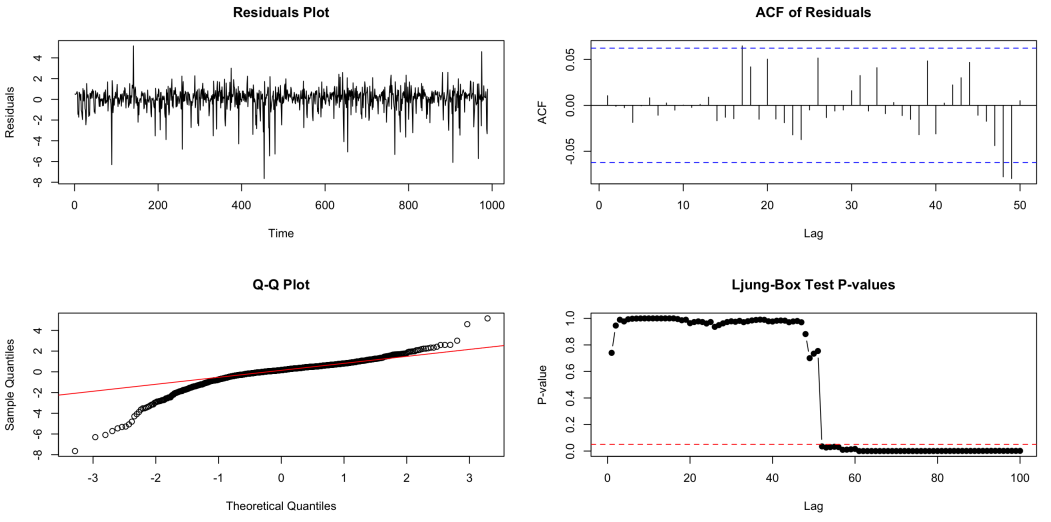

Figure 6. Diagnosis of Model 8, **SARIMA**$(13,0,0)\times(1,0,0)_{52}$

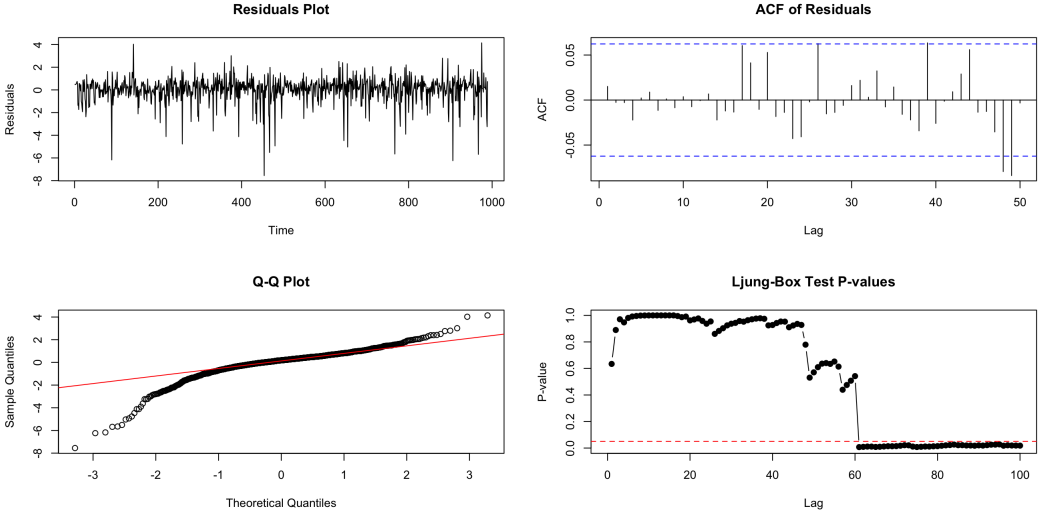

Figure 7. Diagnosis of Model 10, **SARIMA**$(13,0,0)\times(1,0,1)_{52}$

and BIC is

$$(\ln n)\mathbf{df} - 2\ln\widehat{L}$$

where **df** is the number of used parameter in the fitted model and $\ln\widehat{L}$ is the maximized log-likelihood. To minimize AIC and BIC, we must use as less parameter as possible to attain optimal model fitting (i.e. maximize the log-likelihood). Each parameter contributes 2 and $\ln n$ to the total amount of AIC and BIC, respectively. Hence, for close results of AIC, if the difference of AICs of two different models is less than 2, we consider their performance to be equally decent. Likewise, BIC can be applied in a similar manner with 2 replaced with $\ln n \approx 7$.

The AIC and BIC results are shown in Table (2). Since the AIC and BIC of Model 7 and Model

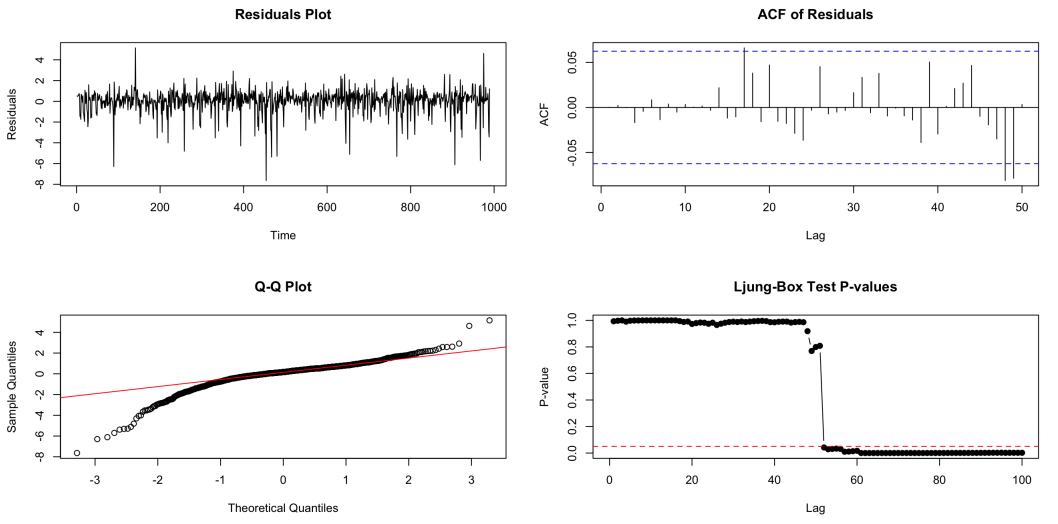

Figure 8. Diagnosis of Model 5, **SARIMA**$(13,0,1)\times(1,0,0)_{52}$

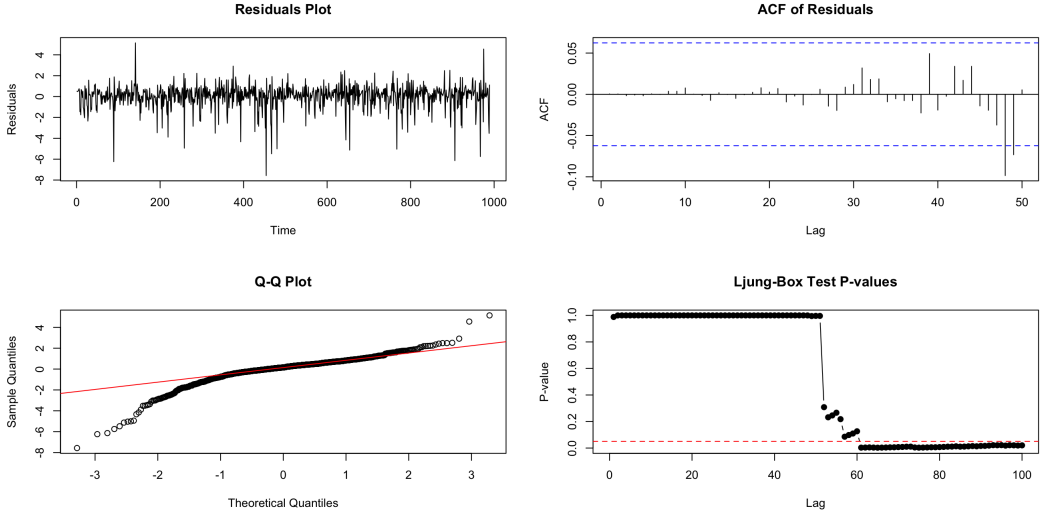

Figure 9. Diagnosis of Model 13, **SARIMA**$(26,0,1)\times(1,0,0)_{52}$

10 are much lower than those of Model 8 and are indistinguishably close in the range discussed (2 and $\ln n \approx 7$), we select Model 7 and Model 10 and further employ them to predict future values.

7. Prediction

In this section, we evaluate the predictive performance models using mean square error (MSE) as our primary metrics. The MSE is defined as

$$\frac{1}{m}\sum_{t=1}^{m}(\widehat{Y}_t - Y_t)^2 = \frac{1}{m}\sum_{t=1}^{m} g(Y_t, \widehat{Y}_t)$$

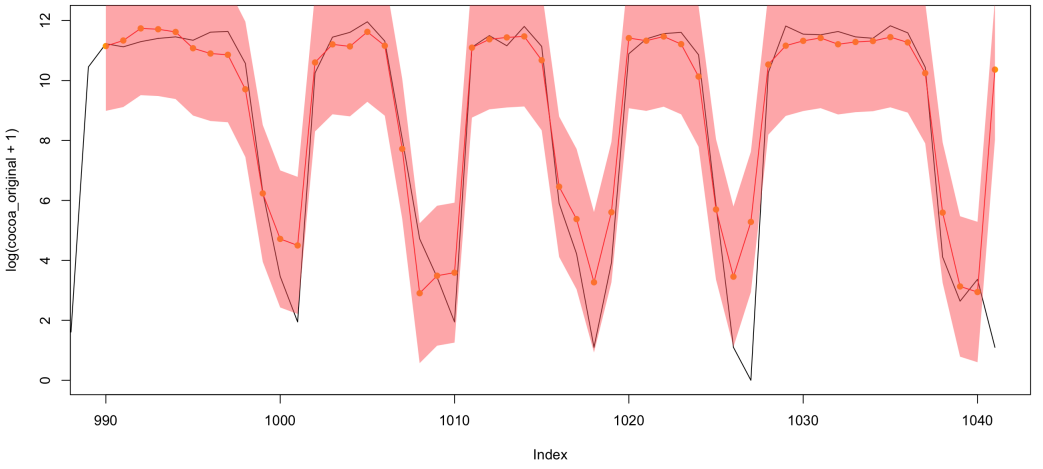

Figure 10. Prediction of **SARIMA**$(13, 0, 1) \times (1, 0, 1)_{52}$

where we define $Y_i = \log X_i + 1$ and $\widehat{Y}_i$ is the corresponding estimate of Y_i and m is the sample size of predetermined test set. Please note that the Y_i here does not involve model fitting. The MSE of Model 7 and Model 10 are 2.896954 and 2.880642, respectively. Since the MSE of Model 7 and Model 10 do not vary largely, we implement a test to verify whether there is a significant difference in their prediction performance. Let $\widehat{Y}^{(7)}$ and $\widehat{Y}^{(10)}$ denote the forecast from Model 7 and 10. We aim to test the following null and alternate hypothesis (Diebold and Mariano, 1995; Harvey et al., 1997):

$$H_0 : g(Y, \widehat{Y}^{(7)}) = g(Y, \widehat{Y}^{(10)});$$
$$H_1 : g(Y, \widehat{Y}^{(7)}) \neq g(Y, \widehat{Y}^{(10)}).$$

The p-value of this test is 0.3937. Depending on this test, we can conclude that the prediction performance of Model 7 and that of 10 do not vary significantly. Consequently, the use of Model 7 or Model 10 cannot be distinguished based on information criteria (Section 6) and statistical test of prediction performance. However, if the user wants to adopt only one model, we recommend Model 10 owing to its slightly better result on MSE, AIC and BIC. Also, Figure (10) and (11) reveal that Model 7 and 10 both have successfully captured the ups and downs in the data despite of some deviation from the true values. Although Model 7 and 10 are capable of recognizing the patterns shown in data, it come to our attention that they fail to reach the downs. Namely, the prediction from Model 7 and 10 do understand the trends but they do not arrive at those low-valued points accurately, or even in a close proximity. The lack in ability of touching the downs may lead to the outliers spotted in Section 5.3 and, furthermore, the non-normality.

8. Conclusion

In this project, we use SARIMA model to capture the trend of cocoa volume. Initially, we transformed the data to be stationary, ensuring that the statistical properties of the series do not change over time. We then identified possible models in Table (2) based on the ACF Figure (3) and PACF Figure (4), considering seasonal factors of 13, 26, and 52, representing quarterly, semi-annual, and annual cycles respectively. Both single and multiple seasonality were explored, including SARIMA and other multiple seasonality models. Subsequent to determination of different combinations of order, we

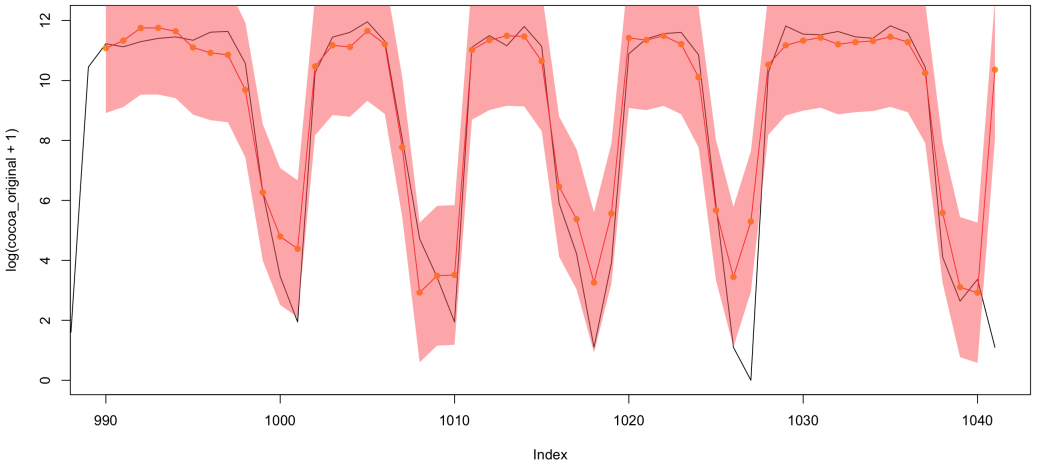

Figure 11. Prediction of **SARIMA**$(13,0,0) \times (1,0,1)_{52}$

then fit those models and conduct diagnostic check, including examining the significance of highest orders, Ljung-Box test and residuals visual checks via ACF plot and Q-Q plot. After performing diagnostic checks, Model 7, Model 8 and 10 are identified as valid models. Also, model selection was implemented using AIC and BIC. Model 7 and 10 are recognized as equally optimal model due to their comparable performance. Furthermore, we provide MSE and prediction interval for these models. The results of prediction are indifferentiable according to the test of MSE. Hence, the performance of Model 7 and 10 are almost identical with Model 10 showing slightly better result on AIC, BIC and MSE.

Recall that one of the primary focus of this study is to offer insights for cocoa farmers. Throughout the modeling and data analysis process, we unveiled that a quarter of a year, namely, 13 weeks, is a critical time period to which cocoa traders and farmers should pay attention. Also, the trading volumes arrive at its valley (i.e., local minimum) for 4-5 times annually and the time range between valleys are trading peaks where a large amount of cocoa futures contracts has been signed.

Also, considering that some data points still fall outside the prediction interval, there is still room for model improvement. To name a few possible approaches for improving the analysis, modern advancements in neural networks and deep learning have shown great performance in capturing patterns in seasonal time series models (Zhang & Qi, 2005). Additionally, non-linear properties should be considered, as some may appear as white noise in linear models. The autoregressive conditional heteroskedasticity (ARCH) model, which has already been applied in futures markets, can be used to explain the non-linearities in the data (Adrangi & Chatrath, 2003). Likewise, the threshold autoregressive (TAR) model (Tong & Lim, 1980) or regime-switching model (Hamilton, 1989) can help handle the nonlinearity in data . Thus, future research could explore the integration of these advanced methodologies to further refine the forecasting accuracy and reliability of cocoa trading volumes.

Acknowledgement

The authors express their ultimate gratitude to Dr. Yu-Wei Chang and Dr. Li-Pang Chen for their valuable comments on this study.

Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this work.

Notes

The authors have found inconsistency in the results of computation. However, the primary conclusions are not affected by the variability of computation.

References

Adrangi, B., & Chatrath, A. (2003). Non-linear dynamics in futures prices: Evidence from the coffee, sugar and cocoa exchange. *Applied Financial Economics*, *13*(4), 245–256.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* (pp. 199–213). Springer.

Assis, K., Amran, A., & Remali, Y. (2010). Forecasting cocoa bean prices using univariate time series models. *Researchers World*, *1*(1), 71.

Beg, M. S., Ahmad, S., Jan, K., & Bashir, K. (2017). Status, supply chain and processing of cocoa-a review. *Trends in food science & technology*, *66*, 108–116.

Box, G. E., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American statistical association*, *106*(496), 1513–1527.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427–431.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business Economic Statistics*, *13*(3), 253–263.

Dragusanu, R., Giovannucci, D., & Nunn, N. (2014). The economics of fair trade. *Journal of economic perspectives*, *28*(3), 217–236.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, 357–384.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, *13*(2), 281–291.

Hassani, H., & Yeganegi, M. R. (2020). Selecting optimal lag order in ljung–box test. *Physica A: Statistical Mechanics and its Applications*, *541*, 123700.

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

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of statistical software*, *27*, 1–22.

Li, W. K. (2003). *Diagnostic checks in time series*. Chapman; Hall/CRC.

Ljung, G. M. (1986). Diagnostic testing of univariate time series models. *Biometrika*, *73*(3), 725–730.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Montagna, M. T., Diella, G., Triggiano, F., Caponio, G. R., Giglio, O. D., Caggiano, G., Ciaula, A. D., & Portincasa, P. (2019). Chocolate,"food of the gods": History, science, and human health. *International journal of environmental research and public health*, *16*(24), 4960.

Osborn, D. R., Chui, A. P. L., Smith, J. P., & Birchenhall, C. R. (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics*, *50*(4), 361–377.

Ouyang, H., Wei, X., & Wu, Q. (2019). Agricultural commodity futures prices prediction via long-and short-term time series network. *Journal of Applied Economics*, *22*(1), 468–483.

Ozturk, G., & Young, G. M. (2017). Food evolution: The impact of society and science on the fermentation of cocoa beans. *Comprehensive reviews in food science and food safety*, *16*(3), 431–455.

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607.

Sarfo, S., & Geman, H. (2012). Seasonality in cocoa spot and forward markets: Empirical evidence. *Journal of Agricultural Extension and Rural Development*, *4*(8), 164–180.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications* (Vol. 3). Springer.

Tong, H., & Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(3), 245–268.

Tsay, R. S. (2005). *Analysis of financial time series*. John wiley & sons.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European journal of operational research*, *160*(2), 501–514.

Appendix 1. Parameters Estimation of All Models

The parameters estimation of each model shown in Table (2) are shown here.

Table 3: Parameter Estimation of Model 1–4.
Note: *p<0.1; **p<0.05; ***p<0.01

Parameter	Model 1	Model 2	Model 3	Model 4
ϕ_1	0.029	1.401***	0.352	0.444
ϕ_2	0.269***	–0.596***		
ϕ_3	–0.281***	–0.133**		
ϕ_4	–0.321***	–0.155***		
ϕ_5	–0.103***	0.279***		
ϕ_6	–0.106***	–0.062		
ϕ_7	–0.134***	–0.086		
ϕ_8	0.104***	0.144**		
ϕ_9	0.270***	0.120**		
ϕ_{10}	–0.164***	–0.340***		
ϕ_{11}	–0.080**	0.198***		
ϕ_{12}	0.011	–0.038		
ϕ_{13}	0.340***	0.500***		
ϕ_{14}		–0.635***		
ϕ_{15}		0.236***		
ϕ_{16}		–0.030		
ϕ_{17}		0.258***		
ϕ_{18}		–0.224***		
ϕ_{19}		–0.037		
ϕ_{20}		0.110*		
ϕ_{21}		–0.047		
ϕ_{22}		–0.017		
ϕ_{23}		–0.066		
ϕ_{24}		0.073		
ϕ_{25}		–0.020		
ϕ_{26}		–0.061*		
θ_1	0.655***	–0.729***	0.270	0.192
θ_2			0.047	0.013
θ_3			–0.229**	–0.234**
θ_4			–0.458***	–0.472***
θ_5			–0.367*	–0.334**
θ_6			–0.178	–0.155
θ_7			0.003	0.015

Continued on next page

Table 3: Parameter Estimation of Model 1-4.
Note: *p<0.1; **p<0.05; ***p<0.01 (Continued)

θ_8			0.343***	0.319***
θ_9			0.508***	0.500***
θ_{10}			0.203	0.176
θ_{11}			0.049	0.033
θ_{12}			–0.219**	–0.200**
θ_{13}			0.111*	0.145***
θ_{14}				–0.049
θ_{15}				–0.047
θ_{16}				–0.102**
θ_{17}				0.001
θ_{18}				0.078
θ_{19}				0.010
θ_{20}				0.007
θ_{21}				–0.028
θ_{22}				0.071*
θ_{23}				–0.062
θ_{24}				–0.060
θ_{25}				–0.066
θ_{26}				0.098**
μ	8.666***	8.668***	8.670***	8.667***

Table 4: Parameter Estimation of Model 5-10. *Note:* *p<0.1; **p<0.05; ***p<0.01

Parameter	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
ϕ_1	0.059	0.045	0.009	0.246***	0.523***	0.218***
ϕ_2	0.088**	0.214***	0.081**	0.043	–0.048	0.038
ϕ_3	–0.048	–0.223***	–0.052*	–0.055*	–0.204***	–0.060*
ϕ_4	–0.090***	–0.266***	–0.095***	–0.079**	–0.142***	–0.080**
ϕ_5	–0.045	–0.107***	–0.051	–0.032	–0.049	–0.038
ϕ_6	–0.054*	–0.115***	–0.047	–0.050	–0.105***	–0.041
ϕ_7	0.028	–0.068**	0.020	0.036	–0.012	0.028
ϕ_8	0.060*	0.106***	0.061**	0.055*	0.125***	0.056*
ϕ_9	0.083**	0.197***	0.119***	0.073**	0.135***	0.106***
ϕ_10	0.020	–0.095***	0.019	0.009	–0.147***	–0.0003

Continued on next page

Table 4: Parameter Estimation of Model 5-10. *Note:* *p<0.1; **p<0.05; ***p<0.01 (Continued)

ϕ_{11}	0.006	–0.055	0.006	0.006	0.043	0.009
ϕ_{12}	–0.091***	–0.009	–0.101***	–0.095***	–0.019	–0.106***
ϕ_{13}	0.215***	0.299***	0.241***	0.222***	0.232***	0.251***
θ_1	0.199	0.524***	0.226*			
Φ_1	0.857***		0.921***	0.860***		0.921***
Θ_1		0.517***	–0.245***		0.539***	–0.239***
μ	8.656***	8.670***	8.655***	8.655***	8.672***	8.654***

Table 5: Parameter Estimation of Model 11-13.
Note: *p<0.1; **p<0.05; ***p<0.01

Parameter	Model 11	Model 12	Model 13
ϕ_1	0.571***	1.280***	–0.200
ϕ_2	–0.044	–0.450***	0.160
ϕ_3	–0.195***	–0.164***	–0.040
ϕ_4	–0.234***	–0.097*	–0.124***
ϕ_5	–0.018	0.148**	–0.072
ϕ_6	–0.039	–0.024	–0.051
ϕ_7	–0.026	–0.002	0.007
ϕ_8	0.082**	0.099*	0.065*
ϕ_9	0.106***	0.056	0.088**
ϕ_{10}	–0.143***	–0.216***	0.038
ϕ_{11}	0.020	0.121**	0.009
ϕ_{12}	–0.006	–0.027	–0.083**
ϕ_{13}	0.384***	0.399***	0.191***
ϕ_{14}	–0.163***	–0.427***	0.079
ϕ_{15}	–0.020	0.098*	–0.025
ϕ_{16}	–0.025	–0.016	–0.013
ϕ_{17}	0.196***	0.214***	0.073**
ϕ_{18}	0.014	–0.120**	0.059
ϕ_{19}	–0.091**	–0.100*	–0.029
ϕ_{20}	0.051	0.113**	0.044
ϕ_{21}	–0.016	–0.053	0.006
ϕ_{22}	0.009	0.012	–0.010

Continued on next page

Table 5: Parameter Estimation of Model 11-13. *Note:* *p<0.1; **p<0.05; ***p<0.01 (Continued)

ϕ_{23}	–0.038	–0.058	–0.021
ϕ_{24}	0.014	0.036	–0.026
ϕ_{25}	–0.009	0.001	–0.021
ϕ_{26}	–0.033	–0.049	0.037
θ_1		–0.715***	0.462
Θ_1	0.498***	0.493***	
Φ_1			0.853***
μ	8.672***	8.673***	8.658***

Appendix 2. Supplementary Information of Diagnostic

We list out the diagnostic plots not included in Section 5.3 here.

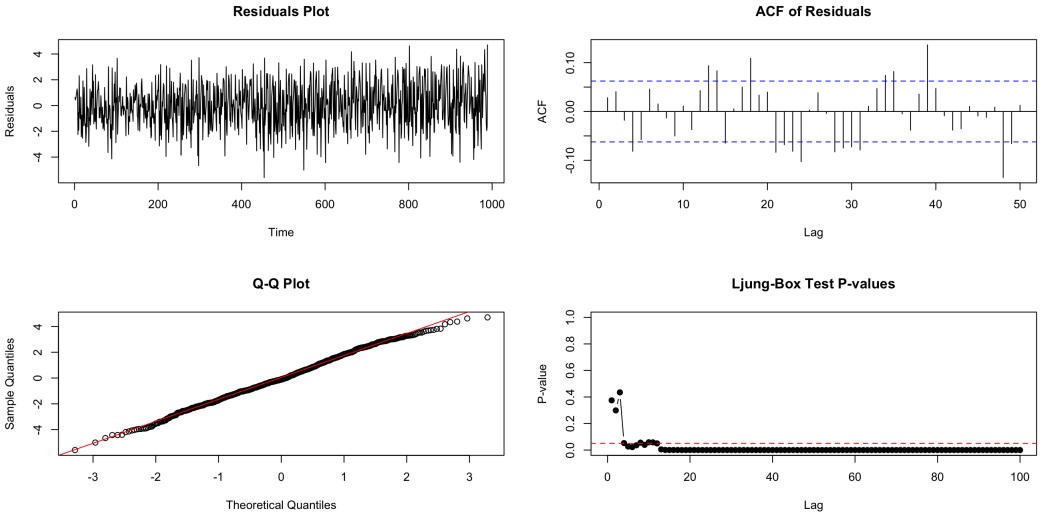

Figure 12. Diagnosis of **ARMA**$(13, 0, 1)$

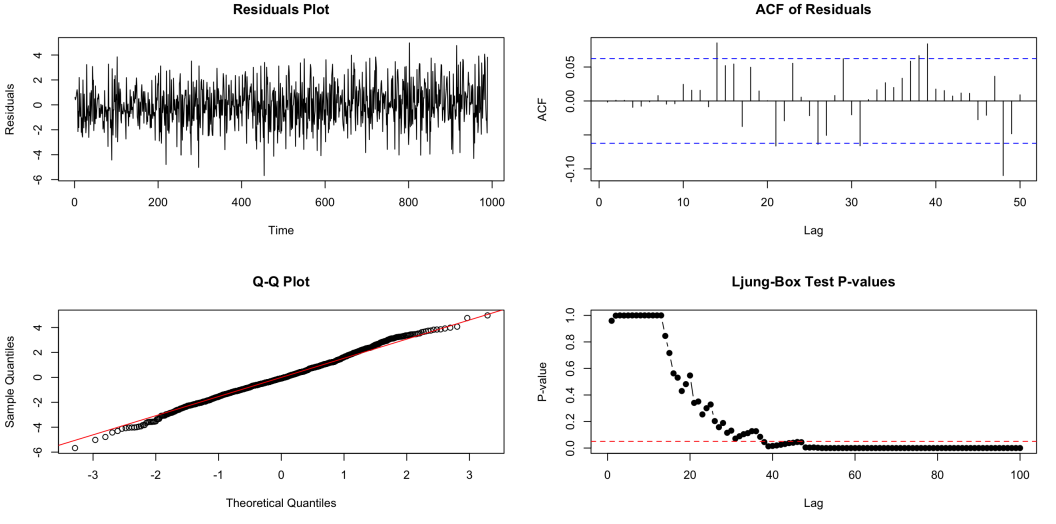

Figure 13. Diagnosis of **ARMA**$(26, 0, 1)$

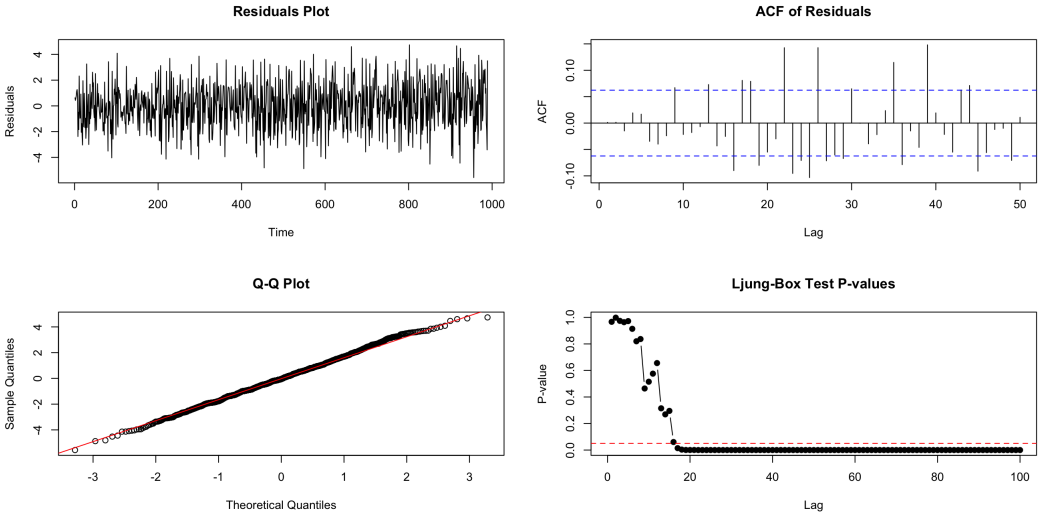

Figure 14. Diagnosis of **ARMA**$(1, 0, 13)$

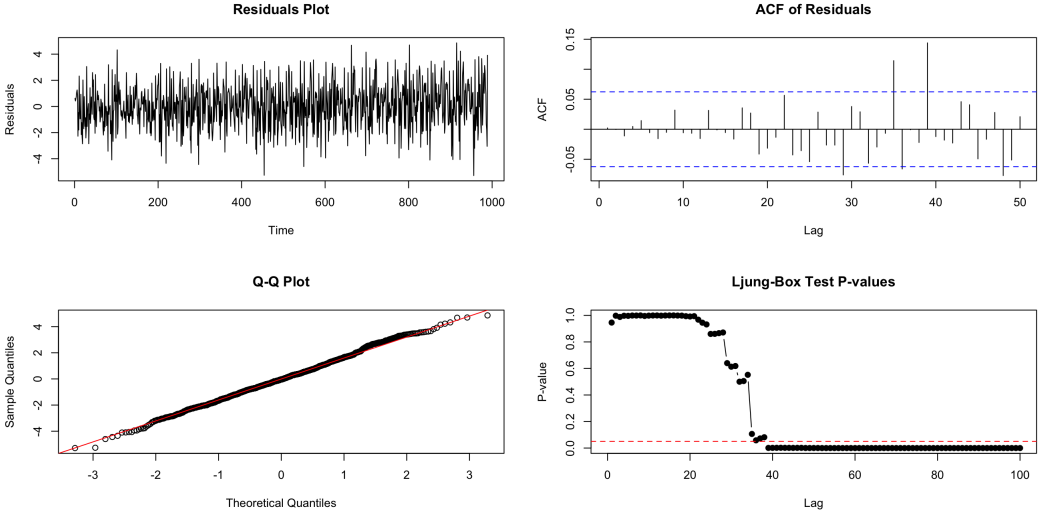

Figure 15. Diagnosis of **ARMA**$(1, 0, 26)$

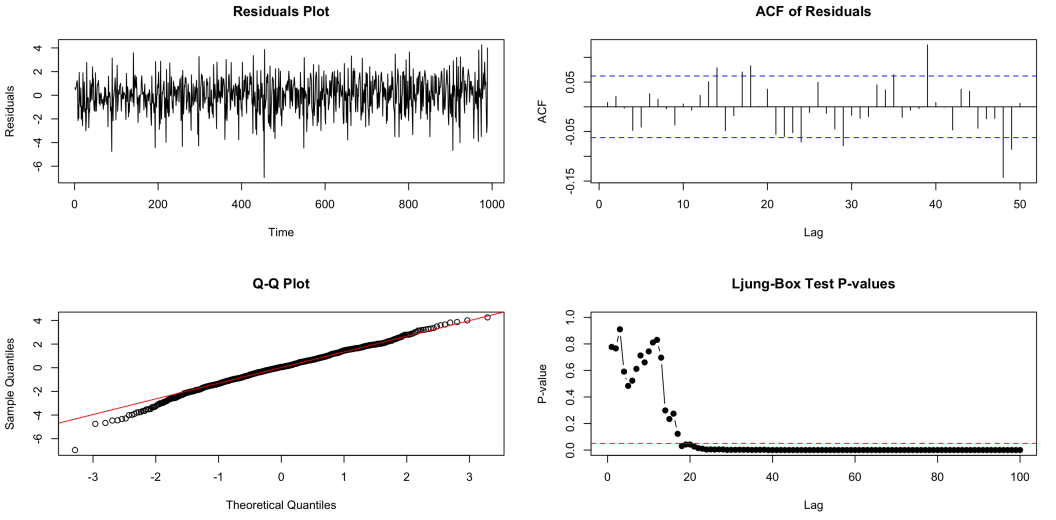

Figure 16. Diagnosis of **SARIMA**$(13,0,1)\times(0,0,1)_{52}$

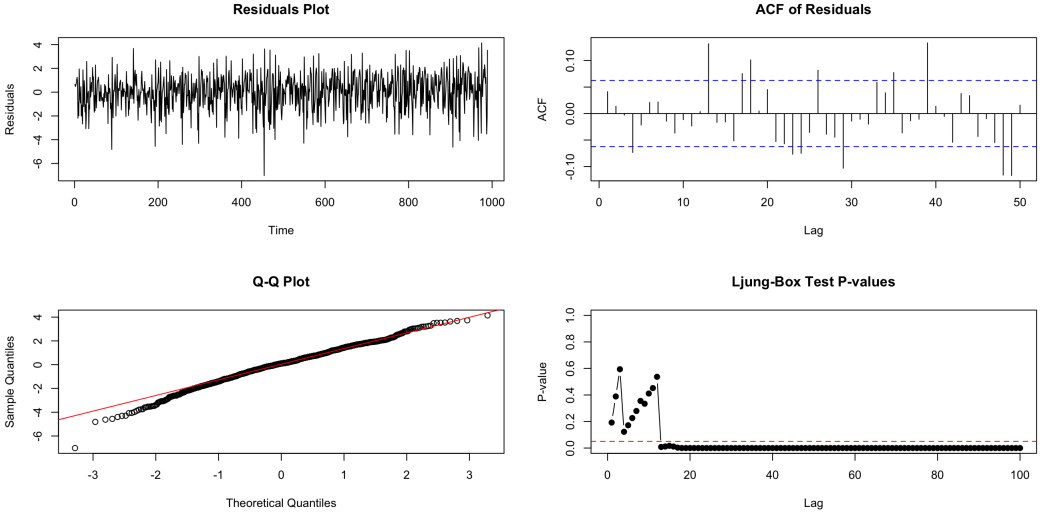

Figure 17. Diagnosis of **SARIMA**$(13,0,0)\times(0,0,1)_{52}$

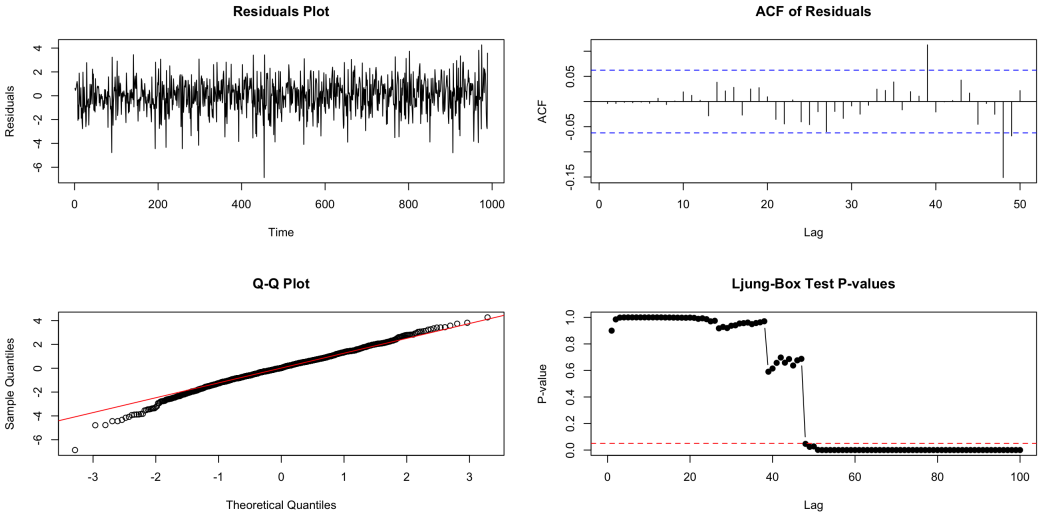

Figure 18. Diagnosis of **SARIMA**$(26,0,0) \times (0,0,1)_{52}$

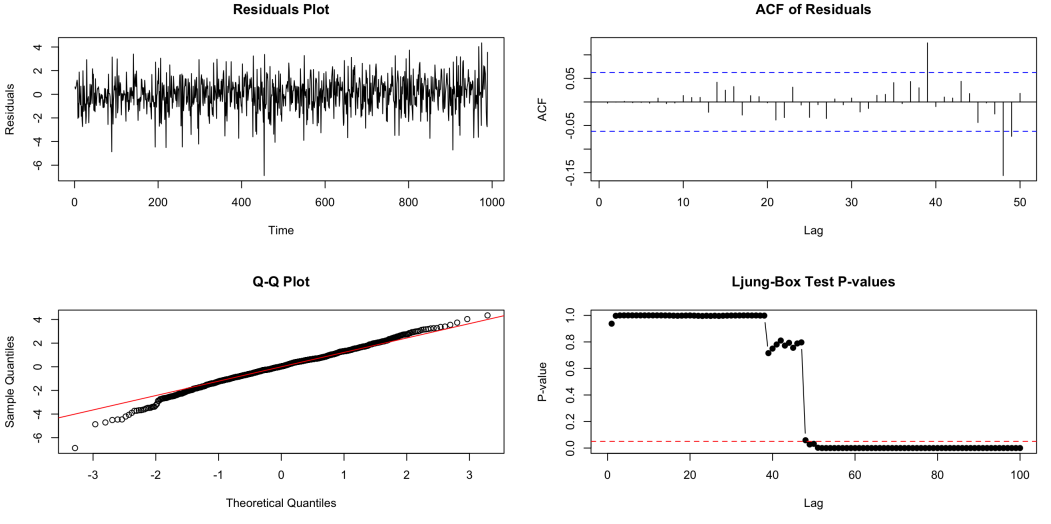

Figure 19. Diagnosis of **SARIMA**$(26,0,1) \times (0,0,1)_{52}$