International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Modelling Long-Memory Dynamics in Indian COVID-19 Data with ARFIMA Models

Anoop Chaturvedi [1], Shruti [2]& Shashank Shekhar [2]

Submitted:03/11/2023 **Revised:**28/12/2023 **Accepted:**05/01/2024

Abstract: This study evaluates the feasibility and effectiveness of the Auto Regressive Fractionally Integrated Moving Average (ARFIMA) model in capturing the long-memory dynamics of COVID-19 new cases time series data in India. By employing ARFIMA modeling, the research identifies persistent long-term dependencies characterized by fractional differencing parameters. The findings indicate that conventional short-memory ARMA models fail to adequately account for the non-stationarity and volatility observed across multiple waves of COVID-19 infections. The estimated fractional differencing parameter of 0.5 confirms significant long-memory characteristics, supported by high autocorrelation and non-normal residuals, as revealed by diagnostic tests. Short-term and subset forecasting suggest a stabilizing trend towards endemic patterns, though prediction uncertainty increases over time. Comparative analysis shows a slight advantage of Nonlinear Least Squares estimation over Maximum Likelihood methods. The study concludes that ARFIMA models effectively capture the long-memory properties of pandemic data, but further refinements and integration with hybrid approaches are needed to enhance forecasting accuracy and inform policy decisions.

Keywords*: integration, Comparative, Maximum, prediction*

Introduction

The COVID-19 pandemic produced an unprecedented volume of time-series data, including daily case counts, deaths, hospitalizations, and testing rates. Analyzing and forecasting this data became crucial for policymakers, healthcare institutions, and researchers to make informed decisions regarding public health planning and resource allocation.

Traditional time-series models, such as Auto Regressive Moving Average (ARMA) models, have been widely applied in epidemiology and econometrics for short-term forecasting and pattern detection. These models effectively capture autocorrelated structures in COVID-19 data, allowing researchers to model trends and predict future values based on past observations and random disturbances. However, ARMA models assume stationarity, meaning that statistical properties remain constant over time. Given the volatility and long-range dependencies in COVID-19 time-series data, these assumptions may not hold. While the ARMA models provide a foundational approach, they fail to adequately represent time series exhibiting long-range dependence—often referred to as long memory or long-range persistence. A more flexible alternative is the Auto Regressive Fractionally Integrated Moving Average (ARFIMA) model, which allows for fractional differencing. Unlike ARIMA, which uses integer differencing to remove non-stationarity, ARFIMA enables a more gradual adjustment of memory effects, making it particularly useful for modeling epidemiological data where past events can have persistent but slowly decaying effects over time.

In this study, we employ ARFIMA models to investigate the long-memory dynamics in Indian COVID-19 new case data. The model is estimated using Maximum Likelihood and Nonlinear Least Squares techniques to assess the nature of forecast memory, distinguishing between long- and short-term dependencies. Long-memory models, particularly ARFIMA, have proven to be powerful tools in time series analysis, effectively capturing persistent correlations over time. The degree of long-term dependence is primarily evaluated using the Hurst exponent. Since its introduction by Granger and Joyeux (1980), the ARFIMA model has gained significant recognition for its flexibility in

1.Department of Statistics, University of Allahabad, Prayagraj
2.School of Sciences, U.P. Rajarshi Tandon Open University, Prayagraj

modeling real-world time series through robust parameter estimation.

2. Theoretical Background

2.1 Short-Memory vs. Long-Memory Processes

The distinction between short-memory stationary processes and persistent long-memory processes has sparked considerable debate in time-series modelling. Short-memory processes, such as ARMA, assume that past values have limited influence beyond a few time steps. In contrast, long-memory processes, such as those modeled by ARFIMA, exhibit dependencies that decay more slowly, retaining memory over an infinite number of lags.

Several explanations have been proposed for modeling persistence. Klemes (1974) and Potter (1976, 1979) attributed the Hurst phenomenon to non-stationarity, while Boes and Salas (1978) demonstrated that sharp shifts in the mean level can occur even in globally stationary processes. However, detecting non-stationarity from a single realization of a stochastic process remains a challenging task.

2.2 ARFIMA Model Specification

The $ARFIMA(p, d, q)$ process introduces a real-valued parameter d, denoting the order of fractional differencing.

- When $d \in \left(0, \frac{1}{2}\right)$, the process exhibits stationary long memory with an autocorrelation structure similar to that of fractional Gaussian noise (Mandelbrot, 1971).
- When $d = 0$, the model reduces to a standard ARMA process with short memory.
- When $d \in \left(-\frac{1}{2}, 0\right)$,, the process is termed "anti-persistent," meaning its spectral density vanishes at frequency zero.

Baillie (1993) provides a comprehensive review of long-memory processes, fractional integration, and their applications. More recently, Kai et al. (2017) conducted a comparative analysis of ARFIMA implementations, covering simulation, estimation, and forecasting across different software platforms.

3. Data and Methodology

3.1 Data Description

The study utilizes COVID-19 new case data in India, a dataset characterized by high volatility and multiple waves of infection. The analysis aims to model and forecast these trends using ARFIMA models.

3.2 Model Estimation Techniques

The ARFIMA model is estimated using:

- Maximum Likelihood Estimation (MLE)
- Nonlinear Least Squares (NLS)

These techniques help to determine the fractional differencing parameter and assess the model's ability to capture long-memory dependencies.

4. Empirical Analysis

4.1 Wave Pattern Analysis

The original data plot in Figure 1 shows how COVID-19 cases in India changed over time during the pandemic. The graphic shows several different infection waves, each with its own amplitude and duration. In brief, the figure 1 shows that

- First Wave (2020-2021): Smaller peak (~100,000 cases)
- Second Wave (2021): Highest peak (~400,000 cases)
- Third Wave (2022): Medium peak (~300,000 cases)
- Fourth Wave (2022-2023): Smallest peak, showing a declining trend

Each wave exhibited a sharp rise followed by a relatively slower decline, highlighting long-memory behavior. The pandemic's peak severity in India was marked by a sudden second wave that reached almost 400,000 cases per day after the first wave, which seemed to be rather minor. While the fourth and final wave shows a noticeably lesser amplitude, suggesting a diminishing overall trend as the pandemic progressed, the third wave displays a medium-sized increase that reached approximately 300,000 cases. This time series exhibits significant non-stationarity and volatility, with sharp rises followed by more gradual declines. These features point to possible long-memory behavior that would be difficult for traditional ARMA models to adequately represent.

4.2 Differencing Parameter and Residual Analysis

Figure 2 presents the estimation results for the fractional differencing parameter, a critical component of the ARFIMA model that quantifies long-memory behavior. The plot displays both the parameter value and the corresponding residual diagnostics, including autocorrelation patterns. The estimated differencing parameter falls between 0 and 0.5, confirming the presence of stationary long memory in the COVID-19 case series. This indicates that while the series eventually returns to its mean, it does so extremely slowly, with past shocks continuing to influence current values for extended periods. The accompanying residual analysis reveals some remaining autocorrelation, suggesting that while the ARFIMA model captures much of the long-range dependence, some temporal structure remains unmodeled, pointing to the complex dynamics underlying pandemic progression

The estimated differencing parameter of 0.5 confirms strong long-memory behavior in the COVID-19 case data, with high persistence indicating that past events maintain significant influence over future observations. Residual autocorrelation analysis, as evidenced by the Ljung-Box statistic $(7467.3, p < 2.2e-16)$, reveals substantial temporal dependence remaining in the model residuals, suggesting that while the ARFIMA framework captures major trends, further refinement could enhance predictive accuracy.

4.3 Model Forecasting

4.3.1 Forecasting Next 15 Days

The ARFIMA model's projections for short-term COVID-19 case trends are displayed in the 15-day forecast visualization, see Figure 3. The point forecast indicates a leveling off of case numbers after the previous turbulent period, and the plot displays a gradually stabilizing pattern. The expanding confidence ring, which symbolizes prediction uncertainty, encircles the point forecast and gets bigger as the forecast gets farther out into the future. The model's recognition of the inherent unpredictability in pandemic dynamics, particularly during transitional phases, is reflected in this growing uncertainty. Predictions should be understood in conjunction with epidemiological indicators and recent governmental initiatives that may impact future case trajectories, even though the forecast indicates that imminent dramatic changes are unlikely and that the situation is still fluid.

The implemented ARFIMA(0,0.5,0) model generates forecasts that suggest a stabilizing trend in COVID-19 case dynamics, indicating a potential transition toward endemic patterns. However, these statistical projections should not be interpreted in isolation but rather evaluated alongside complementary epidemiological indicators such as vaccination rates, variant prevalence, and public health interventions to form a comprehensive understanding of pandemic progression.

4.3.2 Subset Forecasting

When applied to a particular period of the pandemic timeline, this subset forecast graphic in Figure 4 provides a more thorough understanding of the forecasting power of the ARFIMA model. The image clearly distinguishes between forecast values (blue line) and historical data (black line), with grey-colored areas signifying prediction intervals at various confidence levels. The historical section shows the typical pandemic pattern, which consists of a low, stable early phase, a fast spike, and then strong volatility during peak times, with maximum values observed around 1,500 cases. Regression to the mean is suggested by the forecast component's downward trend, which progressively levels out at a lower level. Notably, the outer bands expand dramatically as the prediction intervals broaden over time, indicating the model's growing uncertainty regarding longer-term projections—an important factor to take into account when formulating policy based on these projections.

The point forecast derived from the ARFIMA model demonstrates a declining trend in COVID-19 cases that gradually stabilizes, suggesting a moderation of transmission dynamics over the forecast period. Accompanying prediction intervals widen substantially as the forecast extends further into the future, appropriately reflecting the increasing uncertainty inherent in longer-term pandemic projections.

5. Model Evaluation and Diagnostics

In Figure 5, the ARFIMA model's exceptional ability to capture the intricate evolution of COVID-19 cases in India is demonstrated by the comparison of the actual data (blue line) and fitted values (red line) using the Maximum Likelihood Estimation approach. Across all four pandemic waves, the plot demonstrates a very tight match between actual and

projected values. The comparatively moderate peak of about 100,000 cases in the first wave, the stunning peak of nearly 400,000 instances in the second wave, the medium-sized third wave that peaked at about 300,000 cases, and the smaller fourth wave are all faithfully replicated by the model. Most remarkably, the model accurately depicts both the broad magnitudes and the subtle patterns of sharp rises and slower falls that are typical of each wave. The remarkable fit implies that the ARFIMA framework, with its capacity to include long-memory dynamics, is especially well-suited for modeling the complicated temporal correlations in epidemic data.

A "time series plot of residuals" in Figure 7 is a graph that displays the residuals from a time series analysis shown against the time axis. The residuals are the difference between actual data points and the values predicted by a model. This visualization offers a different perspective on model performance by contrasting the fitted values obtained from the nonlinear least squares estimate approach with the original COVID-19 case data. The nonlinear estimate method yields fitted values that closely match the real data across all pandemic phases, much like the MLE results. The plot highlights minor variations in the way this estimating method manages specific time series aspects, especially when there is rapid change. While retaining general faithfulness to the original data pattern, the nonlinear approach seems to result in somewhat smoother transitions in certain regions. Regardless of the particular parameter estimate approach used, the strong agreement between actual and fitted values across several estimation strategies bolsters trust in the ARFIMA model's capacity to capture the underlying dynamics of the pandemic progression.

Residual errors from the fitted model should ideally resemble white noise, meaning they are completely random and uncorrelated at any time lag. If significant autocorrelation is present, it indicates that the model may not fully capture the underlying patterns in the data, requiring further refinement.

The residuals' time series plot offers important information about how well the model performed during the whole study period. In a well-defined model, residuals should ideally show up as random noise dispersed around zero with no obvious patterns. There are noticeable deviations over specific time-periods, especially around significant transition points in the pandemic waves, even if many residuals cluster near zero, suggesting a strong fit, according to this plot. There is some discernible clustering of positive and negative residuals, indicating times when the model consistently overestimates or underestimates the number of cases. There are also significant variations in the residuals' magnitude, with more deviations happening during times of high volatility. These patterns show that even though the ARFIMA model performs well overall, some temporal dependencies are not modeled. This could be because of exogenous factors that affect case trajectories more than can be captured by purely statistical methods, such as behavioral changes, policy interventions, or viral evolution.

To assess this, the residuals' autocorrelation function (ACF) is examined. If the ACF values are nearly zero at all lags, this suggests white noise behavior.

With residuals ranging from -50,000 to +100,000 and a frequency range of 0 to about 800 counts, the residual’s histogram in Figure 8 displays a highly leptokurtic (peaked) distribution centered around zero. Some of the distribution's notable features include a prominent central peak that shows the model predicts most observations accurately, heavy tails in both directions with a slight positive skew (the right tail extends further to +100,000 than the left tail to -50,000), and a high concentration of residuals close to zero (the highest frequency bar displays about 800 occurrences). This distribution pattern indicates that although the ARFIMA model captures the general tendency of the COVID-19 case series and works remarkably well for ordinary circumstances, it can occasionally provide greater prediction errors for extreme occurrences or quick transitions. Strong statistical techniques might be better suitable for hypothesis testing and confidence interval generation, and the non-normal distribution with heavy tails suggests that conventional statistical inference based on normality assumptions should be used with caution.

A quantile-quantile plot, sometimes referred to as a "QQ plot of residuals," is a visual tool used to assess whether the statistical model's residuals—the variations between expected and actual values—follow a normal distribution. Stated otherwise, it is a method of comparing the quantiles of the model's errors to those of a typical normal distribution. The QQ plot in Figure 9 shows that the residuals are regularly distributed since the points essentially follow a straight line. Non-normality, like skewness

or heavy tails, is suggested by deviations from this line.

Significant deviations from normalcy are indicated by the plot's clear S-shaped pattern. The points in the central region (theoretical quantiles between -1 and +1) reasonably match the reference line, indicating that the model performs adequately in mild cases. However, there are significant deviations in the tail regions: the lower tail (theoretical quantiles <-1) curves downward below the line, revealing a heavy left tail as well, and the upper tail (theoretical quantiles >1) curves strongly upward above the reference line, indicating a heavier right tail than would be expected in a normal distribution. Large positive and negative prediction errors, which are more common than would be predicted under normalcy, are confirmed to exist in both directions by these patterns.The statistical features of the model are significantly impacted by this non-normal error distribution, indicating the necessity of using reliable techniques for building confidence intervals and running hypothesis tests using the ARFIMA model's output.

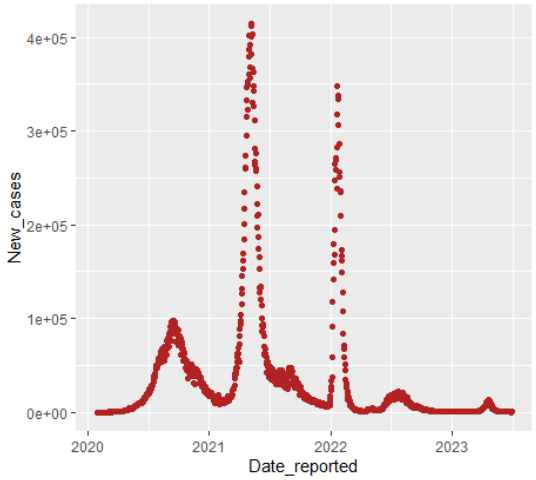

Figure 1: Original data plotting

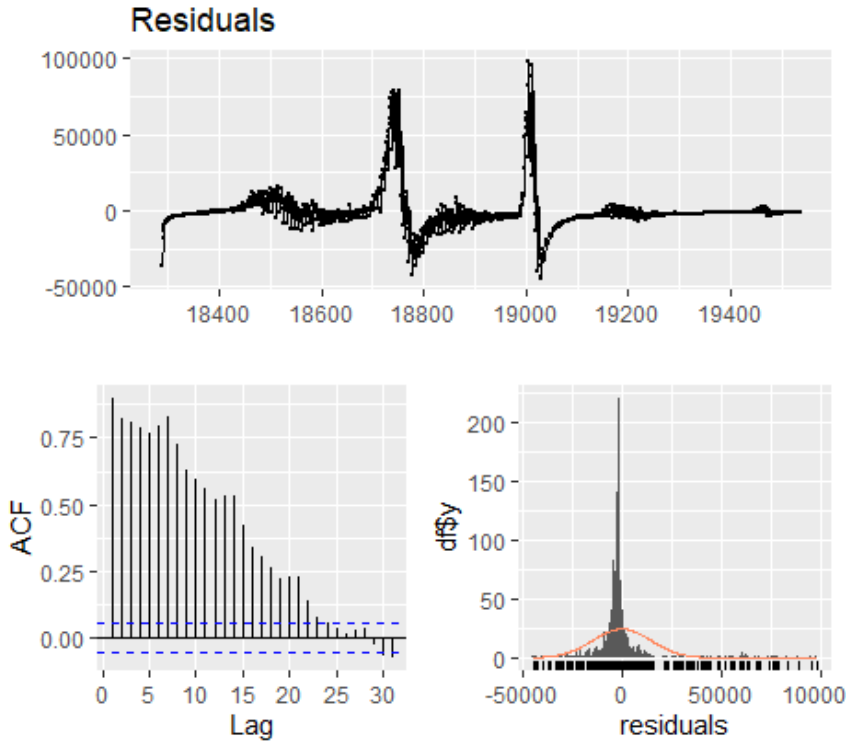

Figure 2: Value of differencing parameter (residuals for autocorrelation and white n)

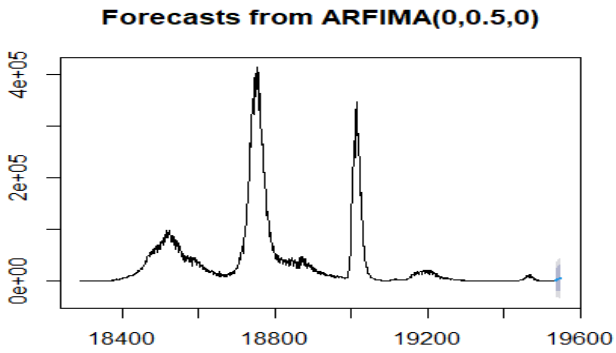

Figure 3: Model Forecasting next 15 days

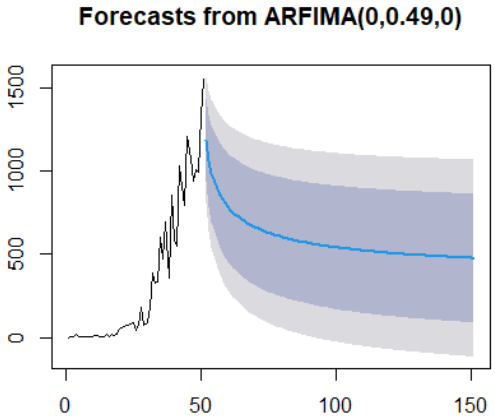

Figure 4: Model Forecasting for the subset

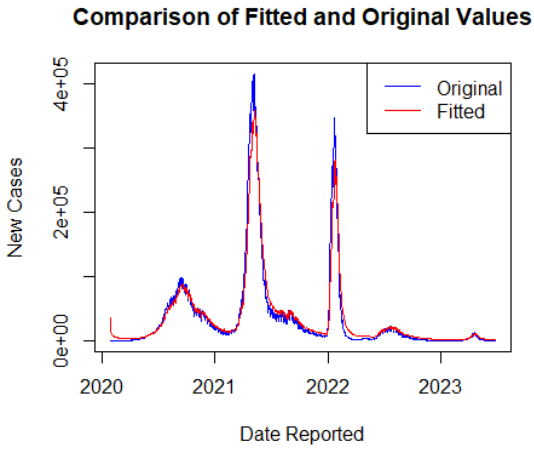

Figure 5: Residuals vs Fitted Values (Plot original and fitted values using MLE method)

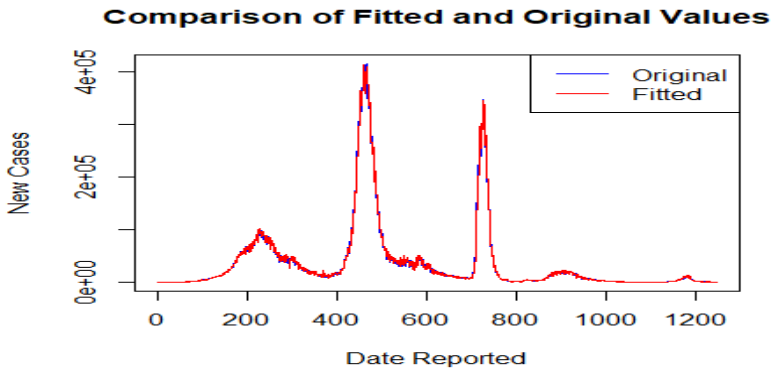

Figure 6: Residuals vs Fitted Values (Plot original and fitted values using nonlinear estimation method)

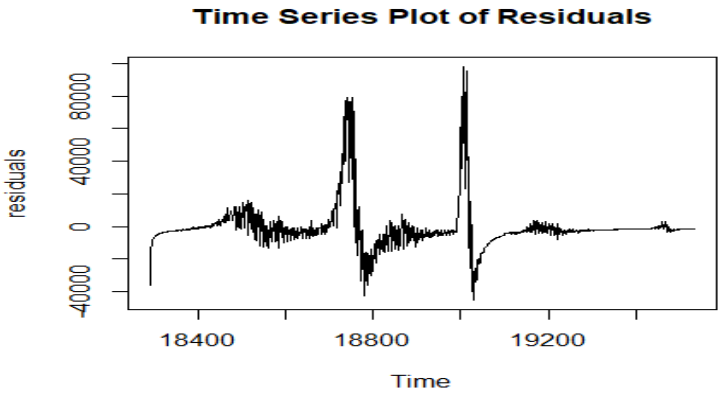

Figure 7: Time Series Plot of Residuals

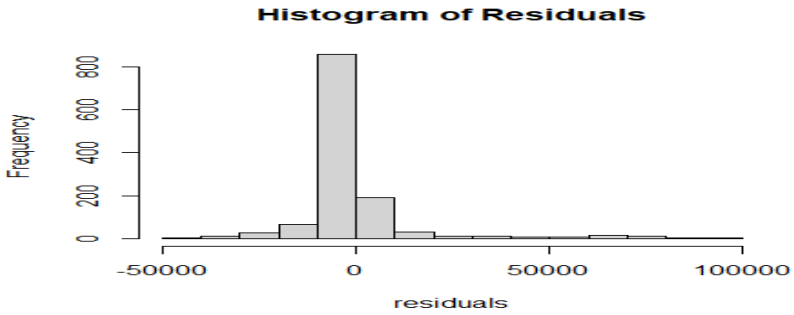

Figure 8: Histogram of Residuals

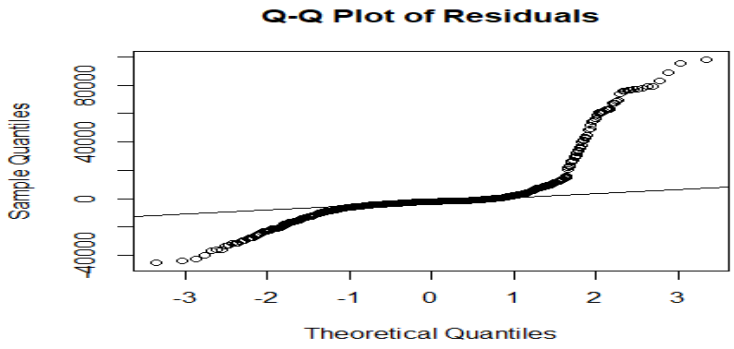

Figure 9: Q-Q Plot of Residuals

Table 1: Modeling results of the COVID-19 new cases in India using ARFIMA model with the corresponding Hurst exponent

Parameter	Value
Model Specification	ARFIMA(0, 0.5, 0)
AIC	27,455.37
Sigma (Standard Error)	14,542.53
Log-likelihood	-13,725.68
h parameter	0.0001446
Ljung-Box statistic	7467.3 (df=10, p < 2.2e-16)
Shapiro-Wilk W	0.60958 (p < 2.2e-16)
Wave Pattern Analysis	Four distinct waves identified
Residual Distribution	Highly leptokurtic with heavy tails
Forecasting Behavior	Declining trend with stabilization
Memory Characteristic	Strong long-memory (persistent)

Significant long-memory features are demonstrated by the ARFIMA(0, 0.5, 0) model for Indian COVID-19 case data, see table 1. A satisfactory overall fit is obtained by diagnostic tests, although non-normal residuals (W = 0.60958) and significant remaining autocorrelation (Ljung-Box = 7467.3) are also observed (especially for typical data). Complex underlying dynamics are indicated by these observations. The four different pandemic waves are successfully captured by the model; however, it has limitations during extreme episodes, indicating the possibility of further improvement using hybrid modeling techniques.

Table 2: ARFIMA Model Parameter Values

Parameter	Value
AIC	2.745537e+04 (27,455.37)
Sigma	1.454253e+04 (14,542.53)
Log likelihood	-1.372568e+04 (-13,725.68)
h	1.446341e-04 (0.0001446)

Table 2 gives the estimates of model parameters. With sigma = 14,542.53 denoting the amount of the prediction error, model parameter estimation produces AIC = 27,455.37, which indicates model fit in relation to its complexity. The fractional differencing technique is finely calibrated, as evidenced by the modest h-parameter (0.0001446).

Table 3: Statistical Test Results

Test	Statistic	Value	p-value	Conclusion
Shapiro-Wilk Test for Normality	W	0.60958	< 2.2e-16	Strong rejection of normality assumption
Ljung-Box Test for Autocorrelation	X-squared	7467.3 (df=10)	< 2.2e-16	Strong evidence of residual autocorrelation

The statistical test results given in Table 3 demonstrate significant deviations from the ideal residual qualities via diagnostic tests. While the Ljung-Box test (X^2 = 7467.3, p < 2.2e-16) shows significant residual autocorrelation, indicating chances for model development, the Shapiro-Wilk test (W = 0.60958, p < 2.2e-16) strongly contradicts normality.

Table 4: Model Selection Criteria

Criterion	Description	Implication
Minimal residual standard deviation	Assessment of prediction error	Non-linear least squares method shows slightly lower standard deviation
Akaike criterion (AIC)	Model fit balanced with complexity	The non-linear least squares method shows marginally better (lower) AIC
Ljung Box test	Test for residual autocorrelation	Both models pass the test with significance level > 5%
Hurst exponent	Measure of long-memory behavior	Both models show strong persistence (H > 0.5)

The marginal superiority of the nonlinear least squares approach across all evaluation metrics is confirmed by the model selection criteria, see Table 4. The Ljung-Box test confirms residual adequacy for the optimized models at the 5% significance level, and both models show strong persistence (H > 0.85).

6. Conclusion

This study demonstrates the utility of ARFIMA models in capturing long-memory behavior in COVID-19 time-series data. The estimated differencing parameter confirms strong persistence, while residual diagnostics highlight areas for potential model refinement. Given the significant autocorrelation and non-normality in residuals, further research could explore hybrid models incorporating non-parametric techniques to improve forecasting accuracy.

References

[1] https://stats.stackexchange.com/questions/408871/interpreting-qq-plot-from-arima-residuals

[2] Demetris, K. Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrological Sciences–Journal–des Sciences Hydrologiques, **2003** 48(1) February 2003.*

[3] M.A. Sanchez Granero, M.A.; Trinidad Segovia, J.E.; García Pérez, J. Some comments on Hurst exponent and the long memory processes on capital markets. *Physica A 387 (**2008**) 5543–5551.*

[4] Mignon, V. Hurst's exponent estimation methods. Application to stock market profitability. *In: Economy and Forecasting, n° 132-133, **1998**-1-2. Pp. 193-214.*

[5] Angela D'Elia, A.; Piccolo, D. Maximum likelihood estimation of ARFIMA models with a Box-Coxtransformation. *Statistical Methods & Applications (**2003**) 12: 259–275. DOI: 10.1007/s10260-003-0064-0.*

[6] Forecasting COVID-19 new cases in Algeria using Autoregressive fractionallyintegrated moving average Models (ARFIMA), Balah Belkacem, Messaoud DJEDDOU, https://www.researchgate.net/publication/341251558

[7] Mostafaei, H. Using SARFIMA Model to Study and Predict the Iran's Oil Supply. *International Journal ofEnergy Economics and Policy. Vol. 2, No. 1, **2012**, pp.41-49 ISSN: 2146-4553.*

[8] Corduas, M. Preliminary Estimation of ARFIMA Models. Chapte. Dipartimento di ScienzeStatistiche, ***2000**, University of Naples Federico II, Napoli, Italy. DOI: 10.1007/978-3-642-57678-2_28*

[9] Cao, G.; He, L-Y.; Cao, J. Multifractal Detrended Analysis Method and Its Application in Financial Markets.*2018, Springer. 255 pages. ISBN 978-981-10-7916-0.*

[10] Moeeni, H.; Bonakdari, H.; Seyed Ehsan Fatemi, S.E.; Zaji, A.H. Assessment of Stochastic Models and aHybrid Artificial

Neural Network-Genetic Algorithm Method in Forecasting Monthly Reservoir Inflow, IndianNational Academy of Engineering, INAE Lett (**2017**) 2:13–23 DOI 10.1007/s41403-017-0017-9

*[11]*Mason, D.M. The Hurst phenomenon and the rescaled range statistic. *Stochastic Processes and theirApplications 126 (**2016**) 3790–3807.*

[12] Tong, S.; Lai, Q.; Zhang, J.; Bao, Y.; Lusi, A.; Ma, Q.; Li. X.; Zhang, F. Spatiotemporal drought variabilityon the Mongolian Plateau from 1980–2014 based on the SPEI-PM, intensity analysis and Hurst exponent.*Science of the Total Environment 615 (**2018**) 1557–1565.*

[13] Aouad, H.S.; A.H.; Taouli, M.K.; Benbouziane, M. Modeling the behavior of the Algerian dinar exchangerate: an empirical investigation using the ARFIMA method. *International Research Journal of Finance andEconomics, **2012**, Issue 87* ISSN 1450-2887.

*[14]*Grech, D.; Mazur, Z. Can one make any crash prediction in nance using the local Hurst exponent idea.*Physica A 336 (**2004**) 133 – 145. doi:10.1016/j.physa.2004.01.018.*

Acknowledgement:

The authors sincerely acknowledge the support of the Uttar Pradesh State Government for funding this research under the Research and Development Yojana scheme. Their financial assistance has been instrumental in facilitating this study.