Time Series Analysis Final Report

Cocoa Bean Volume Prediction

Minen Sun[1], Ching Wang[2] and Kuan-Hsun Wu[3]

Department of Statistics,
National Chengchi University

June 13, 2024

[1] 109304018@g.nccu.edu.tw
[2] 109304015@g.nccu.edu.tw
[3] 110304015@g.nccu.edu.tw

Outline

1. Review
   - Introduction
   - Data Preprocessing
   - ARIMA Order Selection
2. Model Building
   - Model Building Strategies
   - SARIMA
3. Diagnostic
   - Ljung-Box Test
4. Model Selection
   - AIC&BIC
5. Prediction
6. Conclusion
7. References

Introduction

- Cocoa is the *food of the Gods*! (Montagna et al., 2019)
- *"Theobroma cacao normally begins to have berries after 3 years and yield reaches maximum after 8-9 years... some countries have two high production season per year."* (Beg, Ahmad, Jan, and Bashir, 2017)
- The contracts of cocoa futures will be due at March, May, July, September or December.
- *"Seasonality plays an essential role in managing risk in agricultural commodities..."* (Sarfo and Geman, 2012)

Data Source I

ICE (*Intercontinental Exchange)* provides a platform of trading cocoa futures and Yahoo Finance provides data of cocoa futures[4] , it is consists of

- Open price (USD)
- Close price (USD)
- Highest price (USD)
- Lowest price (USD)
- Volume (metric ton)

where the data is collected weekly.

Data Source II

We aim to model on the weekly volume of cocoa futures. In this presentation, we denote the time series of interest by $\{X_t\}$.

- Time period: 2000 01-03 to 2019-12-09, $n = 1041$
- Training Data: 2000 01-03 to 2018 12-03, $n = 989$
- Testing Data: 2018-12-10 to 2019-12-09, $n = 52$

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

Time Series Plot

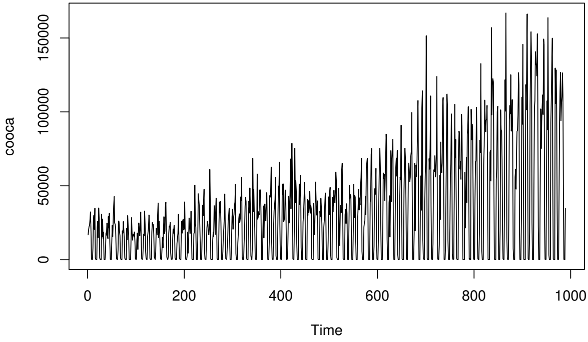

Figure 1: Time Series Plot of X_t

The process performs heteroscedasticity. (Larger variance among time.)

Transformation

Due to heteroscedasticity, we implement log transform to our data. The transformed data seems to be stationary.

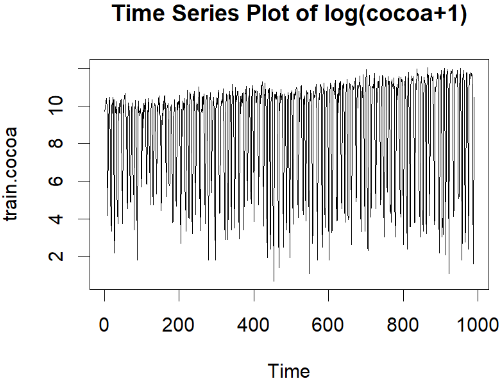

Figure 2: Time Series Plot of $log(X_t + 1)$

ACF plot

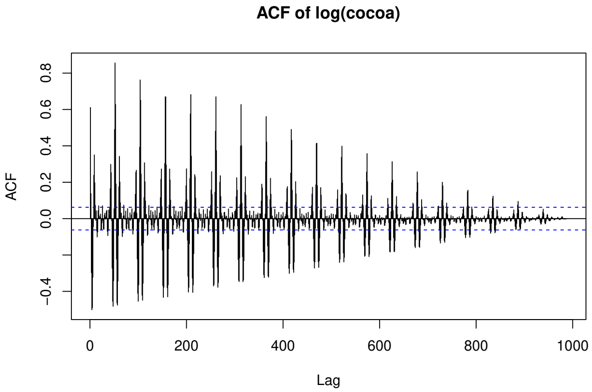

Figure 3: ACF Plot of $log(X_t + 1)$

The ACF does decay. This provides a empirical proof of stationarity in $log(X_t + 1)$.

Stationary Test

To further ensure that $log(X_t + 1)$ is stationary, we implement the augmented Dickley-Fuller test Said and Dickey (1984). The null and alternate hypothesis are stated as follow:

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

We set $\alpha = 0.05$. As the p-value is far smaller than 0.05, there is significant statistical evidence to support that $log(X_t + 1)$ is a stationary time series. Hence, we decide not to conduct first order differencing.

PACF plot

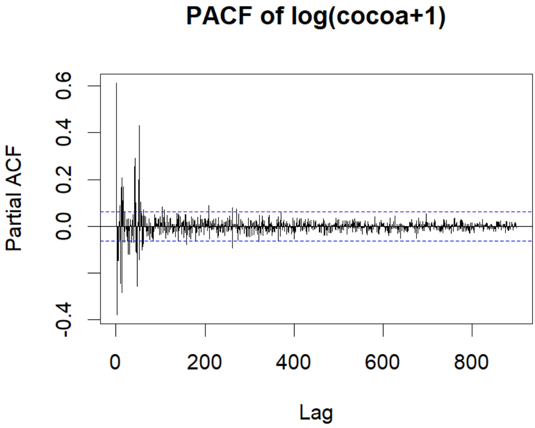

Figure 4: PACF Plot of $log(X_t + 1)$

Significant Lags

	Lags
ACF	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	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 . . .

Table 1: Significant Lags

Seasonality

Figure (3) had shown an annual pattern. 52 weeks form a year.

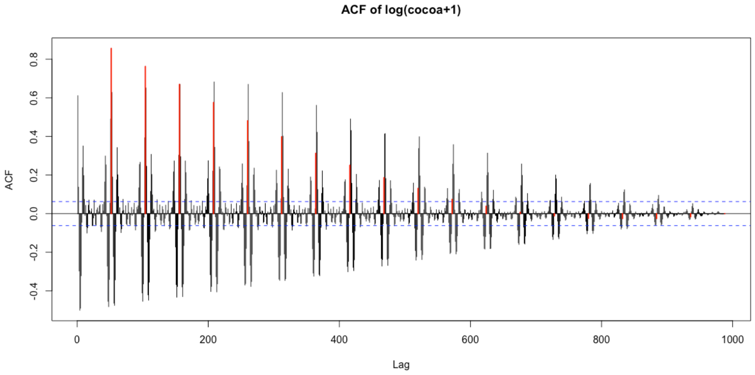

Figure 5: ACF Plot of $log(X_t + 1)$ with multiples of 52 colored

Test for Seasonality

To identify the seasonality in the data, we implemented the test proposed by Osborn et al. (1988). The null and alternate hypothesis are

$$H_0: \text{There exists seasonal unit root.}$$
$$H_1: \text{There does not exist seasonal unit root.}$$

We set our significance level to $\alpha = 0.05$. As the p-value is far smaller than 0.05, there is significant statistical evidence to support that $log(X_t + 1)$ is a seasonal stationary time series. Hence, we decide not to implement seasonal differencing.

Model Building 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 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 multiple seasonalities $(s_1, s_2, \ldots, s_n)$.
  - **SARIMA**$(p, d, q) \times (P, D, Q)_S$: With a more large seasonal factor handled by S, the other seasonal factor can be placed in p and q, the non-seasonal part.

Seasonal ARIMA (SARIMA)

The seasonal arima is of the form

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

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;
- $Z_t \sim wn(0, \sigma^2)$ is white noise.

Order Selection I

- **Scenario 1**: We can regard the behavior of ACF and PACF (Figure 3,5) as tailing off. Hence, we consider ARMA(p, q) models with p and q shown in Table (1).

ARMA$(13, 1)$
ARMA$(1, 13)$
ARMA$(26, 1)$
ARMA$(1, 26)$

Table 2: Possible Models of ARMA

The computation capability allows to extend more cases. Some successful ones are included.

Order Selection II

- **Scenario 2**: We have observed the seasonality of 52 in Figure (3). From a seasonal perspective, the ACF (Figure 3) and the PACF (Figure 5) can be viewed as tailing off and cutting off at lag $h = 52$, respectively. Therefore, we are taking $\text{SARIMA}(p, 0, q) \times (1, 0, 0)_{52}$ into account.

$\text{SARIMA}(13, 0, 1) \times (1, 0, 0)_{52}$	$\text{SARIMA}(13, 0, 0) \times (1, 0, 1)_{52}$
$\text{SARIMA}(13, 0, 1) \times (0, 0, 1)_{52}$	$\text{SARIMA}(26, 0, 0) \times (0, 0, 1)_{52}$
$\text{SARIMA}(13, 0, 1) \times (1, 0, 1)_{52}$	$\text{SARIMA}(26, 0, 1) \times (0, 0, 1)_{52}$
$\text{SARIMA}(13, 0, 0) \times (1, 0, 0)_{52}$	$\text{SARIMA}(26, 0, 1) \times (1, 0, 0)_{52}$
$\text{SARIMA}(13, 0, 0) \times (0, 0, 1)_{52}$	

Table 3: Possible Models of SARIMA

Ljung-Box Test

To test whether the fitted model is adequate, Ljung-Box test (Ljung and Box, 1978) is used. The null and alternate hypothesis are stated as follow:

$$
\begin{aligned}
H_0 &: \text{ The fitted model is adequate;} \\
H_1 &: \text{ The fitted model is not adequate.}
\end{aligned}
$$

The test statistic is

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

and as $n \to \infty, Q \sim \chi^2_{H-(p+q)}$ under H_0.

Selection on H

- Ljung (1986) suggested $H = 5$.
- Shumway and Stoffer (2000) suggested $H = 20$.
- Tsay (2005) suggested
  $H = log(n) = log(989) = 6.896694 \doteqdot 7$.
- Hyndman and Athanasopoulos (2018) suggested
  $H = min(10, \frac{n}{5}) = 10$.

Diagnostic - White Noise

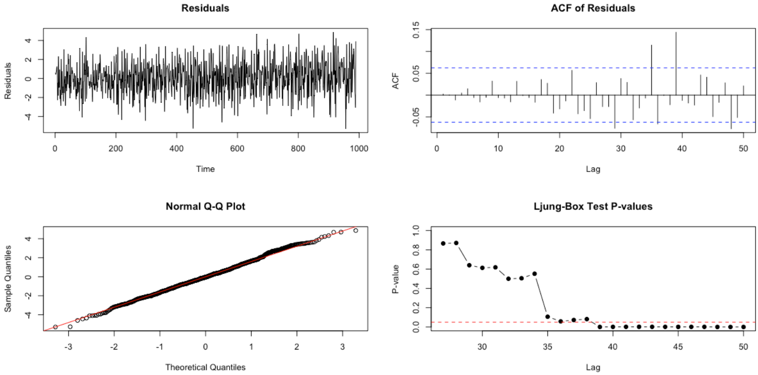

Figure 6: Diagnosis of Gaussian White Noise

Comparison: ARMA(p, q)

Model	Lowest significant lag	p+q
ARMA$(13, 1)$	5	14
ARMA$(1, 13)$	28	14
ARMA$(26, 1)$	38	27
ARMA$(1, 26)$	39	27

Table 4: Results of Ljung-Box Test of ARMA(p, q) Models

Diagnostic - ARMA(13,1)

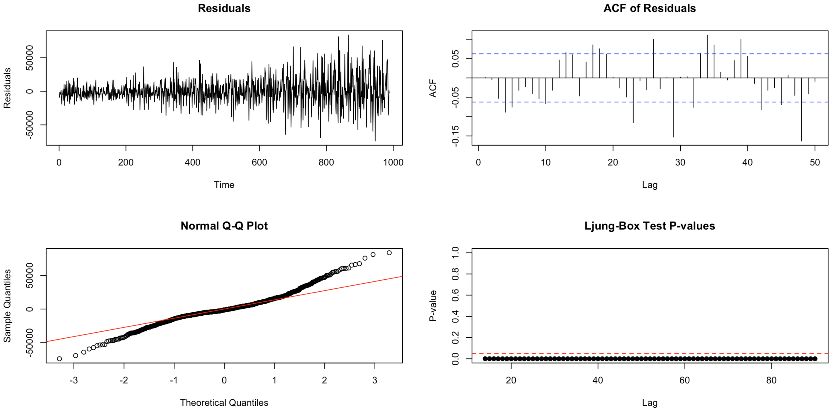

Figure 7: Diagnosis of ARIMA(13,1)

Diagnostic - ARMA(1,13)

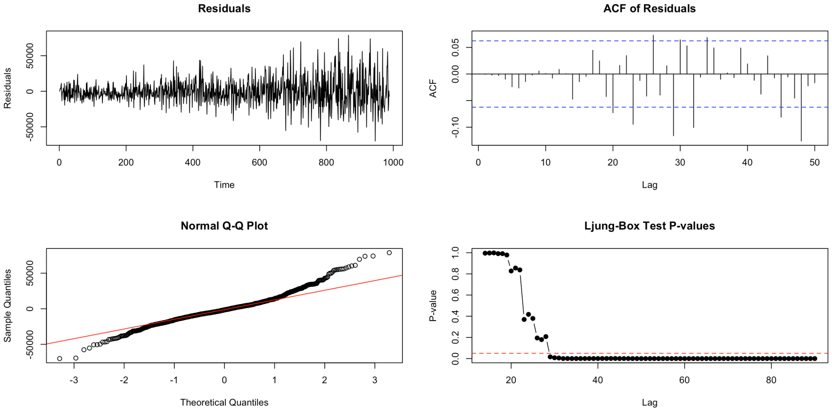

Figure 8: Diagnosis of ARIMA(1,13)

Diagnostic - ARMA(26,1)

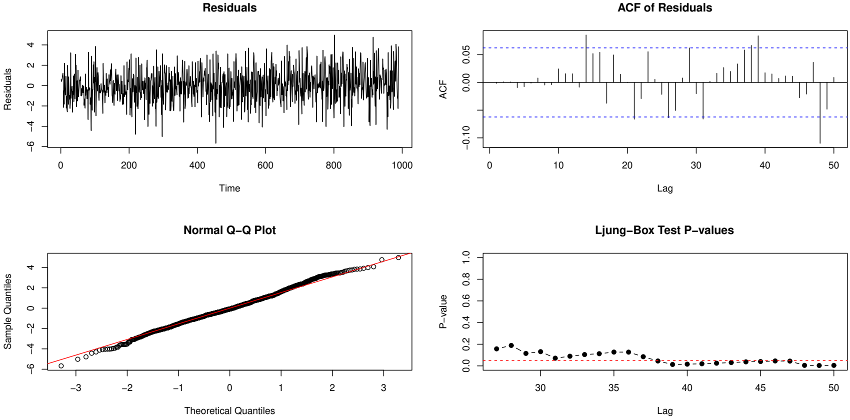

Figure 9: Diagnosis of ARIMA(26,1)

Diagnostic - ARMA(1,26)

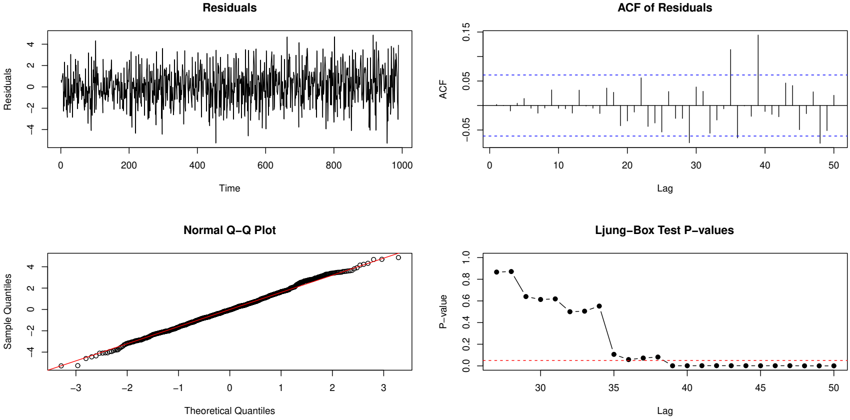

Figure 10: Diagnosis of ARIMA(1,26)

Comparison: SARIMA$(p, 0, q) \times (P, 0, Q)_S$

Model	Lowest significant lag	$p + q + P + Q$
SARIMA$(13, 0, 1) \times (1, 0, 0)_{52}$	18	15
SARIMA$(13, 0, 1) \times (0, 0, 1)_{52}$	52	15
SARIMA$(13, 0, 1) \times (1, 0, 1)_{52}$	52	16
SARIMA$(13, 0, 0) \times (1, 0, 0)_{52}$	52	14
SARIMA$(13, 0, 0) \times (0, 0, 1)_{52}$	18	14
SARIMA$(13, 0, 0) \times (1, 0, 1)_{52}$	61	15
SARIMA$(26, 0, 0) \times (0, 0, 1)_{52}$	46	27
SARIMA$(26, 0, 1) \times (0, 0, 1)_{52}$	46	28
SARIMA$(26, 0, 1) \times (1, 0, 0)_{52}$	47	28

Table 5: Results of Ljung-Box Test of SARIMA$(p, 0, q) \times (P, 0, Q)_S$ Models

Diagnostic - SARIMA$(13, 0, 1) \times (0, 0, 1)_{52}$*

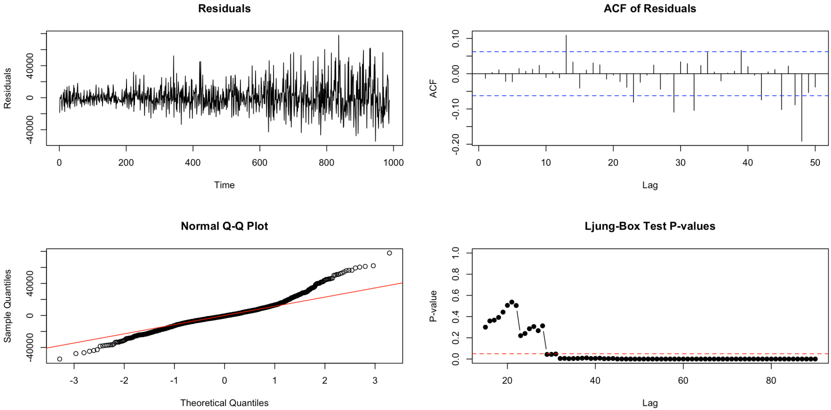

Figure 11: Diagnosis of SARIMA$(13, 0, 1)(0, 0, 1)_{52}$

Diagnostic - SARIMA$(13, 0, 1) \times (1, 0, 0)_{52}$

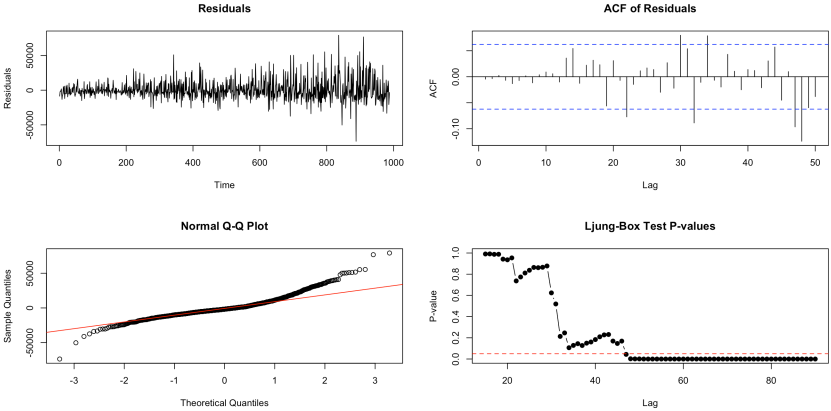

Figure 12: Diagnosis of SARIMA$(13, 0, 1) \times (1, 0, 0)_{52}$

Diagnostic - SARIMA$(13,0,1) \times (1,0,1)_{52}$*

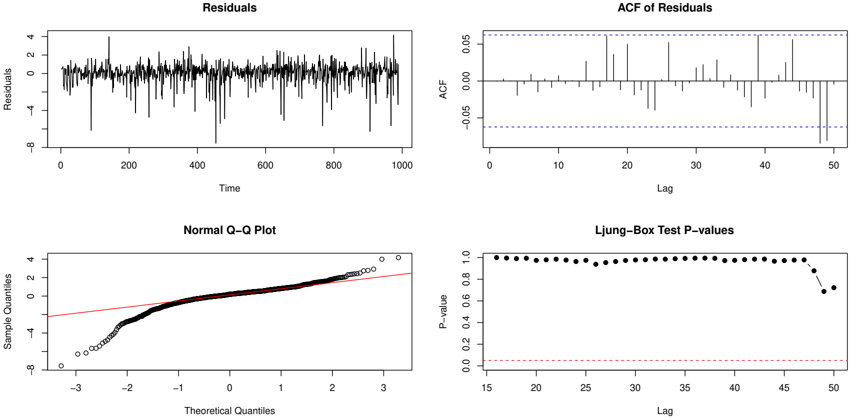

Figure 13: Diagnosis of SARIMA$(13,0,1) \times (1,0,1)_{52}$

Diagnostic - SARIMA$(13,0,0) \times (1,0,0)_{52}$*

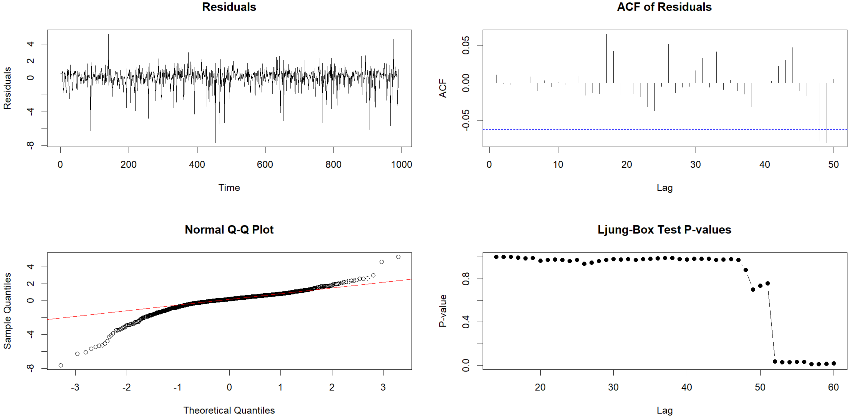

Figure 14: Diagnosis of SARIMA$(13,0,0) \times (1,0,0)_{52}$

Diagnostic - SARIMA$(13, 0, 0) \times (0, 0, 1)_{52}$

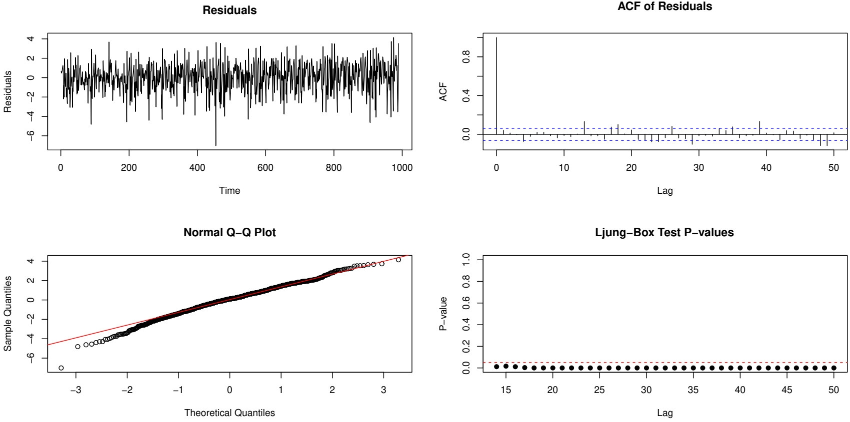

Figure 15: Diagnosis of SARIMA$(13, 0, 0) \times (0, 0, 1)_{52}$

Diagnostic - SARIMA$(13, 0, 0) \times (1, 0, 1)_{52}$*

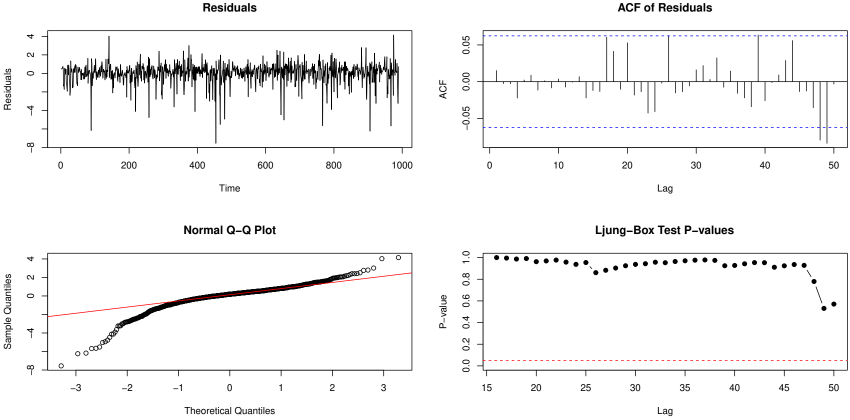

Figure 16: Diagnosis of SARIMA$(13, 0, 0) \times (1, 0, 1)_{52}$

Diagnostic - SARIMA$(26,0,0)\times(0,0,1)_{52}$

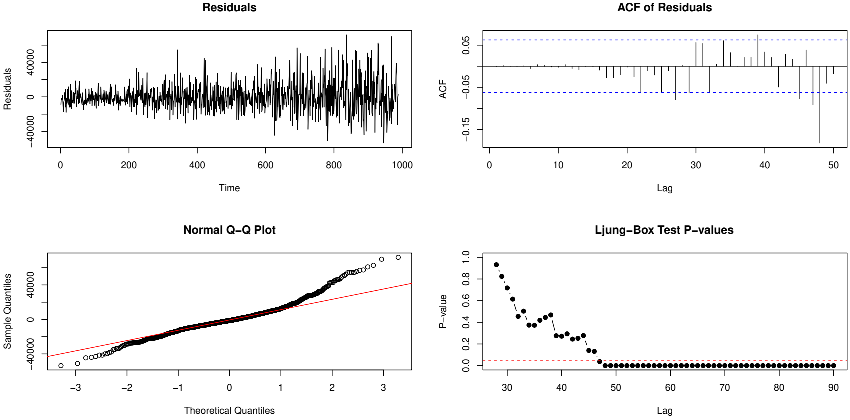

Figure 17: Diagnosis of SARIMA$(26,0,0)\times(0,0,1)_{52}$

Diagnostic - SARIMA$(26, 0, 1) \times (1, 0, 0)_{52}$

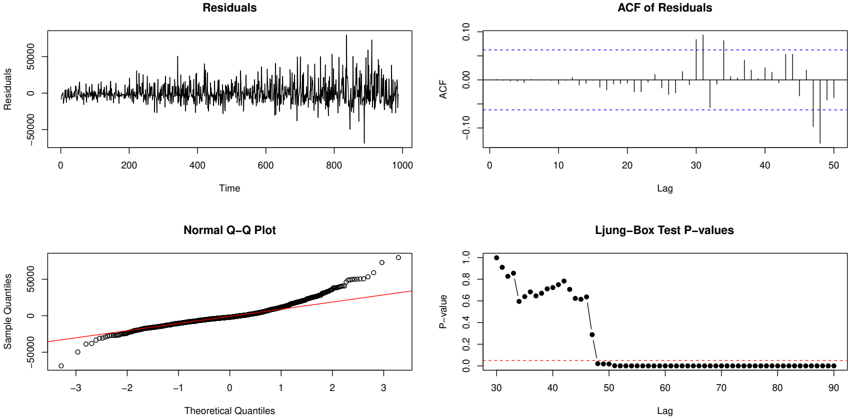

Figure 18: Diagnosis of SARIMA$(26, 0, 1) \times (1, 0, 0)_{52}$

Diagnostic - SARIMA$(26, 0, 1) \times (0, 0, 1)_{52}$

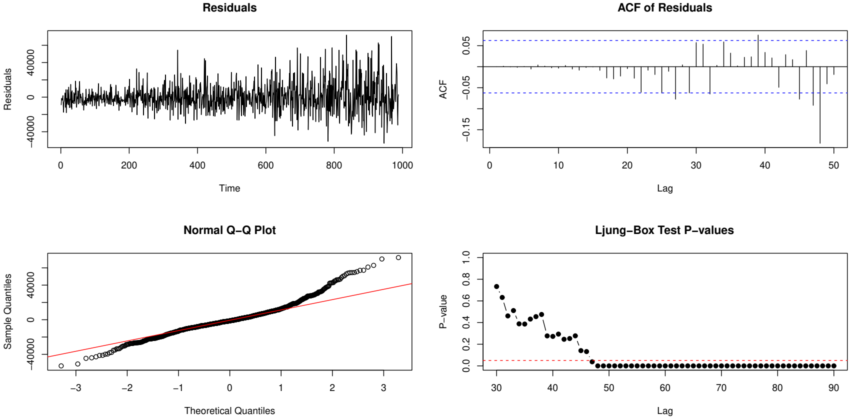

Figure 19: Diagnosis of SARIMA$(26, 0, 1) \times (0, 0, 1)_{52}$

AIC&BIC

Model	AIC	BIC
$\text{SARIMA}(13,0,1) \times (0,0,1)_{52}$	3562.943	3646.187
$\text{SARIMA}(13,0,1) \times (1,0,1)_{52}$	3111.427	3199.567
$\text{SARIMA}(13,0,0) \times (1,0,0)_{52}$	3153.561	3231.908
$\text{SARIMA}(13,0,0) \times (1,0,1)_{52}$	3113.005	3196.249

Table 6: AIC and BIC of Models

Prediction SARIMA$(13, 0, 0) \times (1, 0, 1)_{52}$

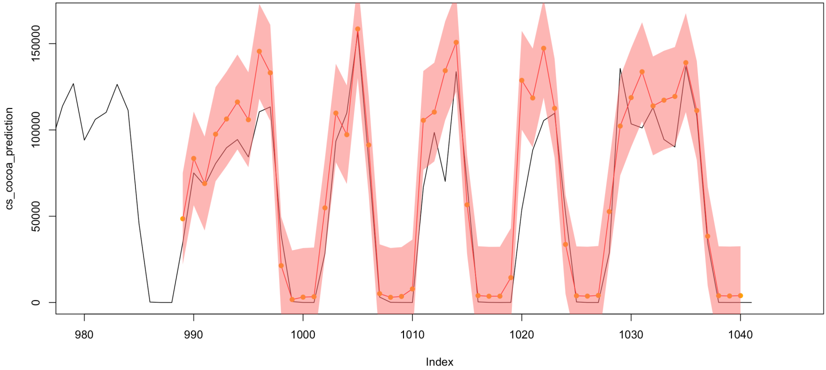

Figure 20: Prediction of SARIMA$(13, 0, 0) \times (1, 0, 1)_{52}$

Prediction SARIMA$(13, 0, 1) \times (1, 0, 1)_{52}$

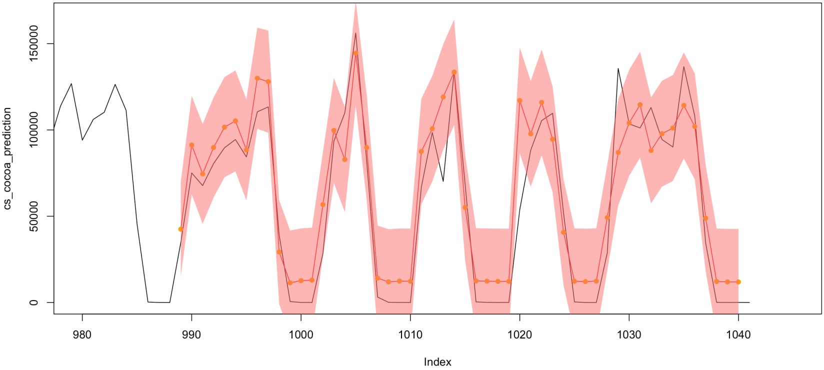

Figure 21: Prediction of SARIMA$(13, 0, 1) \times (1, 0, 1)_{52}$

Prediction

Model	MSE
$\text{SARIMA}(13,0,1) \times (1,0,1)_{52}$	2.896954
$\text{SARIMA}(13,0,0) \times (1,0,1)_{52}$	2.880642

Table 7: Comparison of MSE

Takeaway

- According AIC and BIC, SARIMA$(13, 0, 1) \times (1, 0, 1)$ and SARIMA$(13, 0, 0) \times (1, 0, 1)$ are selected.
- The prediction shows SARIMA$(13, 0, 0) \times (1, 0, 1)$ are slightly better than SARIMA$(13, 0, 1) \times (1, 0, 1)$.
- The use of nonlinear model. For example, threshold autoregressive model (Tong and Lim, 1980) and its extension.

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.

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

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., … 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., et al. (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics*, *50*(4), 361–377.

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.

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.