Proceedings of the 2023 International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/11/20230386

Influenza incidence trend prediction based on ARIMA seasonal multiplicative model

Zhiqian Zhu

Shantou University, Shantou, Guangdong, China

18607109782@163.com

Abstract: This study utilizes a time series ARIMA seasonal multiplicative model to predict the incidence trend of influenza in China, providing valuable early warning references for influenza prevention and control. Monthly incidence data of influenza cases nationwide were collected from January 2014 to August 2022. The data from January 2014 to December 2020 were used as the training set to fit the time series model of influenza incidence. The data from January 2021 to August 2022 were utilized as the testing set to predict the influenza incidence from January 2021 to August 2022 using the fitted model. The predicted values were then compared with the testing set. Through residual white noise testing, significance testing of parameters, and examination of model fit, the final model was determined as $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$, which demonstrated a good fit. The majority of actual data fell within the 95% confidence interval of the predicted values, and the predicted incidence trend aligned closely with the actual trend. The constructed $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$ model holds significant application value in early warning systems for influenza prevention and control, providing crucial insights for public health strategies. In practical applications, this model can be integrated with various factors such as social, natural, and geographic environments to formulate targeted prevention and control strategies, thus enhancing the efficiency and effectiveness of influenza prevention and control measures.

Keywords: influenza, ARIMA seasonal multiplicative model, trend prediction.

1. Introduction

Influenza is a common respiratory infectious disease that causes significant impacts on human health and economic development with large-scale outbreaks occurring worldwide each year [1]. Particularly in China, with the increasing convenience of transportation and global communication in recent years [2], population density and mobility have been constantly rising, leading to an escalating risk of influenza transmission. Therefore, accurate prediction of influenza incidence trend and timely implementation of corresponding prevention and control measures are crucial for safeguarding public health security and promoting economic development. The ARIMA seasonal multiplicative model is a time series analysis method used to handle data with apparent seasonal patterns. It introduces seasonal components based on the traditional ARIMA model to more accurately capture and predict seasonal fluctuations. Domestic scholars have extensively applied the ARIMA model to predict infectious disease surveillance cases in various locations such as Beijing, Shanghai, Guangzhou, and Shanxi [3-9]. This paper aims to construct an ARIMA seasonal multiplicative model for predicting the influenza incidence trend in China, which holds practical significance in influenza prevention and control work.

2. Materials and methods

2.1. Data source and statistical methods

The monthly incidence data of influenza cases nationwide from January 2014 to August 2022 were obtained from the "National Summary of Notifiable Infectious Diseases" published by the National Health Commission.

The data of monthly influenza cases from January 2014 to August 2022 were compiled in Excel, and the SAS software was utilized to establish an ARIMA seasonal multiplicative model for predicting and analyzing the future incidence trend of influenza. The significance level was set at 0.05.

2.2. Study methods

The ARIMA seasonal multiplicative model is a statistical model widely applied in time series forecasting, capable of capturing characteristics such as trends, seasonality, and periodicity in time series data. The multiplicative seasonal $\mathrm{ARIMA}(p,d,q)(P,D,Q)_s$ model is an extension of the ARIMA model designed to handle time series data with seasonality. The model consists of four components: autoregressive terms, differencing terms, moving average terms, and seasonal terms. The notation (p,d,q) represents the orders of the non-seasonal part, while (P,D,Q) represents the orders of the seasonal part, with s denoting the length of the seasonal cycle. In the influenza case data, significant seasonal patterns were observed. Therefore, an ARIMA seasonal multiplicative model can be employed to model and forecast influenza case data.

3. Results

3.1. Time series plot of monthly influenza cases

Based on the time series plot of monthly influenza cases in China from January 2014 to December 2020, the time series is found to be non-stationary. The incidence of influenza in China shows an increasing trend over the years and a seasonal effect with a yearly cycle. Additionally, there is an annual peak occurring predominantly from November to the following February. Around January 2020, there was a sharp increase in reported influenza cases. Several factors contributed to this surge: firstly, the winter and spring seasons are high-risk periods for influenza and other respiratory infectious diseases, leading to an influx of cases as the epidemic spreads across different provinces. Secondly, in November 2019, the National Health Commission issued the "Diagnosis and Treatment Guidelines for Influenza (2019 Edition)", which emphasized the use of rapid diagnostic results as one of the diagnostic criteria for confirmed influenza cases. The implementation of rapid diagnostic services resulted in a significant increase in reported influenza cases. Additionally, increased efforts by health authorities at all levels to strengthen influenza prevention and control measures contributed to heightened public awareness and healthcare-seeking behaviors.

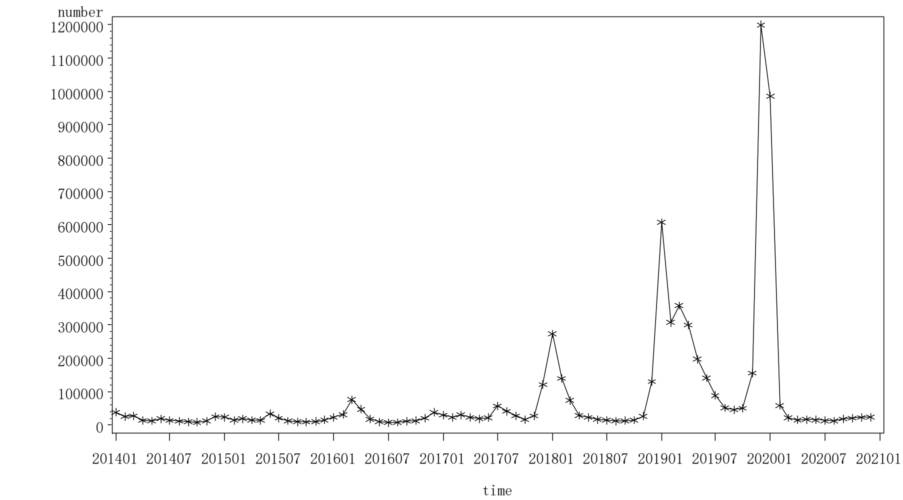


Figure 1. Time series plot of monthly influenza cases in China from January 2014 to December 2020.

3.2. Stabilization of the original series

Based on the aforementioned analysis, the original time series was subjected to ordinary first-order differencing to eliminate the upward trend, and first-order seasonal differencing was performed to remove the seasonal effect. The resulting differenced series is displayed in Figure 2.

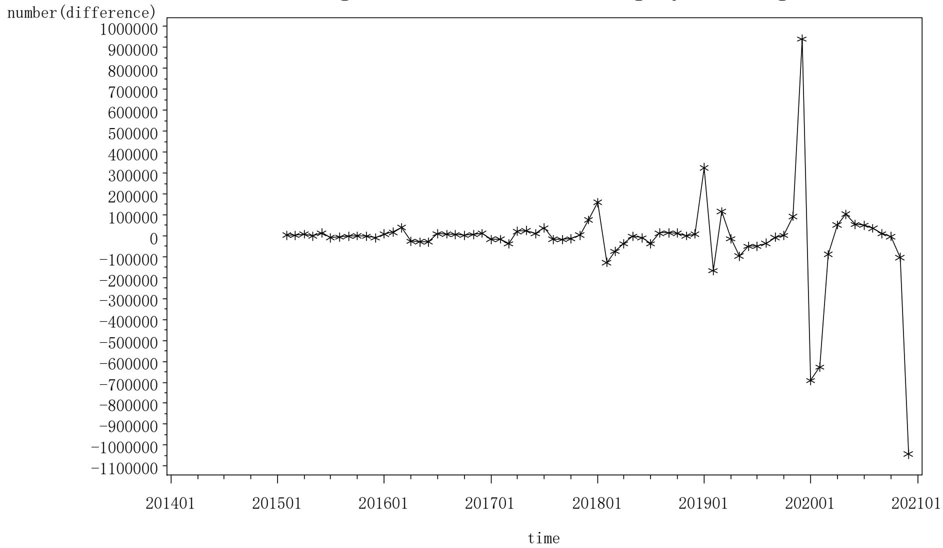


Figure 2. Time series plot after stabilization.

It can be observed that the differenced series exhibits random fluctuations without significant trends or cycles.

3.3. Model identification and parameter estimation

Based on the stabilization process, a preliminary model of $\mathrm{ARIMA}(p,1,q)(P,1,Q)_{12}$ was determined. Autocorrelation and partial autocorrelation plots were generated for the differenced series, as illustrated in Figure 3 and Figure 4, to estimate the model orders based on their characteristics.

```
                                        Autocorrelations

 Lag    Covariance    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      Std Error

   0    4.36782E10        1.00000    |                    |********************|             0
   1    -151570696        -.00347    |               .    |    .               |      0.118678
   2    -8.51623E9        -.19498    |               .****|    .               |      0.118680
   3    -3.83852E9        -.08788    |               .  **|    .               |      0.123109
   4    -2.50577E9        -.05737    |               .   *|    .               |      0.123989
   5    -847384665        -.01940    |               .    |    .               |      0.124362
   6    -1.30111E9        -.02979    |               .   *|    .               |      0.124405
   7    -1.83025E9        -.04190    |               .   *|    .               |      0.124505
   8    1063380757        0.02435    |               .    |    .               |      0.124704
   9    4421134144        0.10122    |               .    |**  .               |      0.124771
  10    6907403043        0.15814    |               .    |*** .               |      0.125922
  11    1.30909E10        0.29971    |               .    |******              |      0.128689
  12    -1.4808E10        -.33903    |             *******|     .              |      0.138171
  13    -4.29639E9        -.09836    |              .   **|     .              |      0.149429
  14    -188413873        -.00431    |              .     |     .              |      0.150338
  15     483422578        0.01107    |              .     |     .              |      0.150340
  16     787958070        0.01804    |              .     |     .              |      0.150351
  17      84481975        0.00193    |              .     |     .              |      0.150382
  18    1169888932        0.02678    |              .     |*    .              |      0.150382
  19    1588601452        0.03637    |              .     |*    .              |      0.150449
  20    -271804990        -.00622    |              .     |     .              |      0.150573
  21    -2.08353E9        -.04770    |              .    *|     .              |      0.150577
  22    1465576597        0.03355    |              .     |*    .              |      0.150790
  23    -648689414        -.01485    |              .     |     .              |      0.150895
  24     179003609        0.00410    |              .     |     .              |      0.150915
```

Figure 3. Autocorrelation plot of the differenced series.

```
                     Partial Autocorrelations

 Lag    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1

   1       -0.00347    |               .    |    .               |
   2       -0.19499    |               .****|    .               |
   3       -0.09290    |               .  **|    .               |
   4       -0.10249    |               .  **|    .               |
   5       -0.06260    |               .   *|    .               |
   6       -0.07805    |               .  **|    .               |
   7       -0.08362    |               .  **|    .               |
   8       -0.01985    |               .    |    .               |
   9        0.06141    |               .    |*   .               |
  10        0.15651    |               .    |*** .               |
  11        0.37717    |               .    |********            |
  12       -0.25636    |               *****|    .               |
  13        0.08473    |               .    |**  .               |
  14       -0.06264    |               .   *|    .               |
  15        0.02837    |               .    |*   .               |
  16       -0.00561    |               .    |    .               |
  17        0.00500    |               .    |    .               |
  18        0.00424    |               .    |    .               |
  19       -0.04373    |               .   *|    .               |
  20       -0.09755    |               .  **|    .               |
  21       -0.09063    |               .  **|    .               |
  22       -0.00378    |               .    |    .               |
  23        0.22985    |               .    |*****               |
  24       -0.10371    |               .  **|    .               |
```

Figure 4. Partial autocorrelation plot of the differenced series.

From the plots, it can be observed that the autocorrelation and partial autocorrelation coefficients decay slowly and exhibit a tailing pattern. Therefore, preliminary values of p and q were set to 1. Considering the seasonal autocorrelation characteristics, the lag-12 autocorrelation and partial autocorrelation coefficients were significantly non-zero, while the lag-24 coefficients fell within two

DOI: 10.54254/2753-8818/11/20230386

standard deviations. Thus, P and Q were tentatively set to 0 and 1. Therefore, the following models were considered: $\mathrm{ARIMA}(1,1,1)(0,1,0)_{12}$, $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$, $\mathrm{ARIMA}(1,1,1)(1,1,0)_{12}$, and $\mathrm{ARIMA}(1,1,1)(1,1,1)_{12}$. Based on the Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC), and considering the goodness of fit as well, the model with the lowest AIC or SBC value was selected as the optimal model. As shown in Table 1, $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$ and $\mathrm{ARIMA}(1,1,1)(1,1,0)_{12}$ exhibited lower AIC and SBC values. By taking into account their predictions for monthly influenza cases from January 2021 to August 2022, the final selected model was the seasonal multiplicative model $\mathrm{ARIMA}(1,1,1)(1,1,0)_{12}$ for influenza forecasting.

Table 1. AIC and SBC values of candidate models.

Model	AIC	SBC
$\mathrm{ARIMA}(1,1,1)(0,1,0)_{12}$	1940.435	1944.961
$\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$	1909.545	1916.333
$\mathrm{ARIMA}(1,1,1)(1,1,0)_{12}$	1902.746	1909.534
$\mathrm{ARIMA}(1,1,1)(1,1,1)_{12}$	1951.698	1960.749

The formula is as follows:

$$\nabla\nabla_{12}x_t = \frac{1-\theta_1 B}{1-\phi_1 B}(1-\theta_{12}B^{12})\varepsilon_t$$

The parameter estimation of this fitted model using the maximum likelihood estimation method in SAS software is:

$$\nabla\nabla_{12}x_t = \frac{1-0.97186B}{1-0.56407\,B}(1-0.78385B^{12})\varepsilon_t$$

3.4. Model verification

The fitted model was subjected to residual white noise tests and parameter significance tests at a significance level of 0.05. The test results indicate that the model is effective, and all three parameters are significant, as presented in Table 2 and Table 3.

Table 2. Residual white noise test results.

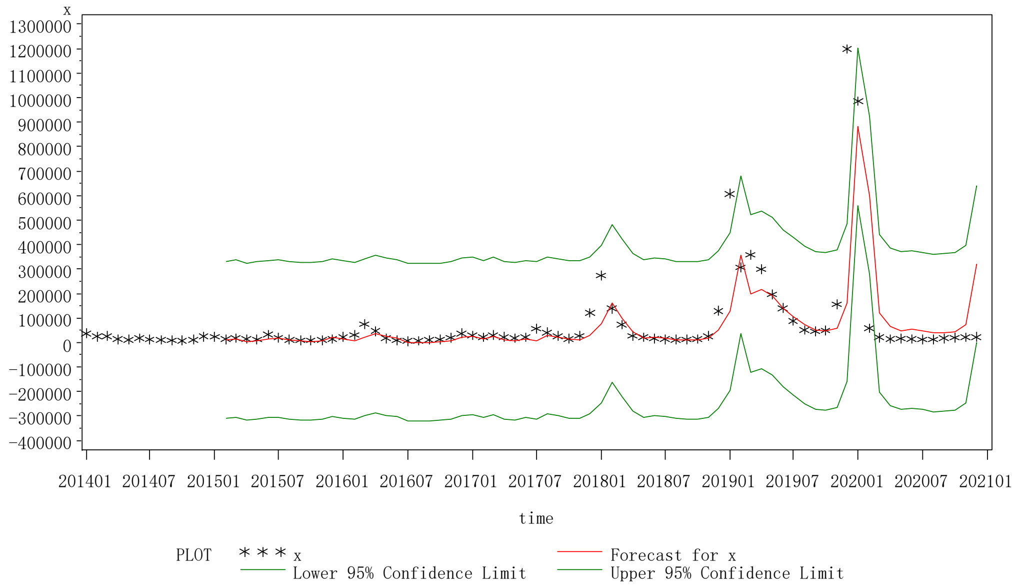
To Lag	Chi-Square	p-value
6	6.26	0.0997
12	10.65	0.3006
18	12.25	0.6597
24	13.18	0.9024

Table 3. Parameter significance test results.

Parameter	Estimate	Standard Error	t-value	p-value	Lag
MA1,1	0.97186	0.03617	26.87	<.0001	1
MA2,1	0.78385	0.13000	6.03	<.0001	12
AR1,1	0.56407	0.12003	4.70	<.0001	1

3.5. Model fit and prediction valuation

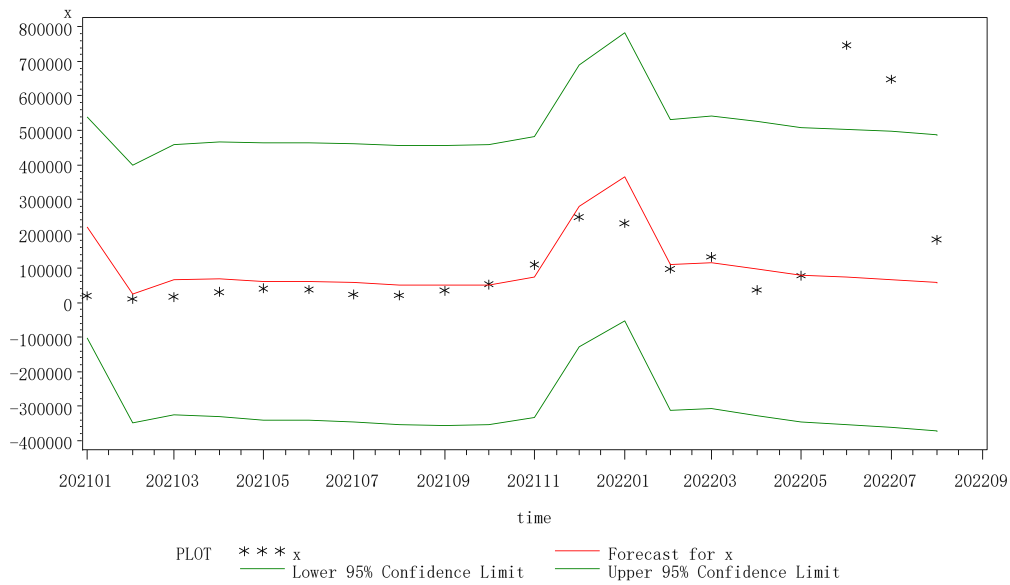
According to the $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$ model applied to re-fit the influenza incidence data in China from January 2014 to December 2020, it was observed that the majority of the actual values fell within the 95% confidence interval of the fitted values, suggesting a strong fit, as illustrated in Figure 5.



Note: Black asterisks represent actual values, red solid line represents fitted values, and the green solid line represents the upper and lower boundaries of the 95% confidence interval.

Figure 5. Fit of influenza incidence data from January 2014 to December 2020.

By comparing the predicted values with the actual values of influenza incidence from January 2021 to August 2022, it was observed that, except for the actual values in June and July 2022, which fell within the 95% confidence interval of the predicted values, the actual incidence for the remaining 18 months was within the 95% confidence interval of the predictions, indicating good prediction accuracy, as depicted in Figure 6.



Note: Black asterisks represent actual values, red solid line represents predicted values, and the green solid line represents the upper and lower boundaries of the 95% confidence interval.

Figure 6. Influenza incidence prediction from January 2021 to August 2022.

4. Conclusion

This study aimed to predict the incidence trend of influenza using a time series ARIMA seasonal model, providing a reference for early warning in influenza prevention and control efforts, thereby improving

DOI: 10.54254/2753-8818/11/20230386

the efficiency of responding to influenza outbreaks.

Based on the collection of monthly influenza incidence data from January 2014 to August 2022 in China, this study utilized the data from January 2014 to December 2020 as the training set to fit the time series model for influenza incidence. Subsequently, the data from January 2021 to August 2022 were used as the test set to forecast influenza incidence during that period, and the predicted results were compared with the test set. Through residual white noise tests, parameter significance tests, and examination of the model's goodness of fit, the $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$ model was determined as the optimal model. This model exhibited good fitting performance, with the majority of actual data falling within the 95% confidence interval of the predicted values. Based on the results, the following conclusions can be drawn: the $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$ model has certain application value in influenza prevention and control early warning, providing important reference for influenza prevention and control efforts. The predicted results closely aligned with the trend of actual data, indicating that the fitted model accurately captured the changes in influenza incidence trends.

However, the model also has some limitations. Firstly, due to the large numerical values of national influenza incidence, errors are magnified. In future research, it is suggested to consider using the incidence rate per month instead of the number of cases to reduce errors. Secondly, future influenza outbreaks may be influenced by various unknown factors, hence uncertainties still exist. For example, from June to August 2022, the influenza peak period in China was delayed by several months compared to previous years, but the incidence was high. It is speculated that the epidemic prevention and control measures had a suppressive effect on the transmission of seasonal influenza, while the decrease in population immunity played a role.

In summary, the $\mathrm{ARIMA}(1,1,1)(0,1,1)_{12}$ model in this study has practical value in influenza prevention and control early warning, providing an important reference for influenza prevention and control efforts. However, to improve prediction accuracy and practicality, further research can consider incorporating more factors, such as integrating meteorological data, population mobility information, and other factors for comprehensive forecasting, coupled with real-time monitoring and adjustments to adapt to the dynamic changes of influenza outbreaks.

References

[1] Historical chronicle - Influenza pandemic in the 20th century. (2009). Disease Surveillance, 24(12), 941.

[2] Lv, X. Q., Zhang, L., Di, H. R., et al. (2018). Overview of research progress on influenza. World Chinese Medicine, 13(02), 278-282.

[3] Liang, J. G., Feng, Y. L., Dong, L., et al. (2023). Application of ARIMA seasonal model in tuberculosis prediction in Shanxi Province. Disease Surveillance, 38(03), 332-338.

[4] Yang, Y., & Zhang, Q. (2023). Application of ARIMA model in forecasting the incidence trend of varicella. Health Education and Health Promotion, 18(02), 149-152. DOI: 10.16117/j.cnki.31-1974/r.202302149.

[5] Xu, L., & Li, X. Y. (2023). Establishment and prediction of ARIMA seasonal model for other infectious diarrheal diseases in Shangcheng District. Practical Preventive Medicine, 30(01), 111-115.

[6] Dai, H. Y., Zhou, N., Ren, X., et al. (2022). Characteristics and trend prediction of various subtypes of influenza based on autoregressive moving average model. Disease Surveillance, 37(10), 1338-1345.

[7] Shi, J. X., Zhang, W. Z., Ji, G. Q., et al. (2010). Application of ARIMA model in predicting and early warning of influenza-like cases. Capital Public Health, 4(01), 12-16. DOI: 10.16760/j.cnki.sdggws.2010.01.010.

[8] Li, Y. T., Zhang, H. W., Ren, H., et al. (2007). Time series analysis and prediction model of influenza-like cases in Shanghai. Chinese Journal of Preventive Medicine, 41(06), 496-498.

[9] Ou, C. Q., Deng, Z. H., Yang, L., et al. (2007). Predicting the variation trend of influenza-like cases using autoregressive models. Chinese Journal of Health Statistics, (06), 569-571.