Comparative Analysis of ARIMA and Random Forest Models for Forecasting COVID-19 Cases in China

Ran Chen *

Department of Applied Mathematics, Case Western Reserve University, Cleveland, United States

* Corresponding Author Email: rxc731@case.edu

Abstract. The COVID-19 pandemic has had a profound global impact on public health systems and global economic situation, making accurate forecasting of infection cases crucial for formulating effective intervention strategies. This study systematically compares the performance of the ARIMA time series model and the Random Forest machine learning model in predicting daily COVID-19 cases in China from 2020 to 2022. The data was partitioned into training and testing sets for model development and evaluation. Results indicate that the Random Forest model significantly outperforms the ARIMA model across all evaluated metrics, including residual mean, standard deviation, and key error indicators, demonstrating a superior ability to capture the timing and amplitude of infection peaks and troughs. Therefore, the value of this study lies in providing clear empirical evidence for model selection in epidemic prediction, indicating that in the face of complex epidemic data, machine learning models may be more reliable than traditional time series methods.

Keywords: COVID-19 forecasting, ARIMA model, random forest model, time series, machine learning.

1. Introduction

The COVID-19 pandemic has caused significant disruptions to social economics and public health systems worldwide [1]. The rapid transmission and recurrent waves of the virus have made accurate forecasting of infection cases a critical tool for governments and health institutions [2]. Reliable predictive models can aid in the proactive allocation of medical resources and provide scientific evidence for government to show what they can do to mitigate the spread of the virus and reduce its impact on public.

China, as a country with a massive population, implemented unique containment strategies during the pandemic, resulting in distinct phasic and regional characteristics in its case data [3]. Compared to global trends, China's epidemic data exhibited a complex pattern of strict initial control followed by localized outbreaks, lending particular relevance to the study of case forecasting in this context [4]. Therefore, predictive modeling of COVID-19 cases in China not only enhances the understanding of its domestic epidemic evolution but also offers a valuable case study for global forecasting efforts under similar conditions.

Therefore, this study employs daily COVID-19 case data from China between 2020 and 2022 to construct and systematically compare an ARIMA model and a Random Forest model. The research aims to identify an effective forecasting method suitable for highly volatile epidemic data.

2. Data

The daily COVID-19 case data represents a comprehensive record of infection patterns throughout the pandemic. The dataset includes complete daily records from January 2020 through December 2025. This specific time period was chosen to comprehensively cover the entire lifecycle of the entire epidemic, including the initial outbreak stage, multiple waves of infection, and the subsequent stable period. Of course, for the later training set data, This project used the specific data from 2020 to 2022. This was because these three years were the periods when the COVID-19 pandemic was most severe, and they also provided the best basis for data comparison.

This data is sourced from https://ourworldindata.org/covid-cases. This website provides the daily new confirmed cases of COVID-19 per million people worldwide from 2020 to 2025. This article selected the data of China from 2020 to 2022 from all the data. The case situation in China during 2020 to 2022 is as follows. Fig. 1 shows the trend of changes in the data. From Fig. 1, it can conclude that there are very few missing values within the time range of 2 complete years (from 2020 to 2022), and the zero-value days have also been properly handled.

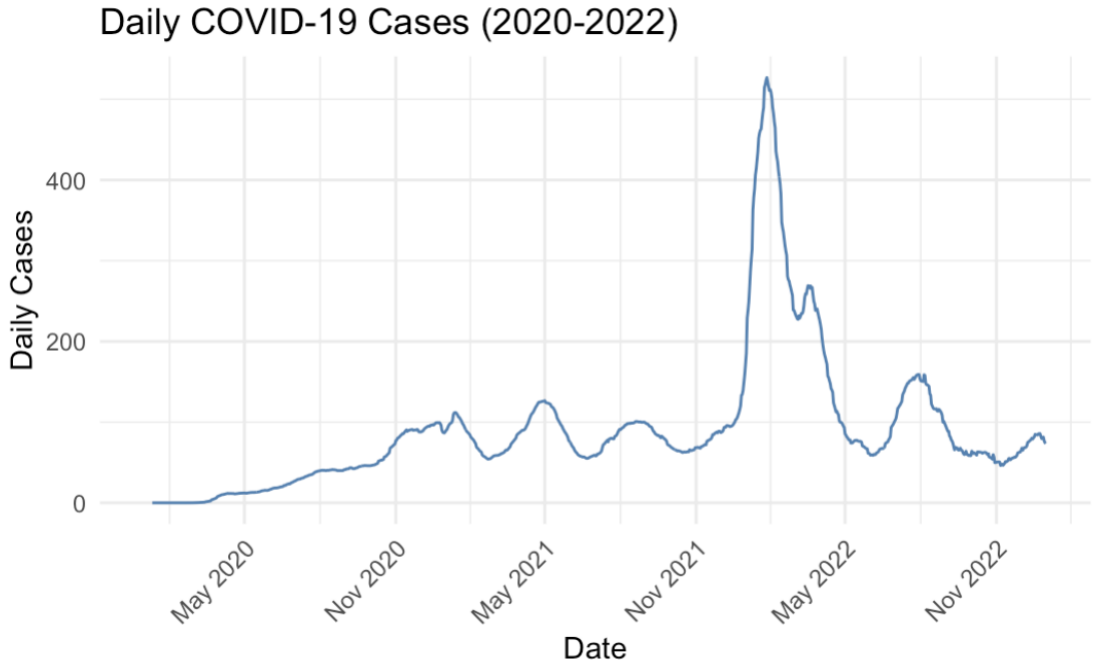

Fig. 1 Daily COVID-19 cases in China from 2020 to 2022

3. Method

3.1. ARIMA Model

For time series data with complex patterns and potential non-stationarity, ARIMA can provide a structured approach to forecasting. The ARIMA model captures temporal dependencies through three main components. First, the autoregressive component, which models the relationship between current values and past observations. Second, the differencing component, which handles non-stationarity by transforming the data to achieve stationarity. Third, the moving average component, which accounts for the relationship between current values and past forecast errors [5-7]. The ARIMA model formulation with parameter triple (p, d, q) can be expressed as:

$$(\lambda - Y\alpha')\left(1 - L\right)^{d} Y_t = c + \sum\left(1 + \theta_i L\right)\varepsilon_t \tag{1}$$

The model considers Y_t, which represents daily COVID-19 cases at time t, as the dependent variable. The lag operator L is defined such that $L^k Y_t$ equals Y_{-}, indicating the value of Y at time t minus k periods. The model incorporates autoregressive parameters ϕ_i for i ranging from 1 to p, which capture the relationship between current and past values of the series. Additionally, moving average parameters θ_i for i from 1 to q accounts for the dependence between the current value and past error terms [8-9]. The order of differencing, denoted by d, determines the number of times the series is differenced to achieve stationarity. The white noise error term at time t is represented by ϵ and follows a normal distribution with mean zero and variance σ^2. Finally, c represents the constant term or drift component in the model.

3.2. Random Forest Model

The random forest model is a supervised machine learning algorithm used for classification and regression tasks. It operates by constructing an ensemble of multiple decision trees during the training process, and outputs the majority class (for classification) or the average prediction value (for regression) of each tree. Compared to using a single decision tree, this ensemble method aims to improve prediction accuracy and reduce overfitting. The Random Forest prediction y for input vector x is given by:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}) \tag{2}$$

The random forest model employs an ensemble approach where B represents the number of decision trees in the ensemble with B set to 100 in the implementation, $T_6(x)$ denotes the prediction of the b-th decision tree for input x, and x constitutes the feature vector comprising temporal and lag features used for forecasting daily COVID-19 cases.

4. Results

4.1. Data Partition

This partitioning strategy was designed to allow the models to learn from the volatile early pandemic periods (2020-2021), which included the initial outbreak phase and multiple infection waves, while testing on more recent, stabilized data patterns from 2022. The comparable zero-case percentages between training and testing sets (15.2% vs. 12.8%) ensure consistent data characteristics across both partitions (See Table 1).

Table 1. Partition of data

Partition	Time period	Duration	Days	0 case percentage	Purpose
Training sets	Jan 2020 - Dec 2021	2 years	731	15.2%	Model development and parameter estimation
Testing sets	Jan 2022 - Dec 2022	1 years	365	12.8%	Model evaluation and performance validation

4.2. Model Configurations

The ARIMA (2,1,1) model configuration reflects specific parameter selections that characterize its forecasting approach. The autoregressive order of p=2 indicates that current COVID-19 case counts depend on the values from the previous two days, capturing short-term temporal dependencies. First-order differencing with d=1 was necessary to transform the raw data into a stationary series, suggesting the presence of underlying trends in the original case data. The moving average component of q=1 incorporates one lag of forecast errors, allowing the model to adapt to unexpected shocks or irregularities in the data pattern.

In contrast, the Random Forest configuration employs a different parameterization strategy. The model utilizes 100 decision trees, striking a balance between computational efficiency and predictive accuracy. At each split, approximately five features are considered, following the established best practice of using the square root of the total feature count [9]. A minimum node size of five prevents overfitting by avoiding excessive partitioning of small data subsets.

These parameter selections represent distinct optimization approaches—the ARIMA parameters were automatically identified as optimal from numerous combinations through an algorithmic search process, while the Random Forest parameters were systematically determined based on feature dimensionality and established regularization practices to ensure model robustness [10].

4.3. Forecasting Results and Visual Analysis

Prior to presenting the forecasting outcomes, it is instructive to examine the models' performance on the training data, as revealed by residual analysis (See Table 2).

Table 2. Residual Analysis of Two Models

Model	Residual Mean	Residual Standard Deviation
ARIMA	-654.1363	335.2434
Random Forest	40.4479	92.7103

From the residual analysis results of the training set shown in Table 2, it can be seen that the mean residual of the random forest model (40.45) is closer to zero, while the mean residual of the ARIMA model (-654.14) shows a significant negative bias. This indicates that the prediction of the random forest model has almost no systematic overestimation or underestimation problems. Moreover, the residual standard deviation of the random forest model (92.71) is much smaller than that of the ARIMA model (335.24), indicating that its prediction results are more stable with a smaller fluctuation range.

Based on the model established through the analysis of the above training set, the data further made predictions for the daily COVID-19 cases from January to December 2022. As shown in Fig. 2, the actual case numbers (black line) exhibit a fluctuating trend of multiple infection peaks. The prediction results of the ARIMA model (red dotted line) can roughly follow the overall trend, but there is a significant underestimation at the peak, and the response is lagging; in contrast, the random forest model (blue dotted line) can more accurately capture the rise and fall of cases, and has better fitting for the timing and amplitude of peaks and troughs, demonstrating excellent real-time tracking capabilities.

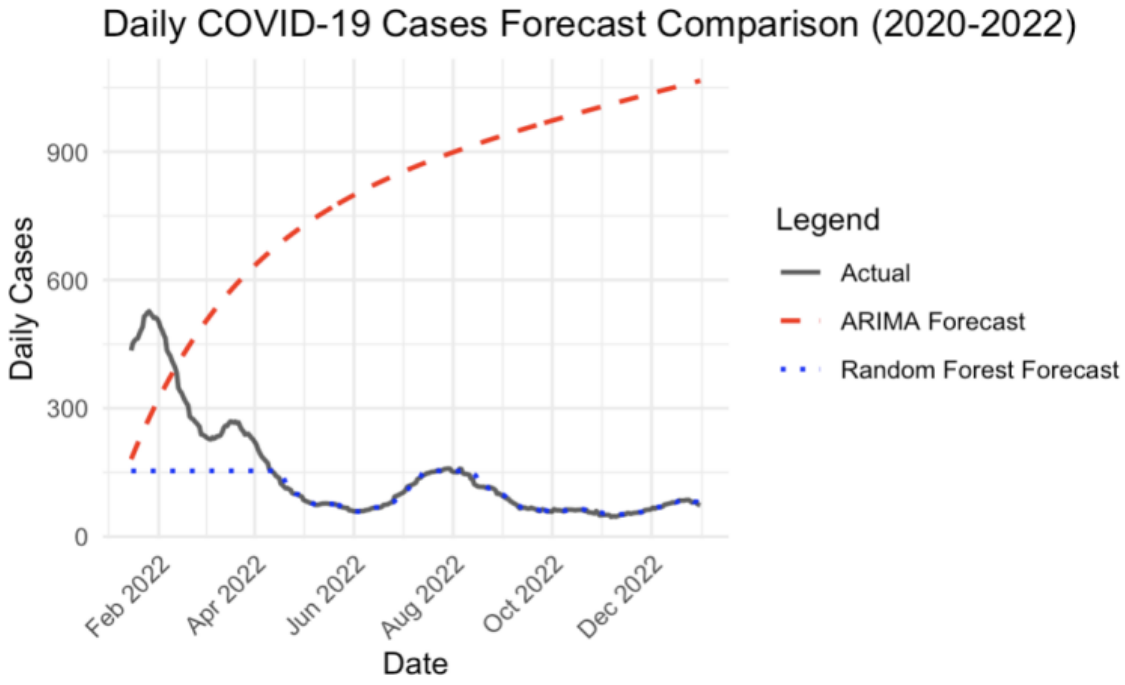

Fig. 2 Daily COVID-19 Cases Forecast Comparison in 2020 in China

To assess the predictive performance of the model more objectively and precisely, the data further calculated three key indicators: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The specific results are presented in Table 3. It can be observed from this that the random forest model significantly outperforms the ARIMA model in all indicators.

Table 3. RMSE, MAE, AND MAPE indicators of two models

Model	RMSE	MAE	MAPE
ARIMA	734.8213	683.0201	848.9
Random Forest	101.0284	43.8677	14.0

Based on the above analysis, whether in terms of residual nature, the degree of prediction graph fitting, or key performance indicators, the random forest model is comprehensively superior to the ARIMA model in the prediction of COVID-19 cases. Its characteristics of low deviation and high precision make it more adaptable to the complex fluctuation patterns of epidemic data, and it has better practical value and promotion potential.

5. Conclusion

This study systematically evaluated the performance of ARIMA and Random Forest models in forecasting daily new COVID-19 cases in China. Through residual analysis, error metric comparison, and visual inspection of forecast fits, the Random Forest model demonstrated superior performance over the ARIMA model in terms of predictive accuracy, stability, and the ability to capture the timing and magnitude of infection peaks. The approach of the Random Forest proved more adept at handling the complex, non-linear patterns inherent in the epidemic data.

Despite these robust findings, this study is not without limitations. The models relied exclusively on historical case data, and future work could be enhanced by incorporating exogenous variables such as mobility data, intervention policies, or genomic surveillance information to improve generalizability.

References

[1] Wang L, et al. A comparative analysis of time series and machine learning models for COVID-19 forecasting. Journal of Medical Systems, 2022, 46 (4): 25.

[2] Chen J, Li K. Forecasting the COVID-19 pandemic: A comparative study of ARIMA and LSTM models. Journal of Healthcare Informatics Research, 2020, 4 (3): 210-225.

[3] Liu Y, Wang Z. Machine learning approaches for epidemic prediction: A case study of COVID-19. IEEE Transactions on Computational Social Systems, 2021, 8 (4): 890-901.

[4] Zhao X, Li X. Predicting COVID-19 outbreaks with random forest and mobility data. Scientific Reports, 2021, 11: 17921.

[5] Box G E P, Jenkins G M. Time series analysis: Forecasting and control. San Francisco: Holden-Day, 1970.

[6] Breiman L. Random forests. Machine Learning, 2001, 45 (1): 5-32.

[7] Zhou L, et al. Evaluating the performance of ensemble methods in epidemic forecasting: Lessons from COVID-19. BMC Medical Informatics and Decision Making, 2022, 22 (1): 98.

[8] Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLOS ONE, 2020, 15 (3): e0231236.

[9] Hyndman R J, Athanasopoulos G. Forecasting: principles and practice. 2nd ed. Melbourne: OTexts, 2018.

[10] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 2003, 50: 159-175.