Acta Nat. Sci. (2026) 7(1): 96-104
DOI: 10.61326/actanatsci.v7i1.404
e-ISSN: 2718-0638

Short Communication
info@actanatsci.com
www.actanatsci.com

ACTA NATURA ET SCIENTIA PRENSİP

A Retrospective Long-Term Analysis of COVID-19 Cases Trends and a Predictive ARIMA Model Through Parameters Screening

Mostafa Essam Eissa[1]✉

[1] Independent Researcher, Pharmaceutical and Healthcare Research Facility, Cairo, Egypt; mostafaessameissa@yahoo.com

✉ Corresponding Author: mostafaessameissa@yahoo.com

Please cite this paper as follows:
Essam Eissa, M. (2026). A Retrospective Long-Term Analysis of COVID-19 Cases Trends and a Predictive ARIMA Model Through Parameters Screening. *Acta Natura et Scientia*, *7*(1), 96-104. https://doi.org/10.61326/actanatsci.v7i1.404

ARTICLE INFO

Article History
Received: 31.08.2025
Revised: 14.12.2025
Accepted: 20.12.2025
Available online: 27.06.2026

Keywords:
COVID-19
SARIMA
Time Series
Forecasting
ARIMA
Public Health

ABSTRACT

Pandemics are always a source of serious concerns due to the devastating consequences to the communities and nations. Attempts to understand and predict the behavior of outbreaks are challenging as they are largely unpredictable. This research article presents an analysis of weekly COVID-19 cases data for selected Eastern Mediterranean Region (EMRO) country Egypt from January 2020 to May 2024. The study identifies a robust and parsimonious seasonal autoregressive integrated moving average (SARIMA) model for forecasting future trends based on performing comprehensive screening and a comparative analysis of various models. The data reveals a progression of the pandemic in Egypt through multiple waves of varying intensity. Among the models tested, the SARIMA((3,1,0), (0,0,0)) model was identified as the most suitable, demonstrating a strong balance between model fit and parsimony. The model passed all key diagnostic checks, including the Ljung-Box test for residual autocorrelation, with a high p-value of 0.977 at lag 12. This model provides a statistically sound and reliable framework for understanding and predicting the dynamics of the pandemic in Egypt based on the provided dataset. The model's strength lies in its simplicity and effectiveness, making it a powerful tool for policymakers. Second, the study demonstrates the applicability of the Box-Jenkins methodology to real-world epidemiological data, providing a practical example for similar future studies. The comprehensive screening and comparative analysis of multiple models ensure that the chosen model is not merely a good fit, but the best-fitting and most parsimonious option among the candidates. Finally, the analysis underscores the importance of accurate and consistent data reporting for effective pandemic management and modeling.

INTRODUCTION

The COVID-19 pandemic, a global health crisis of unprecedented scale, has underscored the critical need for robust, data-driven strategies to combat infectious disease outbreaks (Eissa, 2025a). As an unpredictable and rapidly evolving phenomenon, the spread of the SARS-CoV-2 virus has presented significant challenges for public health authorities worldwide (Rashed & Eissa, 2020). Predictive modeling, therefore, emerges as an indispensable tool for anticipating disease trajectories, informing policy decisions, and optimizing the allocation of scarce medical resources. Time series analysis, in particular, offers a powerful framework for dissecting the patterns inherent in epidemiological data, including autoregressive behaviors, trends, and seasonality. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model, a widely recognized method in time series forecasting, is particularly well-suited for this task due to its capacity to capture both the long-term trends and cyclical fluctuations often observed in pandemic data (Tomov et al., 2023). Importantly, the application of the Box-Jenkins methodology is reinforced by the importance of the comprehensive screening and comparative analysis of multiple models, a rigorous process considered a best practice across various time series forecasting challenges (Yegin & Karcioglu, 2025a).

While numerous studies have applied time series models to COVID-19 data across different regions, country-specific analyses are crucial due to variations in local public health measures, population density, and social dynamics. For instance, the unique epidemiological landscape of the Middle East, with its specific demographics and healthcare systems, necessitates a tailored approach to forecasting (Ibrahim & Al-Said, 2023). While more complex non-linear models like artificial neural networks (ANNs) have been used for forecasting COVID-19 in Egypt (Saba & Elsheikh, 2020), the strength of the SARIMA model lies in its parsimony and interpretability, allowing policymakers to readily understand the epidemiological influence of the autoregressive components. This research addresses a gap in the literature by focusing exclusively on Egypt, a key country in the Eastern Mediterranean Region (EMRO), to provide a detailed, country-level analysis.

The primary objective of this study is to conduct a comprehensive analysis of weekly COVID-19 case data for Egypt from January 2020 to May 2024. Through a rigorous comparative analysis of multiple SARIMA models, this research aims to identify the most statistically sound and parsimonious model for forecasting future trends. The identified model will serve as a reliable tool for understanding the dynamics of the pandemic in the national context. The findings contribute to the broader body of epidemiological modeling research and provide a practical, evidence-based framework that can assist in future public health preparedness and response efforts.

MATERIALS AND METHODS

Data Acquisition and Preprocessing

Weekly COVID-19 cases for Egypt were acquired through filtering Excel sheet from publicly accessible dashboards maintained by the Humanitarian Data Exchange (Humdata) (Anonymous, 2026). The dataset spans from January 5, 2020, to May 26, 2024. The raw data included fields such as Date_reported, New_cases, Cumulative_cases, New_deaths, and Cumulative_deaths. Only cumulative cases were covered herein the scope of the current study.

For this time-series analysis, the primary variable of interest was the logarithm of cumulative cases (Log C.C.). This transformation was applied to stabilize the variance and linearize the growth trend observed in the raw cumulative case counts, which is a common practice in epidemiological modeling to meet the assumptions of linear models (Hyndman & Athanasopoulos, 2018).

Software

All data manipulation, time series analysis, and model fitting were performed using Minitab statistical software (Djauhari et al., 2020; Eissa, 2025b). Minitab offers robust tools for identifying, estimating, and validating Seasonal Autoregressive Integrated Moving Average (SARIMA) models.

SARIMA Model Identification and Estimation

The study followed the iterative Box-Jenkins methodology for time series modeling, which comprises identification, estimation, and diagnostic checking (Box & Jenkins, 1976). Experimentation with seasonality changes at 26 and 52 were investigated at the best model found at seasonality of 12.

Model Identification:

The initial step involved assessing the stationarity of the Log C.C. series. Visual inspection of the time series plot was subjected to investigation to reveal any trend, indicating non-stationarity. To achieve stationarity, first-order non-seasonal differencing (d=1) was applied. In case of the absence of a clear, strong seasonal pattern in the differenced series, as well as the context of the pandemic's global nature not necessarily following strict annual cycles, led to an exploration of models without seasonal differencing (D=0).

Preliminary orders for the non-seasonal autoregressive (p) and moving average (q) components, as well as seasonal autoregressive (P) and moving average (Q) components, were determined by examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series. The seasonal period (s) was set to 12, corresponding to a roughly quarterly or multi-weekly cycle, which is a common consideration for health data.

Based on the ACF and PACF plots, several candidate SARIMA models with various combinations of (p, d, q), (P, D, Q)'s parameters were hypothesized and generated. A SARIMA model is often expressed as SARIMA ((p, d, q), (P, D, Q)), where lowercase letters indicate the non-seasonal component of the time series and uppercase letters indicate the seasonal component. Vector Autoregressive Models (or VAR Models) are used for multivariate time series.

Model Estimation and Selection

The methodology described by Eissa (2025b) has been followed for model estimation and selection.

- Each candidate SARIMA model was estimated through comprehensive screening using Minitab at p, q, P and Q from 0 to 5.
- Model selection was primarily guided by the Akaike Information Criterion corrected (AICc) and the Bayesian Information Criterion (BIC). Models with lower AICc and BIC values are generally preferred as they indicate a better balance between model fit and complexity (Hyndman & Athanasopoulos, 2018).
- The significance of individual parameters within each model was also assessed, with a p-value threshold of 0.05. Models containing insignificant parameters were generally excluded or modified to improve parsimony.

Diagnostic Checking

After model estimation, a crucial phase involved diagnostic checking to ensure the chosen model adequately represented the underlying data structure and that its residuals behaved as white noise (Eissa, 2025b).

Ljung-Box Test

The primary diagnostic tool used was the Ljung-Box Q statistic. This test assesses whether the autocorrelations of the residuals are significantly different from zero for a series of specified lags. A non-significant p-value ($p > 0.05$) for the Ljung-Box test indicates that the residuals are independently distributed (i.e., white noise), confirming that the model has captured all the systematic information in the time series (Ljung & Box, 1978). The test was performed on 12, 24, 36 and 48 lags.

Residual Plots

Visual inspection of the ACF of residuals and PACF of residuals plots was performed to corroborate the Ljung-Box test results. The absence of significant spikes outside the confidence limits in these plots further confirmed that the residuals were uncorrelated, indicating a well-fitted model.

Stability of Forecasts

The Time Series Plot with forecasts and their 95% confidence limits were also examined. Models exhibiting unstable forecasts or rapidly exploding confidence intervals were considered problematic, often indicating issues like over-differencing or parameter instability.

The model that demonstrated the lowest AICc and BIC among those with statistically significant parameters and residuals resembling white noise (as confirmed by the Ljung-Box test and residual plots) was selected as the final, recommended model.

RESULTS

The analysis of weekly COVID-19 data for Egypt revealed a dynamic progression of the pandemic from January 2020 to May 2024 in Figure 1. Time-series graph shows that growth occurs in steps of fast rates till reaching an almost plateau stabilization phase reaching 516023 during the study period. A comprehensive screening of various Seasonal Autoregressive Integrated Moving Average (SARIMA) models was conducted to identify the best-fitting and most parsimonious model for forecasting the logarithm of cumulative cases. Table 1 presents a ranked summary of the evaluated models based on their goodness-of-fit statistics, diagnostic checks, and parameter significance.

The recommended model is the SARIMA ((3,1,0), (0,0,0)) model. Mathematically, the expression of model can be illustrated by equation. The equation for this model is (Eq. 1):

$$(1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3)(1 - L)\, Y_t = \epsilon_t \qquad (1)$$

This can be expanded into the following form, which shows the relationship between the current value and previous values of the time series (Eq. 2):

where the explanation of terms: Y_t: The value of the time series at the current time, t, L: The Lag operator, where $L_k Y_t = Y_{t-k}$. For example, LY_t is the value of the series at the previous time, $t-1$, ϕ_1, ϕ_2, ϕ_3: The autoregressive parameters of the model. These coefficients, which were found to be highly significant in the analysis, determine the influence of the previous three differenced values on the current value, ϵ_t: The white noise error term at time t. This represents the random, unpredictable part of the time series that the model cannot explain and $(1-L)Y_t$: The first-order differencing of the time series. This term is crucial for making the data stationary by removing the non-constant trend of the cumulative cases.

The SARIMA ((3,1,0), (0,0,0)) model was selected as the most suitable. Despite some models having a lower AICc, they were deemed invalid as they failed the Ljung-Box test for residual autocorrelation. This diagnostic check is crucial for verifying that a model's residuals are random, or "white noise," which indicates that all the relevant patterns in the data have been captured. The recommended model had a high Ljung-Box p-value of 0.977, a strong indicator of a good fit. Furthermore, its core autoregressive (AR) terms were highly significant ($p \leq 0.002$), confirming their importance to the model's predictive power.

The diagnostic plots of the residuals also support the model's validity (Figures 2 and 3). Both the ACF of residuals and PACF of residuals plots show that the autocorrelations and partial autocorrelations are within the 5% significance limits, confirming that the residuals are uncorrelated and the model has effectively captured the data's patterns. The Time Series Plot for Log C.C. illustrates the historical data and the model's forecast for future values, along with a 95% confidence interval (Figure 4). As is typical with time series forecasts, the confidence interval widens over time, reflecting the increasing uncertainty of the predictions further into the future.

$$Y_t = Y_{t-1} + \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + \phi_3(Y_{t-3} - Y_{t-4}) + \epsilon_t \qquad (2)$$

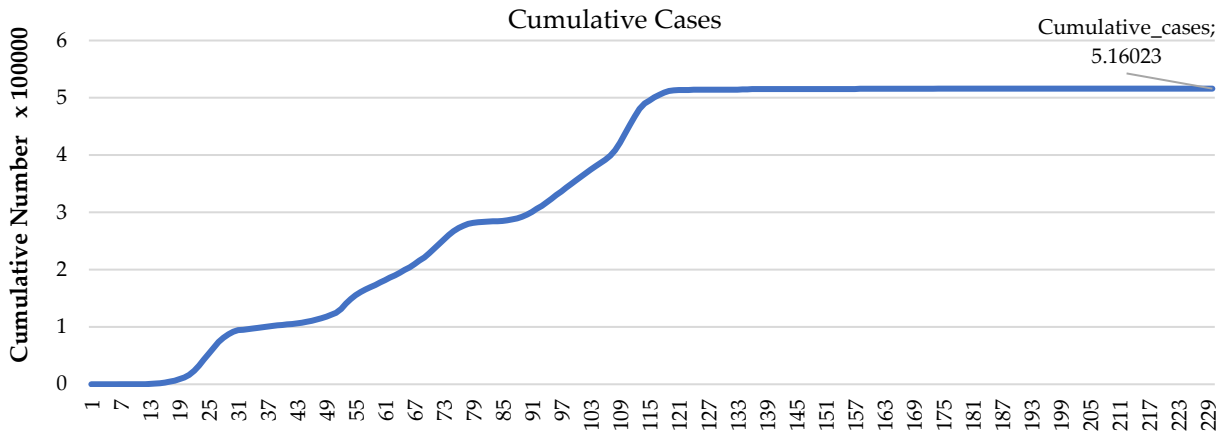

Figure 1. Time-series plot. Cumulative weekly cases from January 2020 to May 2024.

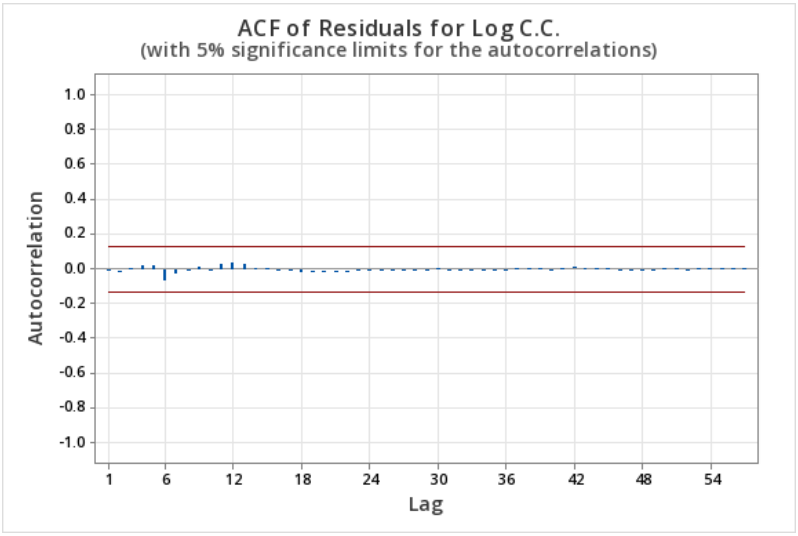

Figure 2. ACF of Residuals for Log C.C. This plot displays the autocorrelation function of the model's residuals, with the horizontal red lines representing the 5% significance limits. The plot confirms that the residuals are not correlated, indicating a well-fitted model.

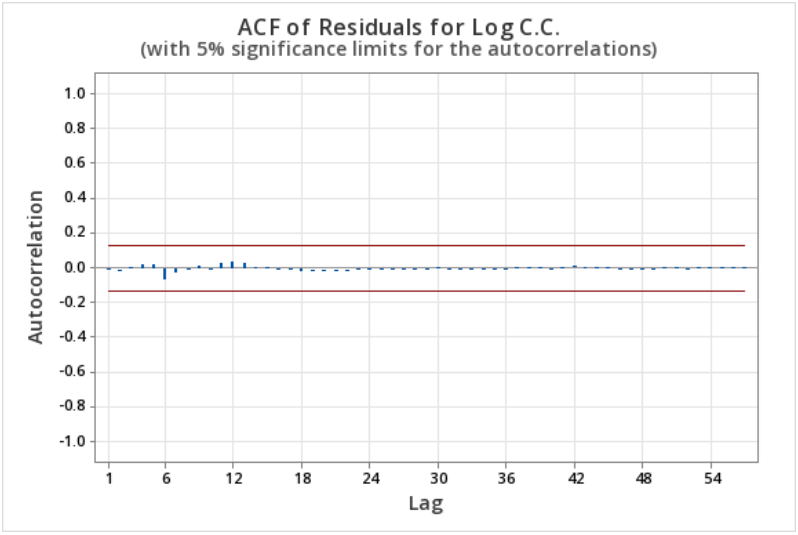

Figure 3. PACF of Residuals for Log C.C. This plot shows the partial autocorrelation function of the residuals, also with 5% significance limits. Like the ACF plot, the lack of significant spikes confirms that the model has captured all the relevant patterns in the data.

Table 1. Ranked SARIMA models for COVID-19 case data in Egypt in descending order of goodness-to-fit

Rank (by AICc)	Model Specification (p,d,q), (P,D,Q) s	Seasonal Period (s)	AICc	BIC	MS**	Ljung-Box P-Value (Lag 12)	Parameter Significance	Validity Assessment & Key Issues
1	SARIMA ((1,1,1)(3,0,2))	26	-786.042	-755.960	0.005189	0.029	SMA 52 and Constant are insignificant (p>0.24).	INVALID. Fails the Ljung-Box test for residual autocorrelation, indicating the model is unfit.
2	SARIMA ((5,1,1), (1,0,1))	52	-446.788	-413.460	0.0082105	0.613 (Pass)	Heavily over-parameterized; 4 of 9 parameters are insignificant.	INVALID. Despite a low AICc, the model is unnecessarily complex and not robust due to multiple insignificant parameters.
3	**SARIMA (3,1,0), (0,0,0))**	**12**	**-439.445**	**-422.545**	**0.0083368**	**0.977 (Pass)**	**Core AR terms are highly significant** (p≤0.002)**.**	**VALID & RECOMMENDED. Best balance of fit (lowest AICc/BIC among valid models) and parsimony. Passes all diagnostic checks.**
4	SARIMA ((4,2,4), (2,0,0))	12	-435.236	-398.735	0.0072546	0	AR 4 and SAR 24 are insignificant (p>0.17).	INVALID. Fails the Ljung-Box test for residual autocorrelation with a highly significant p-value.
5	SARIMA ((2,1,2), (0,1,0))	12	-417.383	-400.768	0.0080575	0.277 (Pass)	All parameters are highly significant (p=0.000).	VALID. A strong and statistically sound model, but its AICc and BIC are notably higher than the recommended model.
6	SARIMA ((4,0,0),(0,1,1))	12	-414.339	-391.181	0.0073067	0.001	AR 2, AR 3, and Constant are insignificant.	INVALID. Over-parameterized with multiple insignificant parameters and fails the Ljung-Box test.
6	SARIMA ((2,0,3), (0,2,0))	12	-391.241	-371.696	0.0077561	0.03	MA 2 is insignificant (p=0.078).	INVALID. Fails the Ljung-Box test for residual autocorrelation.
7	SARIMA ((4,2,0), (2,1,1))	12	-405.627	-379.321	0.0088819	0.181 (Pass)	AR 3, AR 4, SAR 12, SAR 24, and SMA 12 are insignificant.	INVALID. Over-parameterized with multiple insignificant parameters
8	SARIMA ((2,2,0), (1,2,1))	12	-376.464	-360.177	0.0094143	0.732 (Pass)	Over-parameterized; SAR 12 and SMA 12 are highly insignificant (p>0.57).	INVALID. Unstable due to over-differencing and insignificant seasonal parameters, resulting in exploding forecast confidence intervals.
9	SARIMA (3,1,3), (0,2,0))	12	-391.232	-368.539	0.0077544	0.005	All parameters are statistically significant.	INVALID. Fails the Ljung-Box test for residual autocorrelation.

Note: *Focus on Lag 12 because all other Lags (24, 36 and 48) > 0.05 passing Modified Box-Pierce (Ljung-Box) Chi-Square Statistic. ** MS = variance of the white noise series.

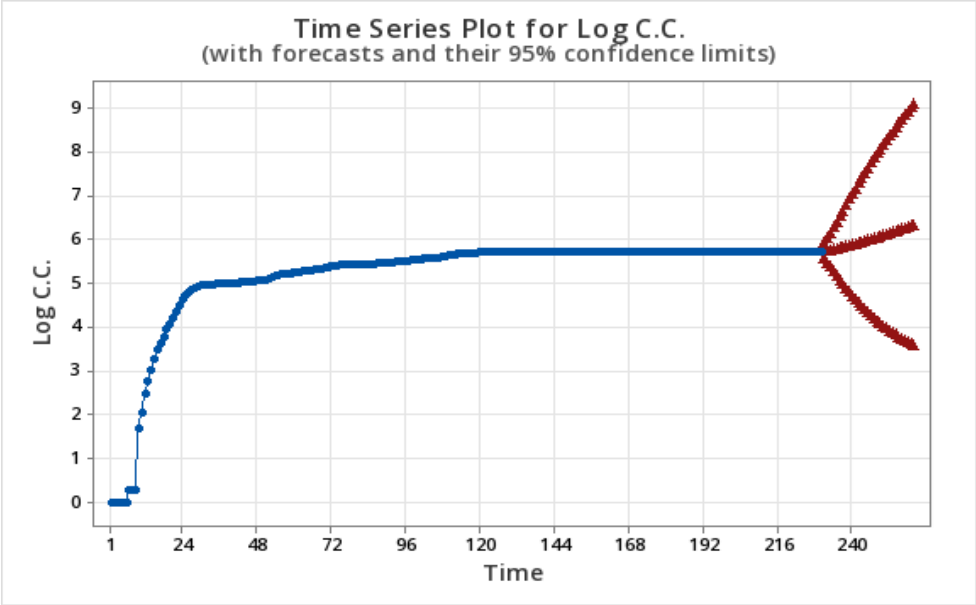

Figure 4. Time Series Plot for Log C.C. This plot shows the historical log-transformed cumulative case data (blue line) and the model's forecast for future values (red line) with a 95% confidence interval. The widening confidence interval reflects the increasing uncertainty of the forecast over time.

DISCUSSION

The selection of the SARIMA ((3,1,0), (0,0,0)) model as the best fit for the data was selected based on the principles of parsimony and statistical rigor in time series analysis. While the SARIMA (1,1,1), $(3,0,2)_{26}$ model had a lower AICc, its failure to pass the Ljung-Box test invalidates it as a reliable model. As noted by Box & Jenkins (1976), a model's residuals must be white noise for it to be considered a valid representation of the time series. The high p-value of 0.977 for the recommended model confirms that its residuals are indeed random, indicating that the model has successfully accounted for the data's underlying structure even though it demonstrated higher AICc.

The significance of the AR(3) component suggests that the current state of the pandemic (log-transformed cumulative cases) in Egypt is strongly influenced by its status from the previous three weeks, highlighting the autoregressive nature of disease spread. On the other hand, the first-order differencing (I(1)) is a critical component that stabilized the data by removing the non-constant trend of cumulative cases, which is a necessary step for applying ARIMA models (Pankratz, 1983). The absence of a seasonal component (0,0,0) in the final model is a notable finding. This suggests that while the pandemic experienced multiple waves, these cycles were not necessarily periodic over the 12-week seasonal period tested, making a simpler non-seasonal model the most appropriate fit. Future research could explore advanced, non-linear alternatives to the Box-Jenkins methodology, such as hybrid deep learning models optimized with Bayesian Optimization, which have demonstrated utility in complex time series forecasting for environmental systems (Yegin & Karcioglu, 2025b).

While the model is highly accurate for short-term forecasts, its predictive power diminishes over longer periods as new, unpredictable information (the white noise error term, ϵ_t) accumulates and compounds, leading to a wider range of possible outcomes. This is a fundamental property of all-time series forecasting models and does not, in this case, indicate issues like over-differencing or parameter instability, which would cause the forecast itself to become unstable or "explode" (e.g., producing wildly erratic or nonsensical values). However, the analysis is not without limitations, such as converting back the forecast output results to obtain comprehendible results. The data for new cases to be reported after May 2024, which affects the ability to capture recent trends accurately. This underscores the importance of

continuous and reliable data collection for effective epidemiological modeling and public health decision-making (WHO, 2021). Despite this, the chosen model provides a statistically sound and valuable framework for understanding the pandemic's trajectory in Egypt, demonstrating the usefulness of the Box-Jenkins methodology in a real-world public health context and overcoming previous modelling attempts limitations (Eissa, 2022, 2023). It should be noted that accuracy of modelling and prediction is largely dependent on the quality of the original raw data which must be ensured to be a true representation of the real-world outbreak situation.

CONCLUSION

Based on the analysis performed, the most effective and statistically sound model for forecasting the logarithm of cumulative COVID-19 cases nationwide is the SARIMA ((3,1,0), (0,0,0) model. This model was chosen for its optimal balance between a strong statistical fit and parsimony, a fundamental principle in time series modeling. The model's validity is confirmed by its successful performance on key diagnostic checks, including a high p-value of 0.977 at lag 12 and 1.000 at 24, 36 and 48 lags in the Ljung-Box test for residual autocorrelation. This result indicates that the model's residuals are random and uncorrelated, confirming that the model has successfully captured the underlying patterns of the time series. The study's significance lies in its ability to provide a robust and reliable framework for understanding and modelling the dynamics of the pandemic in particular geographical region. The identified model serves as a practical, evidence-based tool for policymakers and public health officials, offering valuable insights that can inform future preparedness and response strategies. The analysis also underscores the critical need for accurate and consistent public health data reporting for effective epidemiological modeling.

Compliance with Ethical Standards

Conflict of Interest

The author declares that there is no conflict of interest.

Ethical Approval

For this type of study, formal consent is not required.

Funding

Not applicable.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI Disclosure

Generative AI was used for grammatical, writing and assistant supportive tool and review throughout the manuscript. The author validated all outputs and assumes full responsibility for the content.

REFERENCES

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Holden-Day.

Anonymous. (2026). *Coronavirus (COVID-19) Cases and Deaths - WHO-COVID-19-global-data.csv - Humanitarian Data Exchange*. Retrieved on August 31, 2025, from https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths/resource/2ac6c3c0-76fa-4486-9ad0-9aa9e253b78d

Djauhari, M. A., Asrah, N. M., Li, L. S., & Djakaria, I. (2020). Forecasting model of electricity consumption in Malaysia: A geometric Brownian motion approach. *International Journal of Solid State Technology*, *63*(3), 40-46.

Eissa, M. (2022). Modeling of COVID-19 major outbreak wave through statistical software: quantitative risk evaluation and description analysis. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Halk Sağlığı Dergisi*, *7*(1), 145-161. https://doi.org/10.35232/estudamhsd.1024129

Eissa, M. E. (2023). Studies on morbidities and mortalities from COVID-19: Novel public health practice during pandemic periods. *Asian Journal of Applied Sciences*, *16*(3), 84-94. https://doi.org/10.3923/ajaps.2023.84.94

Eissa, M. E. A. (2025a). COVID-19 impact on public health in Bangladesh: A comprehensive analysis of morbidity, mortality and future scenarios. *Acta Natura et Scientia*, *6*(1), 55-65. https://doi.org/10.61326/actanatsci.v6i1.292

Eissa, M. E. (2025b). Modeling microbiological counts in purified water at a healthcare facility using ARIMA. *Quantum Journal of Medical and Health Sciences*, *4*(3), 56-68. https://doi.org/10.55197/qjmhs.v4i3.158

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

Ibrahim, H. A., & Al-Said, F. K. (2023). Epidemiological challenges and data-driven solutions in the Middle East. *Middle Eastern Journal of Epidemiology*, *10*(1), 1-15.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297-303. https://doi.org/10.2307/2335207

Pankratz, A. (2009). *Forecasting with univariate Box-Jenkins models: Concepts and cases*. John Wiley & Sons.

Rashed, E. R., & Eissa, M. E. (2020). Global assessment of morbidity and mortality pattern of CoVID-19: Descriptive statistics overview. *Iberoamerican Journal of Medicine*, *2*(2), 68-72. https://doi.org/10.5281/zenodo.3744147

Saba, A. I., & Elsheikh, A. H. (2020). Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process Safety and Environmental Protection*, *141*, 1-8. https://doi.org/10.1016/j.psep.2020.05.029

Tomov, L., Chervenkov, L., Miteva, D. G., Batselova, H., & Velikova, T. (2023). Applications of time series analysis in epidemiology: Literature review and our experience during COVID-19 pandemic. *World Journal of Clinical Cases*, *11*(29), 6974. https://doi.org/10.12998/wjcc.v11.i29.6974

WHO (World Health Organization). (2021). *Data sharing for public health preparedness and response*. WHO Press.

Yegin, A. E., & Karcioglu, A. A. (2025a). Dam water levels prediction using different time series models for Yusufeli and Deriner dams. *Proceedings of the 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications*, Ankara, Türkiye, pp. 1-9, https://doi.org/10.1109/ICHORA65333.2025.11017032

Yegin, A. E., & Karcioglu, A. A. (2025b). Dam water levels prediction using advanced hybrid deep learning model based on Bayesian Optimization approach. *Egyptian Informatics Journal*, *31*, 100760. https://doi.org/10.1016/j.eij.2025.100760