Epidemiol. Infect. (2016), **144**, 144–151. © Cambridge University Press 2015
doi:10.1017/S0950268815001144

Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model

L. LIU[1,2]*, R. S. LUAN[1], F. YIN[1], X. P. ZHU[2] AND Q. LÜ[2]

[1] *West China School of Public Health, Sichuan University, Chengdu, Sichuan, People's Republic of China*
[2] *Sichuan Center for Disease Control and Prevention, Chengdu, Sichuan, People's Republic of China*

Received 8 July 2014; Final revision 1 February 2015; Accepted 10 May 2015; first published online 1 June 2015

SUMMARY

Hand, foot and mouth disease (HFMD) is an infectious disease caused by enteroviruses, which usually occurs in children aged <5 years. In China, the HFMD situation is worsening, with increasing number of cases nationwide. Therefore, monitoring and predicting HFMD incidence are urgently needed to make control measures more effective. In this study, we applied an autoregressive integrated moving average (ARIMA) model to forecast HFMD incidence in Sichuan province, China. HFMD infection data from January 2010 to June 2014 were used to fit the ARIMA model. The coefficient of determination (R^2), normalized Bayesian Information Criterion (BIC) and mean absolute percentage of error (MAPE) were used to evaluate the goodness-of-fit of the constructed models. The fitted ARIMA model was applied to forecast the incidence of HMFD from April to June 2014. The goodness-of-fit test generated the optimum general multiplicative seasonal ARIMA $(1,0,1) \times (0,1,0)_{12}$ model ($R^2 = 0{\cdot}692$, MAPE = 15·982, BIC = 5·265), which also showed non-significant autocorrelations in the residuals of the model ($P = 0{\cdot}893$). The forecast incidence values of the ARIMA $(1,0,1) \times (0,1,0)_{12}$ model from July to December 2014 were 4103–9987, which were proximate forecasts. The ARIMA model could be applied to forecast HMFD incidence trend and provide support for HMFD prevention and control. Further observations should be carried out continually into the time sequence, and the parameters of the models could be adjusted because HMFD incidence will not be absolutely stationary in the future.

Key words: ARIMA; forecasting; hand, foot and mouth disease (HFMD).

INTRODUCTION

Hand, foot and mouth disease (HFMD) is a cluster of manifestations with fever and exanthema of the hands, feet and mouth caused by human enteroviruses. It commonly occurs in children aged <5 years throughout the world [1–3]. In China, HFMD had been a nationally notifiable disease since 2008, and cases must be reported to the National Disease Surveillance Reporting and Management System within 24 h. Since 2010, HFMD has become the infectious disease with the largest number of cases in Sichuan province, China. More than 40 000 cases were reported annually with an incidence rate of over 50/100 000 population, causing a serious social and economic burden.

Epidemic forecasting models were regarded as important tools to predict the occurrence of infectious diseases and formulate reasonable short-term or long-term precautions. Certain factors such as speed of pathogen

* Author for correspondence: Mr L. Liu, Sichuan Center for Disease Control and Prevention, West China School of Public Health, No. 6 Zhongxue Road, Chengdu, Sichuan 610041, People's Republic of China, 610041.
(Email: sheva_liulei@126.com)

variation, accumulation of susceptible hosts and environmental change allow infectious diseases to be modelled [4]. More statistical methods have been used for incidence prediction of infectious diseases, including linear regression, correlation coefficient analysis, grey swing models and back propagation artificial neural network models [5–8]. In addition to these, autoregressive integrated moving average (ARIMA) models, which consider changing trends, periodic changes and random disturbances, have been widely used in modelling the temporal dependence structure of time series.

Our study developed a stochastic ARIMA model and then forecast the HFMD incidence in Sichuan province. To the best of our knowledge, this study is the first to apply an ARIMA model to fit and predict HFMD incidence in this area, whereas only Huang *et al.* used ARIMA to predict Chinese HFMD incidence prior to our research [9].

MATERIALS AND METHODS

Data collection

The observed monthly cases of HFMD were extracted from the Chinese National Disease Surveillance Reporting and Management System from January 2010 to June 2014.

Method

The ARIMA model was designed to take advantage of the associations in the sequentially lagged relationships that exist in periodically collected data [10]. Autoregressive (AR) specifies that the output variable depends linearly on its previous values, while moving average (MA) is the linear regression of the current value of the series against current and previous terms. If the raw data showed evidence of non-stationarity, a differencing step (the integrated part of the model) should be applied to remove it.

There are three parameters in the ARIMA model: p, d and q, which refer to the order of the autoregressive, integrated and moving average parts of the model, respectively. As the data showed evidence of seasonal tendency, the general multiplicative seasonal model [ARIMA $(p,d,q)\times(P,D,Q)_s$)] was used. The parameter P indicates the order of seasonal autoregression; D, the degree of seasonal difference; and Q, the order of seasonal moving average [11].

The process of predicting incidence by the ARIMA model consists of four steps. First, the original time series should be transformed by square root or logarithmic algorithm, difference or seasonal difference to induce stationarity. Second, the model identification used autocorrelation function (ACF) and partial autocorrelation function (PACF) analyses to analyse the random, stationary and seasonal effects on the time-series data. Third, the most appropriate model would be selected using parameter estimation and model testing. Diagnostic checking parameters, which include the coefficient of determination (R^2), normalized Bayesian Information Criterion (BIC) and mean absolute percentage of error (MAPE) were used to compare the goodness-of-fit of the models and determine statistical significance ($P < 0{\cdot}05$). The optimum model should have the highest R^2 and the lowest MAPE and BIC. The Ljung–Box test was used to diagnose whether the residual error sequence was a white-noise sequence. Finally, predictive analysis could be conducted with the optimum parameter combination [12–14].

The data were analysed using the appropriate module in SPSS version 19.0 (SPSS Inc., USA) and R version 3.1.2 (R Foundation, Austria).

RESULTS

Spatio-temporal pattern

The monthly HFMD incidence from January 2010 to June 2014 in Sichuan province was between 280 and 10 929, with an incidence rate between 0·3482 and 13·3525/100 000 population.

In terms of the temporal distribution of the HFMD cases, Figure 1 shows that the monthly incidence exhibited obvious seasonal fluctuation, with peaks between May and July.

Figure 2 shows the spatial distribution of the annual incidence rate of HFMD cases at the municipal level. High-prevalence areas included the provincial capital Chengdu and its surrounding cities, as well as the northeastern and southwestern areas of Sichuan province.

Model identification

Time-series data from January 2010 to June 2014 was used as the training set. Given that the raw data had a non-stationary variance, it was converted into a square root to reduce the variance.

Figure 3 shows that the ACF and PACF of the original data had a non-stationary variance, and the effects of linear and seasonal trends were eliminated by taking the difference and seasonal difference. Figures 4 and 5 show that the ACF and PACF of

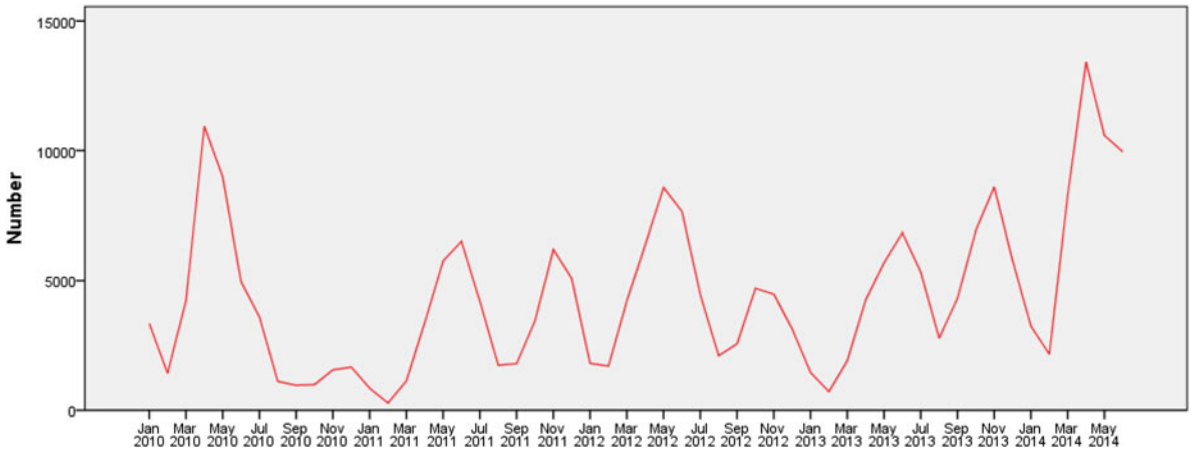

Fig. 1. Monthly HFMD incidence from January 2010 to June 2014 in Sichuan province.

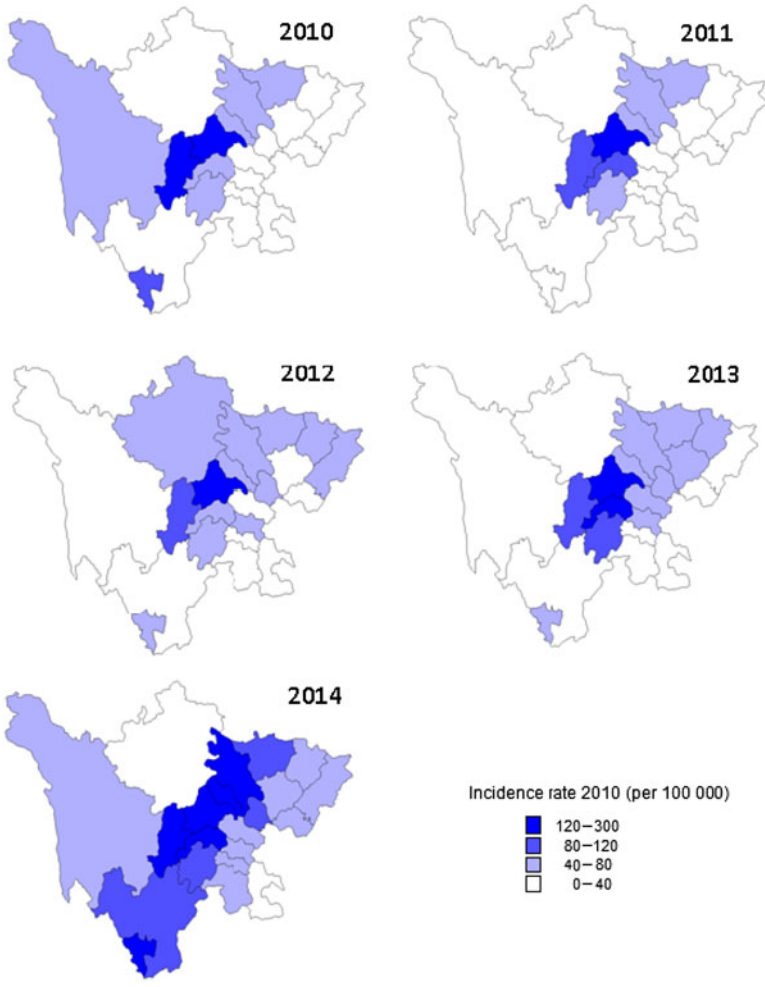

Fig. 2. Annual incidence rates of HFMD at the municipal level.

the new data tended to be stationary after using the first-order seasonal difference, which determined that d and D in the ARIMA $(p,d,q)\times(P,D,Q)_{12}$ have values of 0 and 1, respectively.

Model diagnosis

Based on the distribution characteristics, several model algorithms with the parameters p, q, P and Q with non-negative integer values from 0 to 3 were

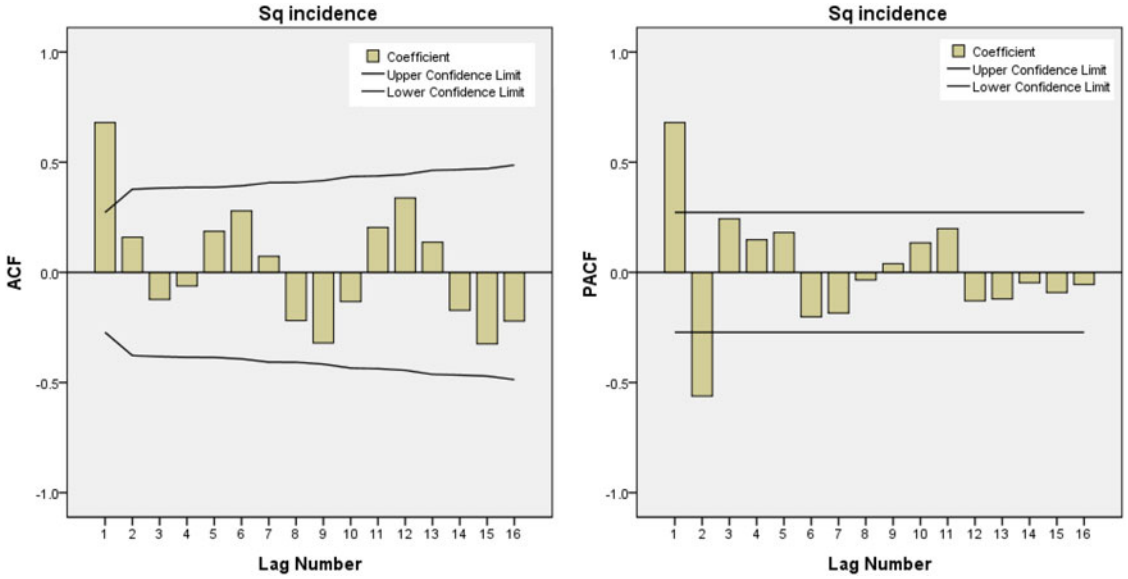

Fig. 3. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the square root of monthly incidence.

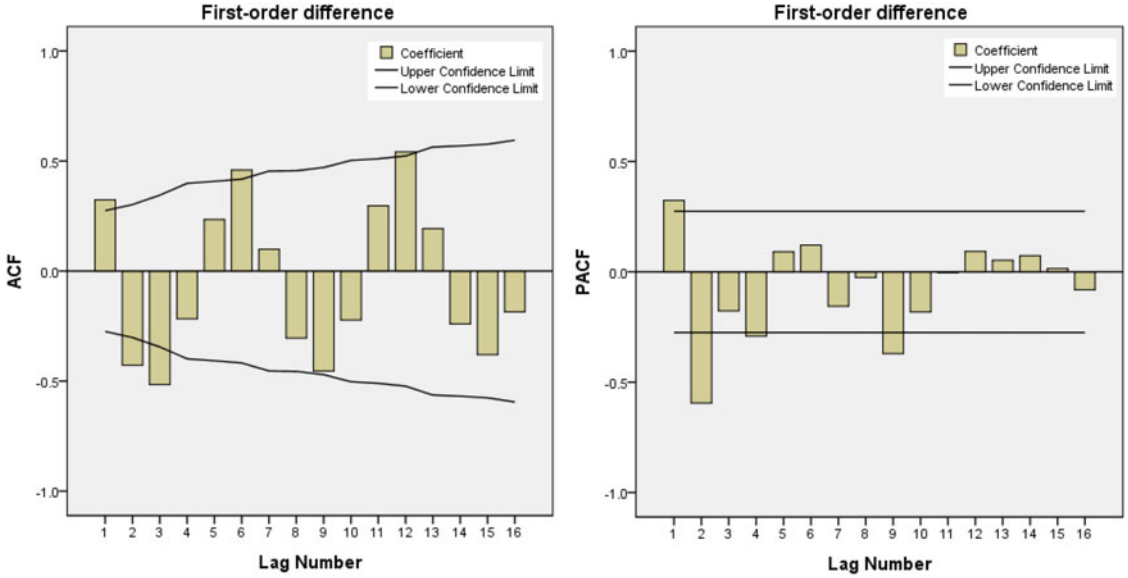

Fig. 4. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the square root of incidence after the first-order difference.

used. Table 1 shows the parameter estimation for plausible ARIMA models. Seven models were found to have statistically significant parameters.

Based on the results of the goodness-of-fit test statistics, we confirmed the optimal ARIMA $(1,0,1) \times (0,1,0)_{12}$ model, which had the highest R^2 (0·692), lowest BIC (15·982) and relatively low MAPE (5·265) among the seven plausible models (Table 2). The Ljung–Box test also showed that its residual was white noise with $Q = 9{\cdot}456$ ($P = 0{\cdot}893$), indicating that the fitted data series was stationary, random and zero-related. Figure 6 shows that the ACF and PACF of the residual sequence fell within the random confidence interval.

Furthermore, with the existence of seasonal trends, we conducted a sine function to fit the same series and obtained two equations:

$$y = 61{\cdot}93284 + 14{\cdot}77089^{*}\sin(0{\cdot}51806^{*}t - 1{\cdot}20027),$$

$$y = 45{\cdot}59231 + 0{\cdot}59733^{*}t + 15{\cdot}39398^{*}\sin(0{\cdot}50603^{*}t, -0{\cdot}94194).$$

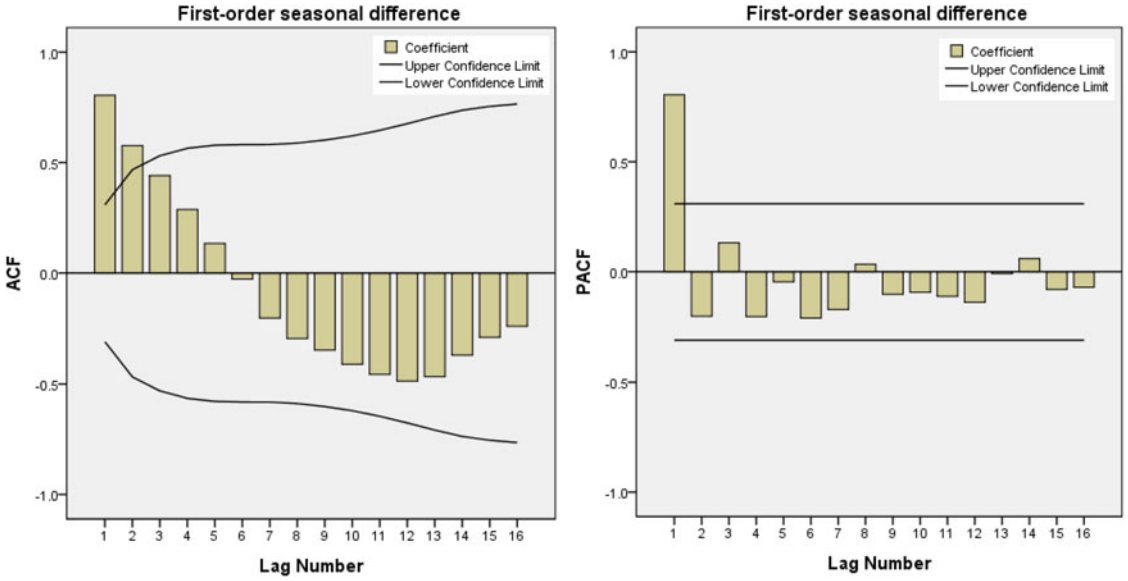

Fig. 5. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the square root of incidence after the first-order seasonal difference.

Table 1. *Parameter estimation for plausible ARIMA models*

	AR		MA		Seasonal AR		Seasonal MA	
	B	*P*	*B*	*P*	*B*	*P*	*B*	*P*
$(0,0,0)\times(1,1,0)_{12}$	—	—	—	—	−0·842	0·000	—	—
$(0,0,0)\times(2,1,0)_{12}$	—	—	—	—	0·376	0·049	—	—
$(0,0,1)\times(0,1,0)_{12}$	—	—	−0·874	0·000	—	—	—	—
$(0,0,1)\times(1,1,0)_{12}$	—	—	−0·805	0·000	−0·733	0·000	—	—
$(0,0,2)\times(0,1,0)_{12}$	—	—	0·501	0·002	—	—	—	—
$(1,0,0)\times(0,1,0)_{12}$	0·845	0·000	—	—	—	—	—	—
$(1,0,1)\times(0,1,0)_{12}$	0·694	0·040	−0·551	0·001	—	—	—	—

AR, Autoregressive; MA, moving average.

Table 2. *Goodness-of-fit statistics for plausible ARIMA models*

Statistic	$(0,0,0)\times(1,1,0)_{12}$	$(0,0,0)\times(2,1,0)_{12}$	$(0,0,1)\times(0,1,0)_{12}$	$(0,0,1)\times(1,1,0)_{12}$	$(0,0,2)\times(0,1,0)_{12}$	$(1,0,0)\times(0,1,0)_{12}$	$(1,0,1)\times(0,1,0)_{12}$
Stationary R^2	0·318	0·341	0·556	0·651	0·650	0·657	0·692
MAPE	24·443	23·471	20·156	17·927	17·677	17·306	15·982
Normalized BIC	5·947	6·027	5·517	5·392	5·395	5·259	5·265

MAPE, Mean absolute percentage of error; BIC, Bayesian Information Criterion.

From the results of the goodness-of-fit, the BIC values of the sine function were significantly larger than those of the ARIMA model.

Forecast and analysis

The ARIMA $(1,0,1)\times(0,1,0)_{12}$ model was applied to forecast HFMD incidence, and its predicted data and 95% confidence limits from July to December 2014 are shown in Table 3 and Figure 7. The predicted data matched the actual data well except for August, but the actual incidence in August still fell within the predicted 95% confidence interval. The average relative error was 15·31, which was also less than that of the other six models (15·52–28·52).

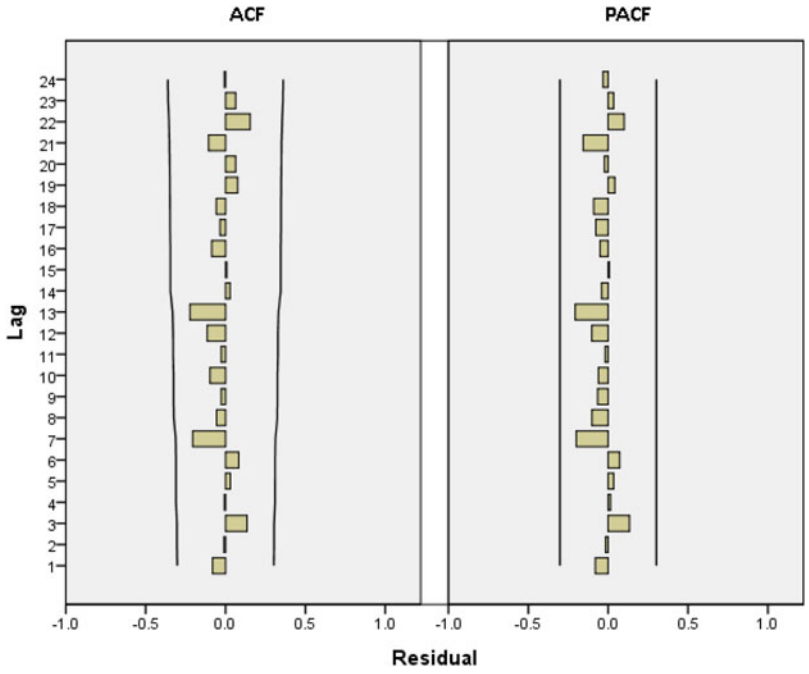

Fig. 6. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the residual series of the ARIMA $(1,0,1) \times (0,1,0)_{12}$ model.

Table 3. *Comparison of the predicted and actual values of the ARIMA* $(1,0,1) \times (0,1,0)_{12}$ *model*

	Actual incidence	Forecast	LCL	UCL
July	7551	7583	4152	12 039
August	2810	4103	778	10 041
September	5236	5656	1163	13 529
October	10 424	8348	2312	18 132
November	11 259	9987	3095	20 804
December	7278	6826	1432	16 230
Average relative error			15·31	

LCL, Lower confidence limit; UCL, upper confidence limit.

DISCUSSION

Since 2008, the incidence of HFMD has continued to rise in China, causing widespread social concern and fear [15]. Hence, exploring effective prediction methods of HFMD has great practical significance. Time-series prediction is very helpful in developing hypotheses to explain and anticipate the dynamics of the observed and subsequent disease status because it is based on the changes in historical datasets over time and produces mathematical models that can be extrapolated [16, 17]. The ARIMA model has been used widely in medical research as it provides a comprehensive model in time-series analysis [18–20]. Through the ARIMA procedure, we can directly foresee the future incidence trend and develop niche targeting preventive and control measures. In addition, the 95% confidence interval range of the predictive value can also be used for early warning of infectious diseases. An infection during a specific time period that is over the upper limit of the confidence interval may be a warning that a certain cluster or outbreak may occur.

Several studies have used ARIMA models to fit and predict the changing trends of infectious diseases and achieved good results [21, 22]. In our study, we obtained a multiplicative ARIMA model to eliminate the influence of seasonal tendency. A seasonal module ensured that the ARIMA model provided the best forecast possible. Based on the testing results, the conducted model $(1,0,1) \times (0,1,0)_{12}$ was reliable with high validity and can be used to forecast the expected numbers of cases. The forecast results also matched the actual data well.

However, few predictions would occasionally have a significant difference from the actual value even by the best-fit model, such as the forecast in August 2014, which was obviously higher than the actual level in our study. Many natural, social and environment factors could affect the incidence of HMFD, including the pathogen, environment and host factor

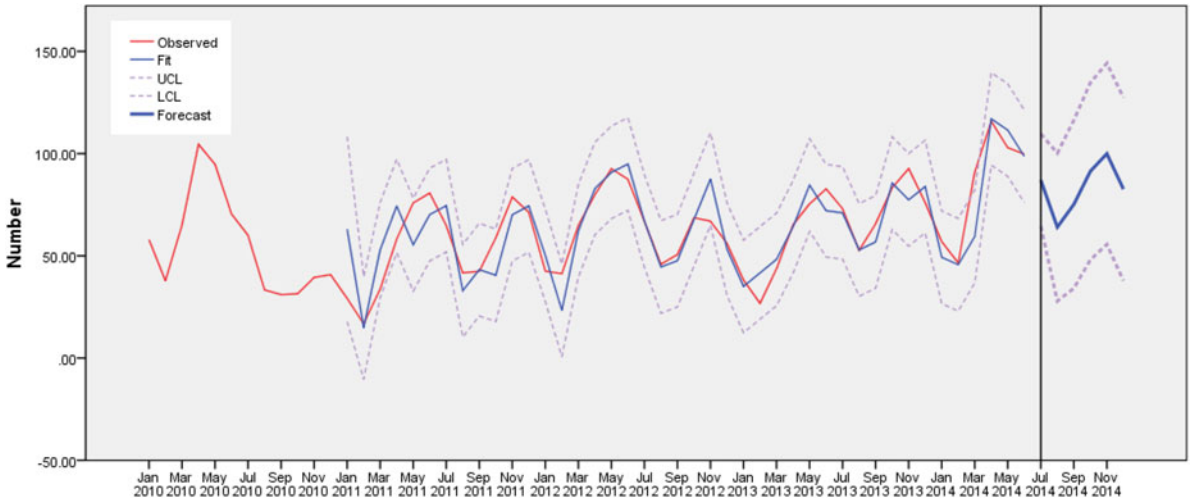

Fig. 7. Observed and predicted value of the ARIMA $(1,0,1) \times (0,1,0)_{12}$ model. UCL, Upper confidence limit; LCL, lower confidence limit.

[23–25]. More than 30 different enteroviruses can cause HFMD; the most common pathogens being EV71 and CoxA16. The capacity of each pathogen to influence disease transmission varies. In Sichuan province, major pathogens presented alternative states in different periods, causing incidence fluctuations. The behaviour of susceptible populations could exacerbate transmission risk, e.g. not washing the hands often or contact with patients and their items. Many studies mentioned that temperature, humidity and other weather conditions could also influence the spread of HFMD. High temperature and relative humidity are generally considered to increase the incidence [26–28]. In addition, Chinese schools have two vacations annually: the summer vacation from January to February and the winter vacation from July to September. During these times, children usually stay at home, with few opportunities for contact with one another; thus, the incidence of HFMD may fall.

Data quality is also significant. First, in order to conduct an effective ARIMA model, the time sequence must be sufficiently long, and observations should be added continually to the sequence over time. Consequently, new parameter combinations of the ARIMA model might be refitted to the new series. Second, missing reports of the Chinese National Disease Surveillance Reporting and Management System also affected the results of the forecast. According to our survey of infectious disease reporting quality in Sichuan province, the missing rates of HFMD were between 1·7% and 8·5% from 2010 to 2014. When evaluating the disease trend, missing reports should also be taken into account.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Seong JK, *et al.*** Risk factors for neurologic complications of hand, foot and mouth disease in the Republic of Korea, 2009. *Journal of Korean Medical Science* 2013; **28**: 120–127.
2. **Ji H, *et al.*** Seroepidemiology of human enterovirus71 and coxsackievirus A16 in Jiangsu province, China. *Virology Journal* 2012; **9**: 248–256.
3. **Peng JP, *et al.*** Sensitive and rapid detection of viruses associated with hand foot and mouth disease using multiplexed MALDI-TOF analysis. *Journal of Clinical Virology* 2013; **56**: 170–174.
4. **Guan P, Huang DS, Zhou BS.** Forecasting model for the incidence of hepatitis A based on artificial neural network. *World Journal of Gastroenterology* 2004; **10**: 3579–3582.
5. **Wang YJ, *et al.*** Applying linear regression statistical method to predict the epidemic of hemorrhagic fever with renal syndrome. *Chinese Journal of Vector Biology and Control* 2006; **17**: 333–334.
6. **Clement J, *et al.*** Relating increasing hantavirus incidences to the changing climate: the mast connection. *International Journal of Health Geographics* 2009; **8**: 1.
7. **Guo LC, *et al.*** Appling grey swing model to predict the incidence trend of hemorrhagic fever with renal syndrome in Shenyang. *Journal of China Medical University* 2008; **37**: 839–842.

8. **Wu ZM, *et al.*** Prediction for incidence of hemorrhagic fever with renal syndrome with back propagation artificial neural network model. *Chinese Journal of Vector Biology and Control* 2006; **17**: 223–226.
9. **Huang XX, *et al.*** Prediction of monthly hand foot and mouth disease incidence in China by using autoregressive integrated moving average model. *Disease Surveillance* 2013; **28**: 396–399.
10. **Akhtar S, Rozi S.** An autoregressive integrated moving average model for short-term prediction of hepatitis C virus seropositivity among male volunteer blood donors in Karachi, Pakistan. *World Journal of Gastroenterology* 2009; **15**: 1607–1612.
11. **Wentong Z.** *The Course of Statistical Analysis with SPSS. Beijing*, China: Hope Electronic Press, 2002, pp. 250–289.
12. **Chatfield C.** *The Analysis of Time Series: Theory and Practice*. London: Chapman and Hall.
13. **Jenkins GW, Reinsel GC, Box GEP.** *Time Series Analysis*, 3rd edn. South Winsdor, New South Wales, Australia: Holden Day.
14. **Bowerman BL, O'Connell R.** *Forecasting and Time Series: An Applied Approach*. Boston: South-Western College Publications.
15. **Chang ZR, *et al.*** Epidemiological features of hand foot and mouth disease in China, 2008–2009. *China Journal of Epidemiology* 2011; **32**: 676–680.
16. **Kuhn L, Davidson LL, Durkin MS.** Use of Poisson regression and time series analysis for detecting changes over time in rates of child injury following a prevention program. *American Journal of Epidemiology* 1994; **140**: 943–955.
17. **Liang W, *et al.*** Prediction of malaria incidence in malaria epidemic area with time series models. *Journal of the Fourth Military Medical University* 2004; **25**: 507–510.
18. **Silawan T, *et al.*** Temporal patterns and forecast of dengue infection in northeastern Thailand. *The Southeast Asian Journal of Tropical Medicine and Public Health* 2008; **39**: 90–98.
19. **Wong J, Chan A, Chiang YH.** Time series forecasts of the construction labor market in Hong Kong: the Box-Jenkins approach. *Construction Management and Economics* 2005; **23**: 979–991.
20. **Mumbare SS, *et al.*** Trends in average living children at the time of terminal contraception: a time series analysis over 27 years using ARIMA (p,d,q) nonseasonal model. *Indian Journal of Community Medicine* 2014; **39**: 223–228.
21. **Earnest A, *et al.*** Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research* 2005; **5**: 36.
22. **Li XJ, *et al.*** A time series model in incidence forecasting of hemorrhagic fever with renal syndrome. *Journal of Shandong University (Health Sciences)* 2008; **46**: 547–549.
23. **Hii YL, Rockl J, Ng N.** Short term effects of weather on hand, foot and mouth disease. *PLoS ONE* 2011; **2**: 1–6.
24. **Wang JF, *et al.*** Hand, foot and mouth disease: spatio-temporal transmission and climate. *International Journal of Health Geographics* 2011; **10**: 25–34.
25. **Zou XN, *et al.*** Etiologic and epidemiologic analysis of hand foot and mouth disease in Guangzhou city: a review of 4,753 cases. *Brazilian Journal of Infectious Diseases* 2012; **16**: 457–465.
26. **Onozuka D, Hashizume M.** The influence of temperature and humidity on the incidence of hand, foot, and mouth disease in Japan. *Science of the Total Environment* 2011; **410**: 119–125.
27. **Zhu SB, Xiang FL.** Multiple linear regression analysis of hand, foot and mouth diseases and the meteorological factors. *Zhejiang Preventive Medicine* 2013; **3**: 37–39.
28. **Feng GS, Yu SC, Hu YH.** Application of panel data model in the study of the relationship between reported hand-foot-mouth morbidity and temperature. *Chinese Preventive Medicine* 2013; **12**: 910–913.