Methods and Applications

An Autoregressive Integrated Moving Average Model for Predicting Varicella Outbreaks — China, 2019

Miaomiao Wang[1]; Zhuojun Jiang[2]; Meiying You[1]; Tianqi Wang[1,3]; Li Ma[4]; Xudong Li[1,#]; Yuehua Hu[1,5,#]; Dapeng Yin[6,#]

ABSTRACT

Introduction: Varicella, a prevalent respiratory infection among children, has become an escalating public health issue in China. The potential to considerably mitigate and control these outbreaks lies in surveillance-based early warning systems. This research employed an autoregressive integrated moving average (ARIMA) model with the objective of predicting future varicella outbreaks in the country.

Methods: An ARIMA model was developed and fine-tuned using historical data on the monthly instances of varicella outbreaks reported in China from 2005 to 2018. To determine statistically significant models, parameter and Ljung-Box tests were employed. The coefficients of determination (R^2) and the normalized Bayesian Information Criterion (BIC) were compared to selecting an optimal model. This chosen model was subsequently utilized to forecast varicella outbreak cases for the year 2019.

Results: Four models passed parameter (all $P<0.05$) and Ljung-Box tests (all $P>0.05$). ARIMA $(1, 1, 1)\times(0, 1, 1)_{12}$ was determined to be the optimal model based on its coefficient of determination R^2 (0.271) and standardized BIC (14.970). Fitted values made by the ARIMA $(1, 1, 1)\times(0, 1, 1)_{12}$ model closely followed the values observed in 2019, the average relative error between the actual value and the predicted value is 15.2%.

Conclusion: The ARIMA model can be employed to predict impending trends in varicella outbreaks. This serves to offer a scientific benchmark for strategies concerning varicella prevention and control.

Varicella, or chickenpox, is a prevalent childhood disease resulting from varicella-zoster virus infection. As the third most reported vaccine-preventable infectious disease in China, varicella imposes a substantial socio-economic burden (*1*). The disease is notable for its tendency to cause outbreaks and epidemics. Since 2006, these outbreaks have been reported through the Public Health Emergency Management Information System in China (*2*). Utilizing this system facilitates the timely detection of epidemiological trends associated with varicella outbreaks offering vital early warning signals. Such early warnings are particularly crucial for the prevention and control of varicella outbreaks, hence highlighting their significant role in public health.

The autoregressive integrated moving average (ARIMA) models, accommodating alterations in trends, variations in periodicity, and random disturbances within a time series, have seen extensive application in predicting infectious diseases (*3*–*5*). Our study aimed to depict the temporal patterns of varicella outbreak cases in China spanning 2005–2018, assess the practicality of employing ARIMA models to project upcoming monthly varicella outbreak cases, and contribute empirical evidence for early alarms and effective prevention measures to suppress varicella outbreaks.

METHODS

Data Source

Per the "*National Public Health Emergency Related Information Reporting Management Standards*" distributed by the Ministry of Health's General Office on December 27, 2005, any instance of more than ten varicella cases within the same school, kindergarten, and other related units in a single week is classified as a varicella outbreak. Such outbreaks are mandated to be reported via the public health emergency information reporting system. Our research involved the extraction of varicella outbreak surveillance data from January 2005 to November 2019. This data was divided into segments for model development and model validation. We used the monthly varicella outbreak cases from 2005 to 2018 to construct the model, while the 2019 monthly data was employed to validate the

model and generate predictions.

Development of the ARIMA Model

ARIMA models take the form of ARIMA (p, d, q)×(P, D, Q)s. Parameters d (the degree of differencing) and D (moving average) are numbers of differences required to stabilize the time series. Parameters p (the order of autoregression) and q (the order of moving average) are simple numeric parameters. Parameters P (seasonal autoregression) and Q (seasonal integration) are seasonal parameters, and s is the length of the seasonal period.

The construction and prediction of the ARIMA model consist of three steps. First, *Time series stabilization*: we assessed stationarity and seasonality by graphing a time series plot of the monthly varicella outbreak cases. The trend and seasonality of the initial sequence were eliminated by taking the ordinary and seasonal differences. The time series' stationarity was then determined through the analysis of the stabilized sequence graph as well as the autocorrelation function (ACF) and partial autocorrelation function (PACF). Second, *model identification and diagnosis*: the values of d and D were determined based on the trend differences and seasonal variations. The values for p and q, and P and Q were permitted to vary between 0 and 2, and were assessed individually in model construction. Each proposed model had to pass the Ljung–Box and parameter tests. The most suitable model was subsequently selected based on the highest coefficients of determination (R^2) and the lowest normalized Bayesian Information Criterion (BIC). Lastly, *prediction*: the fitted model was used to project the number of monthly varicella outbreaks for 2019 (*4*).

Data Analysis

The analysis of the data was performed utilizing the SPSS software (version 26.0, IBM, Armonk, NY, USA). The Mann-Kendall trend test was utilized to evaluate the outbreak trends. A significance level was established at $P<0.05$.

RESULTS

Temporal Analysis

From 2005 to 2018, China reported 246,772 outbreak cases in 8,545 varicella outbreaks. The time series mapping of these cases revealed a statistically significant decline from 2007 to 2011 ($Z=-2.25$, $P<0.05$). However, from 2012 to 2018, there was a notable increase ($Z=2.63$, $P<0.05$). When decomposed, the time series exhibited three components: random errors, periodic factors, and long-term trend factors. The data demonstrated seasonal characteristics, with major and minor epidemic peaks recurring annually (Figure 1).

Stabilization of Time Series

The ARIMA model was developed using monthly instances of varicella outbreaks spanning from January 2005 to December 2018. Upon inspection of Figure 1, it became apparent that the series displayed non-stationary characteristics, thus necessitating the stabilization of the series through the incorporation of one-order ordinal and seasonal differences. The stabilized sequence, as depicted in Supplementary Figure S1 (available in https://weekly.chinacdc.cn/), did not exhibit a pronounced upward or downward trend. Supplementary Figure S2 (available in https://weekly.chinacdc.cn/) illustrates the autocorrelation coefficient of the stationary series experiencing a swift decrease following a brief delay period. This observation suggests that the modified time series tended toward stationarity subsequent to differencing adjustment. Consequently, this procedure assigned the parameters d and D as 1.

Model Identification and Diagnosis

The initial model was designated as ARIMA (p, 1, q)×(P, 1, Q)$_{12}$. The individual values for p, q, P, and Q were adjusted independently, ranging from 0 to 2. Following this iterative adjustment, four models successfully passed the parameter tests, all with $P<0.05$, and the Ljung-Box tests, all with $P>0.05$: ARIMA (2, 1, 0)×(0, 1, 1)$_{12}$, ARIMA (1, 1, 0)×(0, 1, 1)$_{12}$, ARIMA (1, 1, 1)×(0, 1, 1)$_{12}$, and ARIMA (1, 1, 1)×(2, 1, 0)$_{12}$. Due to its superior R^2 and the minimal standardized BIC, ARIMA (1, 1, 1)×(0, 1, 1)$_{12}$ was selected as the optimal model (Table 1). The autoregressive coefficients of the model residuals all fell within the control line, as depicted in Supplementary Figure 3 (available in https://weekly.chinacdc.cn/), suggesting that the residual error was random and affirming the validity of the chosen model.

Prediction

Utilizing the optimal ARIMA (1, 1, 1)×(0, 1, 1) $_{12}$ model, we predicted varicella outbreak cases from January through November 2019. The actual values

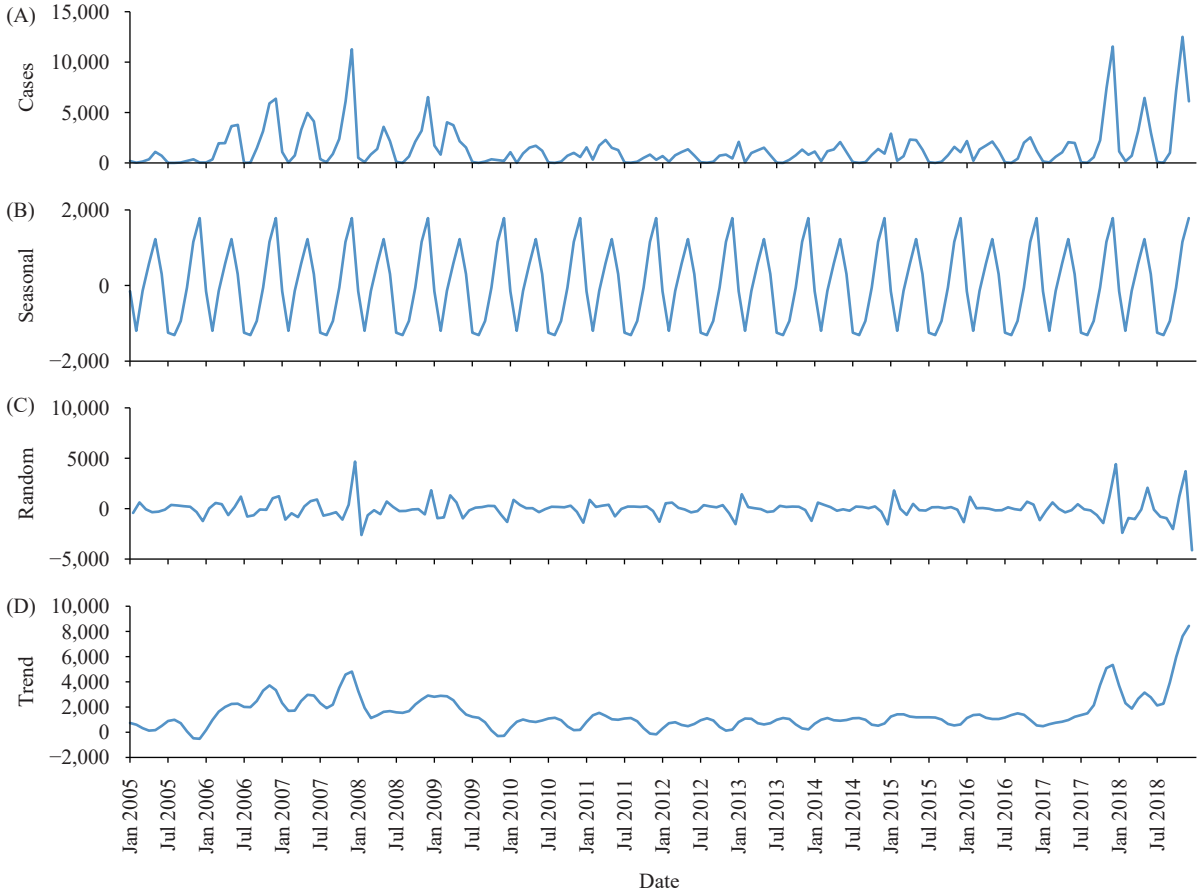


FIGURE 1. The time series graph of monthly varicella outbreak cases in China, 2005–2018. (A) Original time series; (B) seasonal effect; (C) random fluctuation effect; (D) long-term trend effect.

TABLE 1. Estimation of parameters and verification of the ARIMA model.

Variable	ARIMA (2, 1, 0)×(0, 1, 1)$_{12}$		ARIMA (1, 1, 0)×(0, 1, 1)$_{12}$		ARIMA (1, 1, 1)×(0, 1, 1)$_{12}$		ARIMA (1, 1, 1)×(2, 1, 0)$_{12}$	
	Estimate	*P*	Estimate	*P*	Estimate	*P*	Estimate	*P*
AR	−0.346	0	−0.168	0.052	0.379	0	0.381	0
MA	–	–	–	–	0.933	0	0.940	0
Seasonal AR	–	–	–	–	–	–	−0.308	0.006
Seasonal MA	0.306	0.003	0.402	0	0.357	0	–	–
Ling-Box p	0		0		0.005		0.007	
Stationary R^2	0.147		0.048		0.271		0.287	
Normalized BIC	15.127		15.198		14.970		14.987	

Note: "–" represents null values.
Abbreviation: ARIMA=autoregressive integrated moving average; AR=autoregression; MA=moving average; BIC=Bayesian Information Criterion.

aligned closely with the fitted values preceding October 2019 (Figure 2). Even though the subsequent fitted values did not align as closely, they remained within the predicted 95% confidence intervals. The average relative error between the predicted and actual values was 15.2% (Table 2), inferring that the model was deemed suitable for prediction purposes.

DISCUSSION

This study may be the premiere use of an ARIMA model to delineate the epidemic trajectory of varicella outbreaks in China, as it offers a predictive overview of imminent varicella trends. This provides valuable insight for preemptive measures and public health

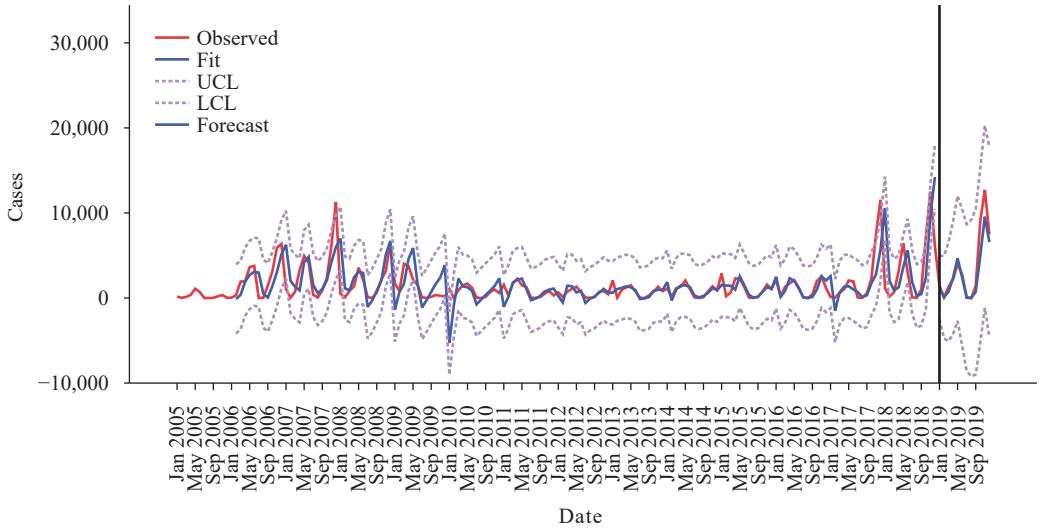


FIGURE 2. Time-series plots of predicted monthly varicella outbreak cases using the ARIMA model, January 2005–December 2019.
Note: The dotted lines represent the 95% *CI*s, with UCL denoting the upper limit and LCL indicating the lower limit of the 95% *CI*.
Abbreviation: ARIMA=autoregressive integrated moving average model; *CI*=confidence interval; UCL=upper confidence limit; LCL=lower confidence limit.

TABLE 2. Comparison between predicted and actual values using ARIMA (1, 1, 1) × $(0, 1, 1)_{12}$ model.

Outbreak cases	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Dec	Nov	Mean relative error
Actual	1,459	91	1,318	2,196	4,038	2,736	93	0	1,428	9,091	12,707	7,492	3,554.083
Predicted	1,256	91	1,335	2,407	4,665	2,421	43	17	745	5,032	9,565	6,570	2,845.583
Absolute error	−203	0	17	211	627	−315	−50	17	−683	−4,059	−3,142	−922	−708.500
Relative error	−0.139	0	0.013	0.096	0.155	−0.115	−0.538	0	−0.478	−0.446	−0.247	−0.123	−0.152

guidance (*5*). Our research reveals an uninterrupted increase in reported varicella outbreak cases since 2012, with a significant surge from 2017 that peaked in 2018. Projections for 2019 continue this rising trend, suggesting that varicella outbreaks have not been fully contained. The low Varicella vaccine (VarV) coverage in China could be a potential catalyst for these increases (*6*). Previous research supports the efficacy of both single and double dose varicella vaccination schedules in mitigating varicella outbreaks (*7*–*9*). A separate study conducted in China (*10*) utilized a modified Delphi technique to gather expert opinions on the potential inclusion of non-program vaccines into China's Expanded Program on Immunization (EPI). VarV emerged as the top non-program vaccine recommended for incorporation into the EPI. Thus, these findings underscore the importance and urgency of integrating VarV into the national immunization framework.

After conducting numerous adjustments to the parameters and running goodness-of-fit tests, it was conclusively determined that the ARIMA (1, 1, 1)× $(0, 1, 1)_{12}$ model was the most compatible with the original time-series data of monthly varicella outbreak cases gathered from 2005 to 2018. This optimal model was subsequently used to predict monthly varicella outbreak cases in 2019. The results revealed that the estimated cases of outbreaks were congruent with the actual reported cases, particularly from January to September. This correspondence indicated the model's ability to accurately predict varicella outbreak cases. From October 2019 onwards, the fitted values did not align as closely, albeit still falling within the 95% confidence intervals. This points to potential influences of large seasonal fluctuations or changes in policy on the model's accuracy, a factor which warrants further analysis. Consequently, it is recommended that the model's data be regularly updated with the most current information to ensure optimal accuracy.

Time series models are instrumental in the prediction of varying trends in infectious diseases such as hand foot and mouth disease (HFMD) (*11*),

coronavirus disease 2019 (COVID-19) (*12*), and influenza (*3*). Our research reinforces the scientific consensus deeming the ARIMA models as proficient tools for synchronous surveillance and forecasting of evolving trends in infectious diseases. Notably, a study conducted in Bulgaria (*13*) illustrated the appropriateness of an ARIMA model in describing varicella incidence trends, and its suitability in projecting near-future disease dynamics, although it didn't account for varicella seasonality. In relation to China, there have been limited studies conducted on varicella incidence prediction, with existing studies only forecasting sporadic varicella incidence in specific regions. For the first time, our study utilizes varicella outbreak data to forecast varicella outbreak occurrences in China on a monthly basis, effectively eliminating the influence of seasonality.

Our study is subject to some limitations. Initially, the use of passive surveillance data can potentially result in an underestimation of the disease burden, which could consequently impact the precision and accuracy of our analyses. Furthermore, the accuracy of our ARIMA model might be subjected to the dynamic changes in key influencing factors such as policy alterations and climate changes. Therefore, the establishment of a dynamic adjustment model is essential to enhance the accuracy of long-term predictions.

In conclusion, the findings from our research indicate the practicality of employing ARIMA models for predicting varicella outbreaks in China. Consequently, these models pose a valuable tool for enhancing varicella prevention and control measures, offering forecasting capabilities for future varicella outbreaks and trend identification within the nation.

Conflicts of interest: No conflicts of interest.

Funding: Supported by Beijing Natural Science Foundation (L202008) and National Science and Technology Major Project of China (2012CB955500, 2012CB955504).

doi: 10.46234/ccdcw2023.134

[#] Corresponding authors: Xudong Li, lixd@chinacdc.cn; Yuehua Hu, huyueer@163.com; Dapeng Yin, yindapeng@hainan.gov.cn.

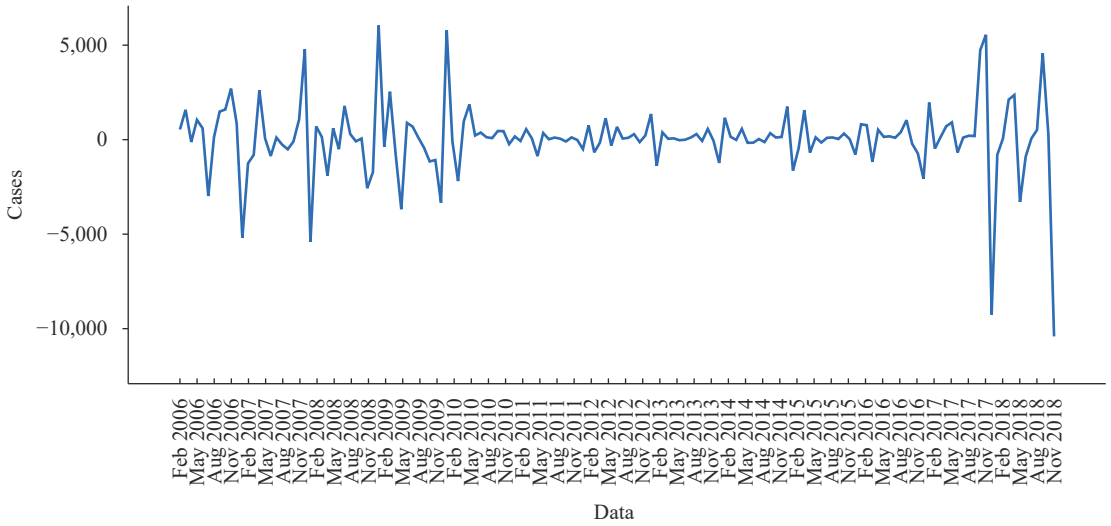
[1] Office of Epidemiology, Chinese Center for Disease Control and Prevention, Beijing, China; [2] Training and Outreach Division, National Center for Mental Health, Beijing, China; [3] Data Resources and Statistics Department, Beijing Municipal Health Big Data and Policy Research Center, Beijing, China; [4] Hefei Center for Disease Control and Prevention, Hefei City, Anhui Province, China; [5] Technical Guidance Office for Patriotic Health Work, Chinese Center for Disease Control and Prevention, Beijing, China; [6] Hainan Center for Disease Control and Prevention, Haikou City, Hainan Province, China.

Submitted: June 28, 2023; Accepted: July 27, 2023

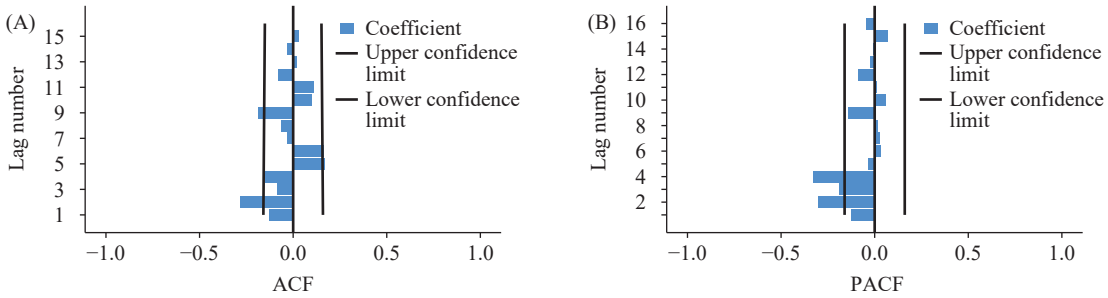
REFERENCES

1. Feng HYF, Zhang HJ, Ma C, Zhang HN, Yin DP, Fang H. National and provincial burden of varicella disease and cost-effectiveness of childhood varicella vaccination in China from 2019 to 2049: a modelling analysis. Lancet Reg Health West Pac 2023;32:100639. http://dx.doi.org/10.1016/j.lanwpc.2022.100639.
2. Ministry of Health of People's Republic of China. Notice of the General Office of the Ministry of Health on the printing and distribution of the national work specification for the management of information reporting related to public health emergencies (trial). 2006. http://www.nhc.gov.cn/cms-search/xxgk/getManuscriptXxgk.htm?id=31353. [2023-4-26] (In Chinese)
3. Chen Y, Leng KK, Lu Y, Wen LH, Qi Y, Gao W, et al. Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010-2018. Epidemiol Infect 2020;148:e29. http://dx.doi.org/10.1017/S0950268820000151.
4. Schaffer AL, Dobbins TA, Pearson SA. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. BMC Med Res Methodol 2021;21(1):58. http://dx.doi.org/10.1186/s12874-021-01235-8.
5. Liu QY, Liu XD, Jiang BF, Yang WZ. Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. BMC Infect Dis 2011;11:218. http://dx.doi.org/10.1186/1471-2334-11-218.
6. Liu AP, Sun TT. Meta-analysis of varicella vaccine coverage among Chinese children. Chin J Vaccines Immun 2017;23(6):698-704. https://d.wanfangdata.com.cn/periodical/zgjhmy201706022. (In Chinese)
7. Leung J, Lopez AS, Marin M. Changing epidemiology of varicella outbreaks in the United States during the varicella vaccination program, 1995-2019. J Infect Dis 2022;226(S4):S400 – 6. http://dx.doi.org/10.1093/infdis/jiac214.
8. Chen YW, Ma R, Zhang YY, Li XD, Yin DP. Effects of varicella vaccine time of first dose and coverage of second dose — Beijing and Ningbo, China, 2012–2018. China CDC Wkly 2020;2(36):696 – 9. http://dx.doi.org/10.46234/ccdcw2020.136.
9. Zhao D, Suo LD, Lu L, Pan JB, Pang XH, Yao W. Effect of earlier vaccination and a two-dose varicella vaccine schedule on varicella incidence — Beijing Municipality, 2007–2018. China CDC Wkly 2021;3(15):311 – 5. http://dx.doi.org/10.46234/ccdcw2021.085.
10. Ma C, Li JH, Wang N, Wang YM, Song YD, Zeng X, et al. Prioritization of vaccines for inclusion into China's expanded program on immunization: evidence from experts' knowledge and opinions. Vaccines 2022;10(7):1010. http://dx.doi.org/10.3390/vaccines10071010.
11. Liu L, Luan RS, Yin F, Zhu XP, Lü Q. Predicting the incidence of hand, foot and mouth disease in Sichuan Province, China using the ARIMA model – CORRIGENDUM. Epidemiol Infect 2016;144(1):152. http://dx.doi.org/10.1017/S0950268815001582.
12. Qi BG, Liu NK, Yu SC, Tan F. Comparing COVID-19 case prediction between ARIMA model and compartment model - China, December 2019–April 2020. China CDC Wkly 2022;4(52):1185 – 8. http://dx.doi.org/10.46234/ccdcw2022.239.
13. Raycheva R, Kevorkyan A, Stoilova Y. Stochastic modelling of scalar time series of varicella incidence for a period of 92 years (1928-2019). Folia Med (Plovdiv) 2022;64(4):624 – 32. http://dx.doi.org/10.3897/folmed.64.e65957.

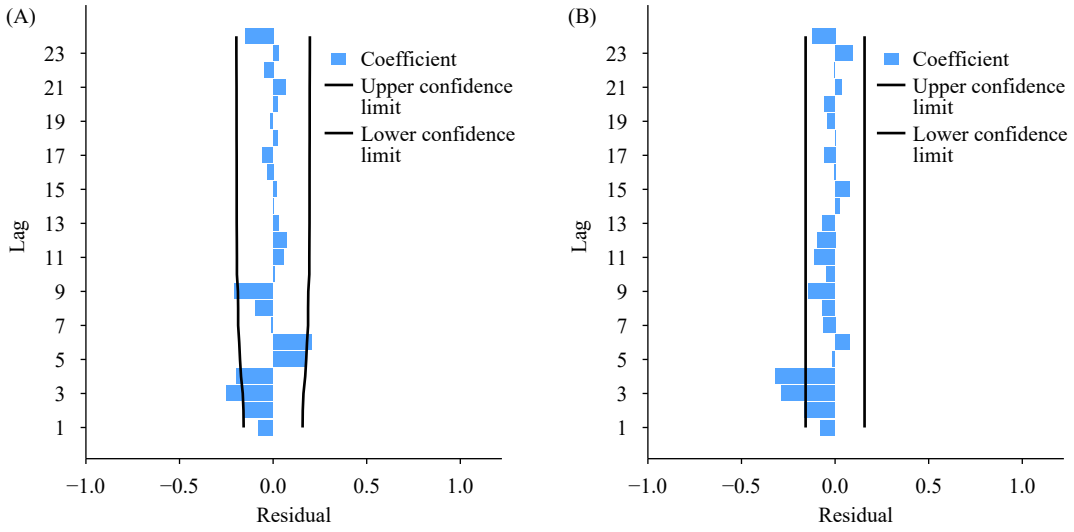
SUPPLEMENTARY MATERIAL



SUPPLEMENTARY FIGURE S1. The sequence following initial ordinary and seasonal first-order differences.



SUPPLEMENTARY FIGURE S2. Diagrams of outbreak cases in China from 2005 to 2018 after first-order differencing and seasonal differencing. (A) ACF; (B) PACF.
Abbreviation: ACF=autocorrelation function; PACF=partial autocorrelation function.



SUPPLEMENTARY FIGURE S3. ACF and PACF graphs of the residuals for the ARIMA (1, 1, 1)×(0, 1, 1)$_{12}$ model. (A) ACF; (B) PACF.

Note: Given that the correlation values fell within the 95% *CI* limits, it can be inferred that the residuals are most likely white noise. This suggests that this model is suitable for prediction.

Abbreviation: ARIMA=autoregressive integrated moving average model; ACF=autocorrelation function; PACF=partial autocorrelation function; *CI*=confidence interval.