Journal of Applied Mathematics and Physics, 2024, 12, 1237-1246
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Building the ARIMA Model for Forecasting the Production of Vietnam's Coffee Export

Duy Quang Phung[1*], Quoc Thang Trinh[2], Quang Truong Do[2], Ngan Giang Nguyen[2], Van Ha Nguyen[2], Gia Khiem Ngo[2], Thi Minh Ngoc Tran[2]

[1]Faculty of Technology and Data Science, Foreign Trade University, Hanoi, Vietnam
[2]School of Economics and International Business, Foreign Trade University, Hanoi, Vietnam
Email: *quangpd@ftu.edu.vn, k61.2211110372@ftu.edu.vn, k61.2211110408@ftu.edu.vn, k61.2211110100@ftu.edu.vn, k61.2211110109@ftu.edu.vn, k61.2211110178@ftu.edu.vn, k61.2211110278@ftu.edu.vn

How to cite this paper: Phung, D.Q., Trinh, Q.T., Do, Q.T., Nguyen, N.G., Nguyen, V.H., Ngo, G.K. and Tran, T.M.N. (2024) Building the ARIMA Model for Forecasting the Production of Vietnam's Coffee Export. *Journal of Applied Mathematics and Physics*, **12**, 1237-1246.
https://doi.org/10.4236/jamp.2024.124076

Received: March 11, 2024
Accepted: April 25, 2024
Published: April 28, 2024

Abstract

Coffee is a significant industry, accounting for 3% of Vietnam's GDP, with annual export turnover consistently exceeding USD 3 billion. Despite global economic challenges affecting purchasing power at various times, Vietnam's coffee exports in December 2023 continued to surge, reaching the highest level in the past 9 months at 190,000 tons, a 59.3% increase compared to November 2023, but still a slight 3.5% decrease from the same period last year. The export turnover reached USD 538 million, a 51% increase from November 2023 and a 26.4% increase from the same period last year. Therefore, forecasting the coffee export volume holds significant importance for coffee producers nationwide. This research employs the Box-Jenkins method to construct an ARIMA model for forecasting Vietnam's coffee export volume based on annual data published by the General Statistics Office. Results indicate that among the models considered, the ARIMA(1, 1, 2) model is the most suitable. The study also provides short-term forecasts for Vietnam's coffee export volume. However, the current model is limited to forecasting and is not yet optimized, as the assumed linearity in the model is a simplification.

Keywords

ARIMA, Forecasting, Coffee Export Volume, Data Science

1. Introduction

Over the past three decades (counting since the reform in 1986), coffee has been one of the most important contributors to the revenue of Vietnam's Agriculture in particular and to the entire national GDP in general. The Coffee Industry has created thousands of direct and indirect jobs, and is the main livelihood of many

households in agricultural production areas. The value of coffee exports usually accounts for about 15% of total agricultural exports and the proportion of coffee has always exceeded 10% of agricultural GDP in recent years.

In recent years, Vietnamese coffee has developed strongly with export orientation and has now become the second largest producer and exporter in the world. Vietnam coffee has exported to more than 80 countries and territories, accounting for 14.2% of the global coffee export market share (after Brazil) and 9.1% of the market share (ranked 5th; after Brazil, Indonesia, Malaysia, and India), creating many opportunities and prospects for Vietnam's coffee industry to penetrate deeper into international markets, through signed free trade agreements. The EU is Vietnam's largest coffee market—accounting for 40% of the total quantity and 38% of the total export turnover of the country, followed by Southeast Asia—accounting for 13% of the total volume and total turnover. However, Vietnamese coffee is facing many challenges from the climate to competing with other crops and production costs are rising higher while world coffee prices are at very low levels. In order to effectively exploit the strengths of coffee exports to limit the risks in the coffee bean business, the ARIMA model is built with the purpose of forecasting Vietnam's coffee export output in the coming time, helping producers monitor to come up with export plans in accordance with the output situation obtained during the production process of the year.

2. Literature Review and Research Gap

In fact, in recent years, research using the ARIMA model often focuses on economic issues such as disbursement, stock price forecasts, unemployment rate forecasts, and crop output forecasts, planting, forecasting export value, etc. The work of Ahmad Farooqi [1] builds an ARIMA model following the approach of Box and Jenkins for the total annual imports and exports of Pakistan from 1947 to 2013. Research by Bui Thi Minh Nguyet *et al.* [2] uses the ARIMA model to forecast Vietnam's export value in the last 6 months of 2018 using Eviews software. Experimental research results show that the most suitable model to forecast Vietnam's export value with data from January 2010 to June 2018 is ARIMA(1, 1, 16). The model results have predictive value with a fairly small error level compared to reality. The work of Joseph Lwaho and Bahati Ilembo [3] sets out to develop a model for forecasting maize yields in Tanzania using the ARIMA model. Research by Nguyen Thi Hien *et al.* [4] uses the ARIMA model combined with data collected in the period from July 2009 to January 2023 to forecast inflation in Vietnam in the first half of 2023. The study of Omkar Poudel *et al.* [5] aims to determine the model. ARIMA model and National Consumer Price Index (NCPI) forecast for Nepal, using annual data from fiscal years 1972-1973 to 2022-2023. Ramakrishna and Vijaya [6] used ARIMA model for forecasting of rice production in India. Author Vo Van Tai [7] used different mathematical models of regression and ARIMA time series to forecast Vietnam's rice output. Depending on the list of data statistics, forecasting methods have advantages and disadvan-

tages. The above scientific works all develop methods of time series theory introduced in the book of Box and Jenkins [8] and the book of Brockwell and Davis [9].

Each study involves the application of statistical techniques, particularly time series analysis, to analyze historical data and make forecasts or predictions for future trends in the respective economic indicators. However, very few studies related to the agricultural sector in general and the coffee import-export sector in particular. Not to mention that this is such an important field contributing to the development of the Vietnamese economy. If there are research articles on coffee export, it was a long time ago and the forecast time is limited. Aware of a crucial need, we decide to develop the topic: "Building ARIMA model for forecasting Vietnam's coffee export output".

3. Methodology

3.1. Data Collection Methodology

To gather data on Vietnam's coffee export volume from 2000 to 2023, we conducted a meticulous search process on the Internet. This involved a modeling procedure executed through the utilization of various online data sources. Specifically, we extracted data from reputable sources such as the Journal of Industry and Trade and the website of the Tuoi Tre Weekend newspaper. The process commenced with establishing a set of search criteria and proceeded with a thorough analysis of the results obtained from these sources.

During the search process, a range of relevant articles and documents were examined to ensure comprehensive and reliable data collection. Data sources were chosen based on their credibility and quality content, and all collected information underwent rigorous verification to ensure the highest level of accuracy and reliability.

The data used in the article is Vietnamese coffee export data for the period 2000-2023 compiled from the reports "Coffee exports" and "number 1 in the world" and "Coffee export challenges in Vietnam currently" website of the Ministry of Industry and Trade (VietNam).

3.2. Data Processing

Following data collection, the gathered data on coffee export volume underwent a meticulous processing procedure. This included verification, cleansing, and standardization of the data to eliminate any inaccuracies, deficiencies, or unreliable data points. Once the data was cleaned, we proceeded to organize and classify it by year, generating a structured dataset that is easily accessible for subsequent analysis and comparison.

3.3. Model Specification

ARIMA is a renowned family of time-series models that originated for its usage in economics. This family of models, capable of predicting future points in a time

series dataset, is appreciated for their statistical traits, their capacity to implement a range of exponential smoothing models, and the integration of the Box-Jenkins method during the model training phase.

An autoregressive model is a model where the dependent variable is regressed on at least one lagged period of itself. If an autoregressive model includes one lagged period of itself, it follows a first-order autoregressive stochastic process, denoted AR(1). Furthermore, if the model includes p number of lagged periods of the dependent variable, it follows a pth-order autoregressive process, denoted AR(p).

An autoregressive process may be used to forecast a time series. As mentioned earlier, a first-order autoregressive model is denoted AR(1) and is Y_t regressed on Y_{t-1}. An autoregressive model of the pth-order is denoted AR(p) and takes the form of:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \delta + \varepsilon_t \tag{1}$$

where the constant is denoted by δ and ε_t is white noise [9].

In a moving average process, the dependent variable is regressed on current and lagged error terms and is therefore estimated through a constant and a moving average of the error terms. If the dependent variable is regressed on the current and one lagged error term, it follows a first-order moving average process, denoted MA(1). Moreover, a model that includes q number of error terms follows a qth-order moving average process, denoted MA(q). A MA(q) process is defined as:

$$Y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{2}$$

where the error terms ε are assumed to be white noise and μ is the constant [9].

The ARIMA model is denoted ARIMA(p, d, q), which means that the AR is of the pth-order, the time series is integrated d number of times, and the moving average is of the qth-order. It is important to note that an ARIMA model is not derived from any economic theory, that is, it is a theoretic model. The Box-Jenkins methodology can be followed to determine p, d, and q and estimate an ARIMA model [9], and this model is defined as:

$$Y_t = \varphi_1 Y_{t-1} + \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{3}$$

The Box-Jenkins methodology consists of four consecutive steps that should be followed when building an ARIMA model.

The first step is called *identification*, and the purpose of this step is to determine appropriate values for p, d, and q. The ACF and the Partial Autocorrelation Function (PACF) with their respective correlograms are used for pattern detection of p, d, and q in the first step. The PACF measures the autocorrelation between observations in a time series that are separated by k number of lags and the intermediate autocorrelation between the lags are held constant. The choice of the AR(p) model depends on the PACF graph if it has a high value at lags 1, 2, ..., p and decreases suddenly afterwards, and the PACF functional form gradually disappears. Similarly, choosing the MA(q) model is based on the ACF graph if

it has high values at lags 1, 2, ..., q and decreases sharply after q, and the PACF functional form gradually disappears.

Estimation of the parameters in the model is the second step. The parameters of the ARIMA model will be evaluated according to the least squares method.

Step three is *diagnostic checking*, which tests the chosen ARIMA model's goodness of fit, usually done by testing if the residuals are white noise. In the case of residuals that are not white noise, step one, two, and three should be repeated using new values for p, d, and q. However, if the residuals are white noise, the model should be accepted and it is possible to proceed to step four.

Forecasting is the fourth step where the model may be used to predict desired periods for the time series [9].

4. Results and Discussions

4.1. Abbreviations and Acronyms

In mathematics, stationarity is used as a tool in time series analysis. To develop a statistically meaningful model, it is necessary to check the stationarity of the time series data beforehand. A process is said to be stationary if it is a random process, characterized by a constant mean and variance of errors over time. In reality, most economic time series (raw series) are non-stationary. This implies that these time series have changing means and variances over time. However, when differencing is applied, time series often become stationary.

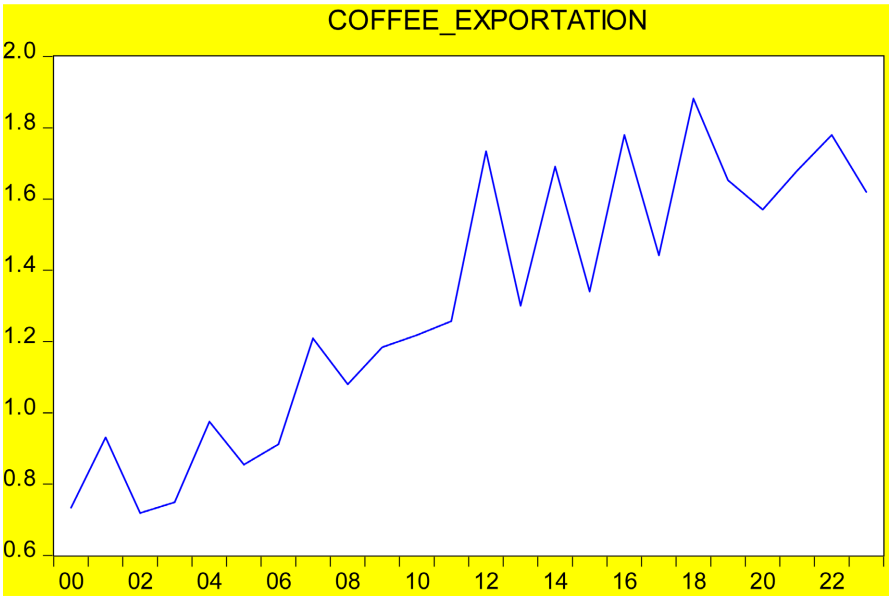
The time series used in the ARMA model are assumed to be stationary. Therefore, to forecast the number of international visitors to Vietnam using the ARMA model, we need to consider whether these series are stationary. To affirm this, we can first rely on direct observation of the graph of the time series data and then proceed to test it.

Figure 1 illustrates the fluctuations in the quantity of coffee exportation from Vietnam over the years (from 2000 to 2023), showing instability and an upward trend. Specifically, its mean tends to increase over each period. Therefore, it can be inferred that the coffee exportation quantity series is non-stationary. However, upon taking the first difference of this series, we obtain a new series, the variation in the coffee production across months (abbreviated as COFFEE-EXPORTATION), which does not exhibit a clear trend and revolves around a certain mean value (**Figure 2**). This is indicative of a stationary series.

To confirm these hypotheses, the ADF and PP tests are employed to examine the consistency of the results. We use two different tests for unit roots: the Dickey-Fuller (DF) (1979, 1981) test and the Phillips-Perron (PP) (1988) test. The Augmented Dickey-Fuller (ADF) test is based on the following regression equation:

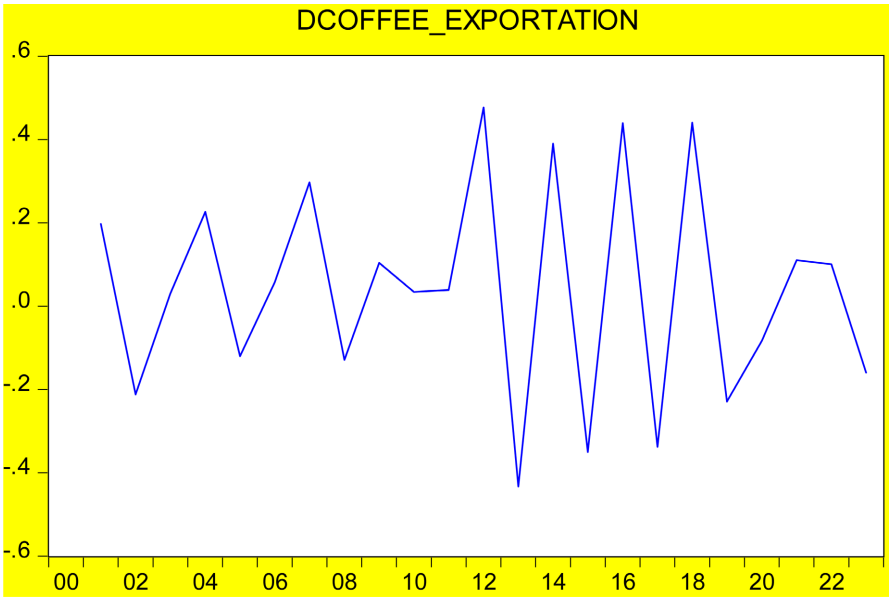
$$\Delta x = \alpha_0 + \varnothing x_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta x_{t-1} + e_t, \tag{4}$$

where x_t is the variable tested for unit root; Δ is the first difference operator; α is the constant; and p is the number of lags included to avoid the problem of autocorrelation in the residuals. The lag length in the ADF regression is selected



Source: From data and calculations on Eviews software by the authors.

Figure 1. Evolution of Vietnam's coffee export from 2000 to 2023.



Source: From data and calculations on Eviews software by the authors.

Figure 2. Fluctuations in Vietnam's coffee export volume from 2000 to 2023.

based on the minimum SBC. The null hypothesis in the ADF tests is that the series (which should be in level form) is nonstationary, *i.e.* it contains unit roots. To reject the null, the calculated test value has to be greater than the critical value. [9] is an alternative to the ADF test. It controls for serial correlation when testing for unit root and is based on the non-augmented DF-test equation. The key focus of this method is on modifying the t-ratio so that serial correlation does not affect the asymptotic distribution of the test statistic.

For the original series, both the ADF and PP tests show test statistics exceed-

ing critical values at the 1%, 5%, and 10% significance levels. Therefore, the null hypothesis H_o (the original series is non-stationary) cannot be rejected, indicating that the COFFEE_EXPORTATION series is non-stationary. This is understandable as original time series data often exhibit instability.

However, for the first differenced series, the null hypothesis H_o is rejected as both the ADF and PP tests yield test statistics smaller than critical values at the 1%, 5%, and 10% significance levels. Thus, both testing methods are consistent, and it can be concluded that the first differenced series or DCOFFEE_EXPORTATION is stationary.

4.2. Building ARIMA Model for the Chain of Fluctuations in Vietnam's Coffee Export Volume

To build the ARIMA model, we used a series of 20 observations from 2000 to 2023. The past data is named COFFEE_EXPORTATION and then takes the most classified error of COFFEE_EXPORTATION, characterized as DCOFFEE_EXPORTATION.

Step 1: *Identification* (*Determining the Trials p*, *d*, *q*)

The COFFEE_EXPORTATION data chain tested above shows that this chain stops at the wrong scale 1, we have $d = 1$.

To determine *p*, [8] introduced the following method of identification: a sequence of stops of self-correlation of the p-level if: 1) the self-correlation coefficient decreases slowly in the form of a hood or a synchronous shape, 2) the individual part-coefficients decrease abruptly by zero meaningful immediately after the delay of *p*.

From autocorrelation diagram and the separated correlation of the DCOFFEE_EXPORTATION sequence showing the existence of another coefficient 0, meaning at delay 1, in which after delay, a separated cofficient suddenly decreases in value equal to a meaningful zero. As a circle, *p* has a value of 1. Similar to defining *p*, observing the autocorrelation chart and the separate correlation of the DCOFFEE_EXPORTATION series, we find that *q* can carry one of the values: 1, 2. So, we have the ARIMA model (*p*, 1, *q*) with the combination of *p* and *q* found:

$$p = 1 \quad \text{and} \quad q \in \{1, 2\}$$

Step 2: *Estimate Model Quantity*

To estimate the coefficients of the ARIMA(*p*, 1, *q*) models as identified above, the Eviews software was used.

Step 3: *Check*

To test the suitability of the models we are based on the AIC standard as small as possible and the significant variables. After testing the ARIMA models, the statistical results are summarized in **Table 1**.

Tested, compared several models and found the ARIMA(1, 1, 1) model to be the most suitable. The model estimate results are presented in **Table 2**.

Set Z_t = DCOFFEE_EXPORTATION.

We have:

$$Z_t = 0.042621 - 0.641698 Z_{t-1} - 0.482246 U_{t-1} + U_t \quad (5)$$

Step 4: *Forecast*

The short-term forecasts of Vietnam's coffee export production based on the ARIMA model (1, 1, 1) are presented in Table 3.

Table 3 shows that the forecast of Vietnam's coffee export production in 2020 is close to reality. This suggests that this ARIMA model (1, 1, 1) explains the volatility of Vietnam's coffee export production. However, the forecast for the next points is greater in error, which is why it is necessary to update the data regularly in order to give more realistic forecasts.

5. Conclusions

The fluctuation in coffee export volume in Vietnam follows a time series that adheres to an integrated Autoregressive Moving Average (ARIMA) process with a

Table 1. Statistical results of a number of standardized ARIMA test models.

ARIMA model (p, d, q)	AIC	Adjustable correlation factor (R^2)	Significant variables
(1, 1, 1)	−0.649434	0.633517	3
(1, 1, 2)	−0.576212	0.603664	2

Source: From data and calculations on Eviews software by the authors.

Table 2. ARIMA model estimate result (1, 1, 1) for the COFFEE_EXPORTATION series.

Variable	Coefficient	Std. error	t-Statistic	Prob.
C	0.042621	0.011252	3.787949	0.0012
AR(1)	−0.641698	0.207181	−3.097276	0.0057
MA(1)	−0.482246	0.242750	−1.986593	0.0608

Source: From data and calculations on Eviews software by the authors.

Table 3. Vietnam's coffee export production forecast results.

Time	Forecast	Reality	Standard error
2020	1.57	1.57	0.61
2021	2.01	1.68	0.61
2022	1.78	1.78	0.61
2023	1.70	1.62	0.61
2024	1.81		0.61
2025	1.91		0.61
2026	1.75		0.61

Source: From data and calculations on Eviews software by the authors.

lag of 1, specifically characterized as ARIMA(1, 1, 1). Through the research, we observe that the fluctuation in coffee export volume annually is unstable and tends to increase. And through model validation, we can conclude that the first-order differenced series of COFFEE_EXPORTATION is a stationary series. The forecasted data for Vietnam's coffee export volume in 2020 closely aligns with reality. This indicates that the ARIMA(1, 1, 1) model has effectively captured the variability of Vietnam's coffee export volume. Utilizing this model allows us to provide short-term forecasts for Vietnam's coffee export volume to global markets in the coming years. Forecast results for the years 2024-2026 reveal a margin of error ranging from 0.61 units, which is not surprising given the economic volatility during that period. However, it is essential to note that while the ARIMA model serves as a forecasting tool, it is not yet optimized due to the assumed linearity in model dependency.

Therefore, the forecast results from the ARIMA(1, 1, 1) model for coffee export volume from 2024 to 2026 can serve as supplementary reference material for management units, policy planners, and businesses seeking guidance on directions, strategies, and proposed solutions for the development of coffee-producing regions nationwide. This includes a particular emphasis on the application of advanced technology in production to enhance productivity and contribute to sustainable development in the future.

In the upcoming years, there is a need to intensify exports to key markets such as Germany, the United States, Italy, Spain, and Japan. Leading up to 2030, diversifying coffee export products (Robusta, Arabica, Excelsa...) and prioritizing high-value processed coffee products will be crucial. Building Vietnam's coffee export brand and reducing intermediary exports are also important goals. Additionally, efforts should focus on increasing exports to potential markets such as Russia, Belgium, Algeria, the United Kingdom, Malaysia, Thailand, France, South Korea, India, and markets with existing Free Trade Agreements. Exploring new markets aligns with the target of achieving a USD 6 billion coffee export goal by 2030.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Farooqi, A. (2014) ARIMA Model Building and Forecasting on Imports and Exports of Pakistan. *The Pakistan Journal of Statistics and Operation Research*, **10**, 157-168. https://doi.org/10.18187/pjsor.v10i2.732

[2] Nguyet, B.T.M., Cham, N.T.Q. and Nga, N.T.Q. (2019) Sử dụng mô hình ARIMA trong dự báo giá trị xuất khẩu của Việt Nam. *Tạp chí Nghiên cứu Tài chính Kế toán*, **1**, 58-65.

[3] Lwaho, J. and Ilembo, B. (2023) Unfolding the Potential of the ARIMA Model in Forecasting Maize Production in Tanzania. *Business Analyst Journal*, **44**, 128-139.

https://doi.org/10.1108/BAJ-07-2023-0055

[4] Hien, N.T., Trang, L.M., Vu, P.L., Hoang, P.V.D. and Huy, L.Q. (2023) Ứng dụng mô hình ARIMA để dự báo lạm phát của Việt Nam và một số khuyến nghị. *Tạp chí Ngân hàng*, **2023**, 65-76.

[5] Poudel, O., Kharel, K.R., Acharya, P., Simkhada, D. and Kafle, S.C. (2024) ARIMA Modeling and Forecasting of National Consumer Price Index in Nepal. *Interdisciplinary Journal of Management and Social Sciences*, **5**, 105-118. https://doi.org/10.3126/ijmss.v5i1.62666

[6] Ramakrishna, G. and Vijaya, K.R. (2017) ARIMA Model for Forecasting of Rice Production in India by Sas. *International Journal of Applied Mathematics & Statistical Sciences* (*IJAMSS*), **6**, 67-72.

[7] Tai, V.V. (2012) Dự báo sản lượng lúa Việt Nam bằng các mô hình toán học. *Tạp chí Khoa học*, *Trường Đại học Cần Thơ*, **23b**, 125-134.

[8] Box, G.E.P. and Jenkins, G. (1970) Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco.

[9] Brockwell, P.J. and Davis, R.A. (2002) Introduction to Time Series and Forecasting. 2nd Edition, Springer, New York. https://doi.org/10.1007/b97391