Letter to the Editor

Now- and Fore-casting the Secular Epidemiological Trends and Seasonality of the Comeback of Scarlet Fever in China: A 16-year Time Series Analysis*

WANG Yong Bin[1,&,#], LI Yan Yan[1,&], LU Hao[1,&], TAO Ying Jun[1], LI Yu Hong[2], WANG Lei[3], and LIANG Wen Juan[1,#]

Scarlet fever (SF) is a common communicable disease that results from group A *Streptococcus* (GAS) infections[1]. SF accounted for the global loss of life among children 5–15 years of age in the 18th and 19th centuries[2]. A rapid reduction in SF morbidity and mortality occurred due to the scale-up of effective antibiotics and improvements in sanitation and nutrition[3]. The unexpected increase in the incidence of SF has attracted a renewed interest in infectious diseases[3]. Because the triggers that cause SF outbreaks are not fully understood and there is a scarcity of available vaccines protecting susceptible populations from GAS infections, effective prevention and control plans are required to stop the continued spread of SF.

Time series analysis assists in the development of hypotheses to explain the temporal patterns of different diseases and to analyze the spread, therefore, facilitating the creation of a quality forecasting system. The seasonal autoregressive integrated moving average (SARIMA) model has been widely applied to estimate the epidemiological patterns of contagious diseases because this model has a simple structure, fast applicability, and a relatively high forecasting reliability level[4]. It has been shown that the SARIMA model is able to satisfactorily estimate a simple time series[4], but it is difficult to manage complex time series, such as the data with multiple seasonal periods, high-frequency seasonality, non-integer seasonality, and dual-calendar effects. By comparison, the innovation state-space modelling framework that combines Box-Cox transformations, Fourier series with time-varying coefficients, and autoregressive moving average (ARMA) error correction (known as the TBATS method) is customized for use with the patterns included in a complex time series described above[5]. In addition, the TBATS model is used for linearity and some types of non-linearity in a complex time series based on Box-Cox transformations[5], which makes it possible to perform a multistep ahead prediction. Moreover, the TBATS model is able to decompose a complex seasonal time series into its trend, seasonal, and irregular components[5], which is not able to be undertaken by use of the SARIMA model. Importantly, SF morbidity has been shown to display dual seasonal patterns in some countries. Therefore, this study analyzed the long-term epidemic patterns using the TBATS model. The forecasting power under the TBATS model was compared with the SARIMA model.

We obtained the monthly SF incidence and population data between January 2004 and December 2019 from the Chinese CDC and the Statistical Yearbook of China, respectively. Then, we partitioned the SF morbidity series into two segments comprising a training dataset from January 2004 to December 2017 to construct the SARIMA and TBATS models and a testing dataset from January 2018 to December 2019 to test the generalization of both models. Two additional datasets were provided to test the robustness of both models: the first 180 data sets from January

doi: 10.3967/bes2022.076

*This work was supported by the Natural Science Foundation of Henan, China, Innovation and Entrepreneurship Training Project for University Students of Henan Province and Xinxiang Medical University [222300420265, S202110472047, S202010472007, and XYXSKYZ201932] and the Key Scientific Research Project of Universities in Henan [21A330004].

1. Department of Epidemiology and Health Statistics, School of Public Health, Xinxiang Medical University, Xinxiang 453003, Henan, China; 2. National Center for Tuberculosis Control and Prevention, China Center for Disease Control and Prevention, Beijing 102206, China; 3. Center for Musculoskeletal Surgery, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität Zu Berlin and Berlin Institute of Health, Berlin, Germany

2004 to December 2018 and 156 data sets from January 2004 to December 2016 were treated as training datasets, respectively; and the remainder were testing datasets. We estimated the changing epidemiologic trends of SF based on annual percentage change (APC) and average APC (AAPC) using the Joinpoint regression program (version 4.8.0.1). We constructed the TBATS model with the "forecast," "tseries,", and "FinTS" packages in R software (version 3.4.3). The incidence rate ratio (IRR) with a 95% uncertainty limit (UL) before and after the SF outbreak was calculated using the method proposed by Armitage and Berry[6]. The mean absolute deviation (MAD), root mean square error (RMSE), mean absolute percentage error (MAPE), mean error rate (MER), and root mean square percentage error (RMSPE) were computed to compare the forecasting ability of both models.

During the study period, a notable increase in SF morbidity was detected, with an AAPC = 8.942 (95% UL: 5.995–11.971; t = 6.697, P < 0.001; Supplementary Figure S1, available in www. besjournal.com), and the highest morbidity (5.930/100,000 persons) occurred in 2019, which increased by a factor of nearly four compared with the lowest level (1.489/100,000 population) in 2004 (Supplementary Figure S2, available in www. besjournal.com). The SF epidemics remained relatively steady from 2004–2010 (average, 1.937/100,000 persons annually), with an AAPC = −0.840 (95% UL: −7.678 to 6.505; t = −0.231, P = 0.817). An unexpected outbreak was witnessed in 2011, and since then a rapidly increasing trend occurred (average, 4.580/100,000 persons annually [excluding 2013]; the exact causes regarding this annual drop are unknown; Supplementary Figure S2), with an AAPC = 5.952 (95% UL: 0.239–11.992; t = 2.466, P = 0.043), which showed good agreement with a resurgence of SF in Hong Kong, China[7], but inconsistent with England where the resurgence occurred in 2014[8]. There has been a doubling in the SF incidence during the post-resurgence periods compared with the pre-resurgence periods (IRR = 2.364, 95% UL: 2.358–2.370). Nevertheless, the driving force associated with the increased pathogenicity of GAS fails to be elucidated. A possible explanation may be due to the acquisition of novel prophages harboring new hybridizations of toxin genes and antimicrobial resistance genes, which is related to the emergence and expansion of the predominant genotypes of *emm12* and *emm1* in China[3]. Another explanation may be associated with the natural periodicity of the SF incidence (SF epidemics are characterized by a cyclic change of approximately 6 years[3]). A third explanation may be linked to the relaxation of the 2-child policy in 2011[3], which led to an increase in the number of susceptible individuals. A fourth reason may be attributed to improvements in the diagnostic capacity and the increased awareness of medical workers in reporting SF[3]. A fifth reason may be a result of the deterioration of air quality in China[9], despite the gradual improvements in the last 2 years. Finally, there are no vaccines available to prevent infections with GAS until now.

A marked semi-annual seasonal behavior occurred in the monthly SF incidence, with a strong peak between May and June, and a weak peak between November and December (Supplementary Figures S3–S4, available in www.besjournal.com). We surmised that different climatic features and beginning of spring and autumn semesters contributed to this difference in the margin of peak activities. Our seasonal profile correlates well with previous findings from Hong Kong, China[7]; however, discordant with that in England, which peaked between February and March[8]. This inconsistency may be due to the different school breaks, population density, and different GAS *emm* gene types in east Asia and Europe[1,8]. In addition, the SF epidemics retain the lowest level in February every year (Supplementary Figure S3), attributable to the winter holidays and the Spring Festival.

The forecasts under the TBATS approach rely largely on the number of harmonics k_i applied for each seasonal pattern. As a result, in selecting the number of harmonics k_i, considering one seasonal component each time, we then fitted the model on the target data repeatedly *via* gradually increasing the number of harmonics k_i, but holding the remaining harmonics constant for each i until the optimal AIC is obtained. In determining the most suitable orders (p and q) of the ARMA model, we used the automatic procedure proposed by Hyndman and colleagues[10] to fit the forecasting residuals. If the selected model with the ARMA (p, q) residual component generates a smaller AIC than the one model without the ARMA (p, q) residual component, this selected specification would be considered as the best possible model; otherwise, the ARMA (p, q) residual component is deleted. After modelling by trial and error, the TBATS (0.04, {4,0}, 0.882, {<12,5>}) specification was selected as the optimal model in that the minimum AIC (−197.965) was detected in this model, and the identified key parameters of this best TBATS model are reported in

Supplementary Table S1, available in www.besjournal.com. Additional statistical diagnoses for the forecast errors are provided in Supplementary Table S2 and Supplementary Figures S5–S6, available in www.besjournal.com. The Ljung-Box Q statistics of the forecast errors produced a $Q_{(18)}$ = 11.442 with a *P*-value of 0.875, indicating no serial correlations in this residual series. Moreover, the ARCH effect was largely removed because the $LM_{(18)}$ = 23.808 with a *P*-value of 0.302. These results confirmed the adequacy of the model specifications. Similarly, based on the modelling steps described above, the TBATS (0.01, {0,0}, 0.898, {<12,5>}) and TBATS (0.048, {0,0}, 0.902, {<12,5>}) specifications tended to be the preferred models for forecasting the 12- and 36-holdout periods (Supplementary Tables S3–S4 and Supplementary Figures S7–S10, available in www.besjournal.com).

Similarly, following the SARIMA modelling steps, the optimal SARIMA models on different datasets were identified (Supplementary Tables S2–S5, and Supplementary Figures S5, S11–S12, available in www.besjournal.com). Subsequently, the best SARIMA and TBATS models could be used to perform multistep ahead predictions (Figure 1 and Supplementary Figures S13–S14, available in www.besjournal.com). Table 1 lists the measurement metrics, which indicate the forecasting reliability levels on different time windows under the preferred SARIMA and TBATS models. The optimal TBATS models provided a smaller MAD, MAPE, RMSE, RMSPE, and MER compared with the optimal SARIMA models, with a performance improvement of almost 50% in the forecasting abilities for estimating both short- and long-term epidemiological trends, albeit the predictive potential showed a slight reduction with the increase in prediction time windows. We further compared the forecasting abilities of both methods for 48- and 72-step ahead predictions, and the comparative results are listed in Supplementary Tables S1, S6–S7, and Supplementary Figures S15–S16 (available in www.besjournal.com), which show a similar finding, but the predictive results

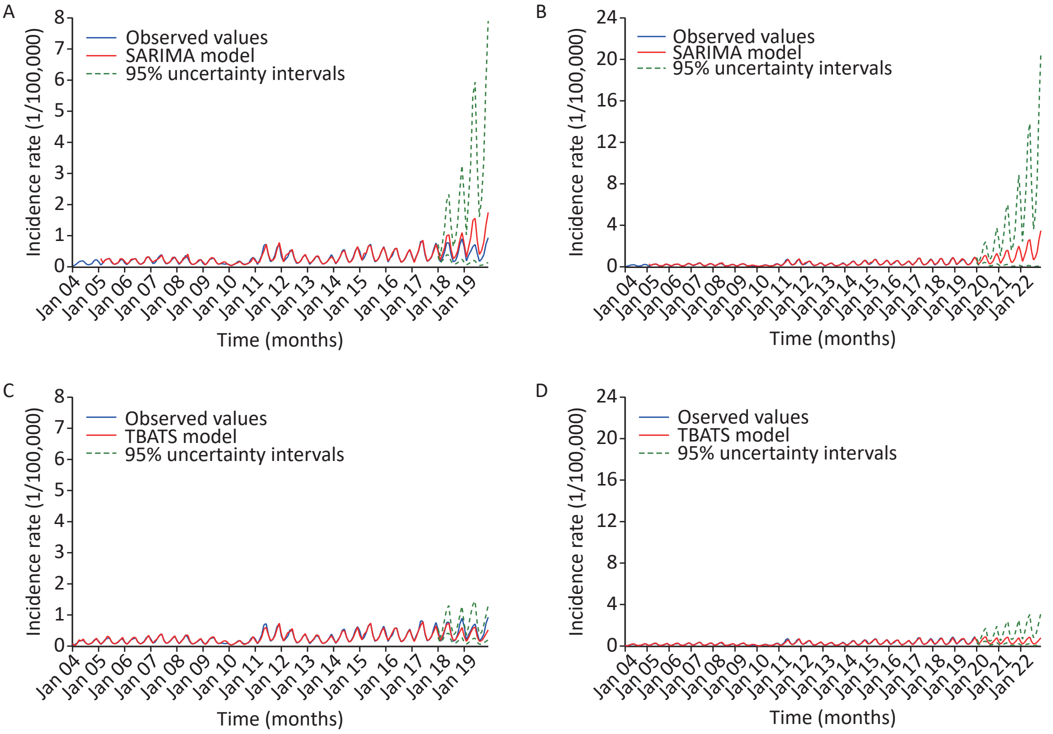

Figure 1. Comparative results of the forecasts based on the SARIMA and TBATS models. (A) The comparison between the 24-step ahead forecasts of the SARIMA model and the observed values. (B) The predicted upcoming 36-month values from January 2020 to December 2022 using the SARIMA model. (C) The comparison between the 24-step ahead forecasts of the TBATS model and the observed values. (D) The predicted upcoming 36-month values from January 2020 to December 2022 using the TBATS model.

deviated from the epidemic trajectories. In addition, we used the SF incidence data in Liaoning, Heilongjiang, and Shandong provinces, and Inner Mongolia (which are the hardest hit areas by SF in China in the last decade[2,3]) to assess the predictive quality of these two methods. Likewise, the TBATS method produced lower error rates in all the datasets (Supplementary Table S8, available in www.besjournal.com). Our recent study indicated that the Error-Trend-Seasonal (ETS) model also has a powerful potential in estimating the long-term epidemic behaviors of diseases[4]. As a result, we further developed the ETS model based on the SF morbidity to predict the epidemiological trends, and the results also showed similar findings (the computed MAPE values were 16.101% *vs*. 38.511%, 21.142% *vs*. 28.273%, and 23.984% *vs*. 26.735% in the 12-, 24-, and 36-step ahead forecasts, respectively; Supplementary Table S9, available in www.besjournal.com). These findings further substantiated the utility of the TBATS model. The TBATS model was introduced by adding the trigonometric representation of seasonal components based on the Fourier series into the traditional BATS model, which enabled handling of all complex time series, as well as linear and non-linear information[5], thus indicating the suitability and adequacy of this model. Considering the attractive advantages of the TBATS model, this model can be recommended as a flexible and useful long-term predictive tool in assessing the epidemic patterns of SF in other countries or other contagious diseases; however, further work is required for validation. Moreover, with the rapid advances in the forecasting domain of time series, many hybrid prediction models (e.g., SARIMA-BPNN, SARIMA-GRNN, and SARIMA-LSTM) have also been reported to show an attractive advantage in estimating the long-term epidemic trajectories of diseases. Therefore, what is now needed are studies involving comparisons of the predictive reliable level between the TBATS model and the above-mentioned models.

This study had some limitations. First, SF is a mild illness and has rarely led to death since the 20th

Table 1. The comparisons of the predicted results between the SARIMA model and the TBATS model on different testing datasets

Models	Testing horizons				
	MAD	MAPE	RMSE	MER	RMSPE
24-step ahead predictions					
SARIMA	0.295	64.346	0.382	0.607	0.551
TBATS	0.114	21.142	0.160	0.235	0.061
Reduced percentage (%)					
SARIMA *vs*. TBATS	61.356	67.143	58.115	61.285	88.929
12-step ahead predictions					
SARIMA	0.212	42.951	0.257	0.429	0.467
TBATS	0.087	16.101	0.113	0.177	0.193
Reduced percentage (%)					
SARIMA *vs*. TBATS	58.962	62.513	56.031	58.741	58.672
36-step ahead predictions					
SARIMA	0.258	55.527	0.396	0.546	0.758
TBATS	0.133	23.984	0.174	0.282	0.271
Reduced percentage (%)					
SARIMA *vs*. TBATS	48.450	56.807	56.061	48.352	64.248

Note. SARIMA, seasonal autoregressive integrated moving average method; TBATS, an advanced innovation state-space modelling framework by combining Box-Cox transformations, Fourier series with time-varying coefficients and autoregressive moving average (ARMA) error correction; MAD, mean absolute deviation; MAPE, mean absolute percentage error; RMSE, root mean square error; MER, mean error rate; RMSPE, root mean square percentage error.

century[3]. Therefore, infected individuals with mild clinical manifestations sometimes fail to seek medical aid, resulting in under-reporting and under-diagnosis. Second, to ensure that the model obtained a satisfactory forecasting result, it is important to note that this model should be updated with new incidence data. Third, to investigate whether our TBATS model was adequate for estimating the SF epidemics in other study regions or other infectious diseases, much work is still needed. Finally, integrating the factors influencing the SF epidemics may improve the predictive power. Nevertheless, we are not able to perform such an analysis due to the unavailability of a multivariate TBATS method and SF-related factors.

In summary, SF had dual seasonal behaviors, peaking in May–June and November–December, with a recurrence in 2011 in China; since then it started to be increasing in the SF incidence. The TBATS method was advantageous in analyzing the long-term epidemiological seasonality and trends of SF, which can be considered a useful and flexible alternative to aid stakeholders to develop practical solutions to stop the ongoing spread of SF in China. In addition, we re-established the preferred TBATS (0.023, {0,0}, 0.895, {<12,5>}) specification based on the 16 years of data to predict the SF incidence into 2022, although the SF incidence was predicted to reach a plateau in the next 3 years [Supplementary Tables S1 and S10 (available in www.besjournal.com), and Figure 1], the SF incidence remained at a high level, suggesting that additional or comprehensive interventions must be developed to manage this evolving scenario.

Data Availability All the data supporting the findings of the work are contained within the Supplementary Material (Supplementary Table S11, available in www.besjournal.com).

[&]These authors contributed equally to this work.

[#]Correspondence should be addressed to WANG Yong Bin, MD, Tel: 86-373-3831646, E-mail: wybwho@163.com; LIANG Wen Juan, MD, Tel: 86-373-3831646, E-mail: wenwen3_1@126.com

Biographical notes of the first authors: WANG Yong Bin, male, born in 1989, Professor, MD, majoring in infectious disease epidemiology; LI Yan Yan, female, born in 1996, master’s, majoring in infectious disease epidemiology; LU Hao, male, born in 1999, Undergraduate, majoring in preventive medicine.

Received: November 12, 2021;

Accepted: June 2, 2022

REFERENCES

1. Chen M, Cai J, Davies MR, et al. Increase of emm1 isolates among group A Streptococcus strains causing scarlet fever in Shanghai, China. Int J Infect Dis, 2020; 98, 305−14.
2. Li WT, Feng RH, Li T, et al. Spatial-temporal analysis and visualization of scarlet fever in mainland China from 2004 to 2017. Geospat Health, 2020; 15, 831.
3. Liu YH, Chan TC, Yap LW, et al. Resurgence of scarlet fever in China: a 13-year population-based surveillance study. Lancet Infect Dis, 2018; 18, 903−12.
4. Wang YB, Xu CJ, Yao SQ, et al. Estimating the prevalence and mortality of coronavirus disease 2019 (COVID-19) in the USA, the UK, Russia, and India. Infect Drug Resist, 2020; 13, 3335−50.
5. De Livera AM, Hyndman RJ, Snyder RD. Forecasting time series with complex seasonal patterns using exponential smoothing. J Am Stat Assoc, 2011; 106, 1513−27.
6. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th ed. Blackwell Publishing. 2002, 127.
7. Lee CF, Cowling BJ, Lau EHY. Epidemiology of reemerging scarlet fever, Hong Kong, 2005-2015. Emerg Infect Dis, 2017; 23, 1707−10.
8. Lamagni T, Guy R, Chand M, et al. Resurgence of scarlet fever in England, 2014-16: a population-based surveillance study. Lancet Infect Dis, 2018; 18, 180−7.
9. Liu YH, Ding H, Chang ST, et al. Exposure to air pollution and scarlet fever resurgence in China: a six-year surveillance study. Nat Commun, 2020; 11, 4229.
10. Hyndman R, Athanasopoulos G, Bergmeir C, et al. Forecast: forecasting functions for time series and linear models. https://cran.r-project.org/web/packages/forecast/. [2022-05-10].