Original Article

Using an Autoregressive Integrated Moving Average Model to Predict the Incidence of Hemorrhagic Fever with Renal Syndrome in Zibo, China, 2004–2014

Tao Wang, Yunping Zhou, Ling Wang, Zhenshui Huang, Feng Cui, and Shenyong Zhai*

Zibo Center for Disease Control and Prevention, Zibo, People's Republic of China

SUMMARY: Hemorrhagic fever with renal syndrome (HFRS) is highly endemic in mainland China, where human cases account for 90% of the total global cases. Zibo City is one of the most seriously affected areas in Shandong Province, China. Therefore, there is an urgent need for monitoring and predicting HFRS incidence in Zibo to make the control of HFRS more effective. In this study, we constructed an autoregressive integrated moving average (ARIMA) model for monthly HFRS incidence in Zibo from 2004 to 2013. The ARIMA (3,1,1) $\times$ (2,1,1)₁₂ model is reliable with a high validity, which can be used to predict the next year's HFRS incidence in Zibo. The forecast results suggest that the HFRS incidence in Zibo will experience a slight growth in the next year.

INTRODUCTION

Hemorrhagic fever with renal syndrome (HFRS), a rodent-borne disease caused by hantaviruses (family Bunyaviridae), is characterized by fever, acute renal dysfunction, and hemorrhage manifestations (1–3). Various rodent species are natural hosts and serve as sources of infection (4). In China, the causative agents of HFRS are predominately Hantaan virus (HTNV) and Seoul virus (SEOV), which result in fatality rates of approximately 10% and 1%, respectively (5). HTNV causes a more severe form of HFRS than SEOV and is associated with *Apodemus agrarius*, whereas SEOV is typically carried by *Rattus norvegicus* (6–8). Hantaviruses are primarily transmitted from rodent host to human by aerosols generated from contaminated urine and feces, and possibly from contaminated food or rodent bites (9–11).

HFRS was first recognized in northeastern China in 1931 and has been prevalent in many other parts of China since 1955. At present, it is highly endemic in mainland China, accounting for 90% of the total cases reported in the world (9,11–13). In response to the spread of HFRS in China, the Chinese Center for Disease Control and Prevention (CDC) established the National Notifiable Disease Surveillance System in 2004, which made the surveillance data for HFRS more accurate and comprehensive. Surveillance and early warning are essential for controlling or reducing the risk of outbreaks (14). Early warnings of infectious diseases should be provided based on the analysis of surveillance information. These early warnings can provide a scientific basis for better decision making. Thus, epidemic modeling and forecasting can be essential tools to prevent and control HFRS (15). Autoregressive integrated

moving average (ARIMA) models, which take into account changing trends, periodic changes, and random disturbances in time series, are very useful in modeling the temporal dependence structure of a time series (15). In epidemiology, ARIMA models have been successfully applied to predict the incidence of infectious diseases, such as HIV (16), influenza (17,18), malaria incidence (19), and others (20–22).

This study aimed to fit ARIMA models and predict the HFRS epidemic trend by using SAS version 9.2 (SAS Institute, Cary, NC, USA). Our study was based on HFRS epidemic data from Zibo City, China, where the data can provide a basis for HFRS prevention and control.

MATERIALS AND METHODS

Materials: The study site is located in Zibo City (latitude 35°55'~37°17'N and longitude 117°32'~118°31'E), in the central part of Shandong province. Monthly HFRS cases reported during 2004–2014 were provided by the Zibo CDC and we were permitted to use the data. In China, HFRS is a nationally notifiable disease and hospital physicians must report every case to the local health authority within 24 hours. All HFRS cases were initially diagnosed according to the clinical symptoms. Patient blood samples were also collected and sent to local CDC laboratories for serological confirmation. Finally, data were collected by case number according to the sampling results.

Methods: ARIMA models are the most commonly used time series prediction models (16). We constructed ARIMA models for monthly HFRS incidence in Zibo from 1983 to 2012. ARIMA was designed to deal with highly seasonal data (23). An ARIMA (p, d, q) model comprises 3 types of parameters (15,23,24): the autoregressive parameters (p), number of differencing passes (d), and moving average parameters (q). The multiplicative seasonal ARIMA (p, d, q) $\times$ (P, D, Q)_s model is an extension of the ARIMA method of time series in which a pattern repeats seasonally over time (6,22,23). Analogous to the simple ARIMA parameters, the seasonal parameters are: seasonal autoregressive (P), seasonal

Received December 12, 2014. Accepted July 6, 2015.
J-STAGE Advance Publication September 11, 2015.

DOI: 10.7883/yoken.JJID.2014.567

*Corresponding author: Mailing address: Zibo Center for Disease Control and Prevention, Zibo, People's Republic of China. Tel: (+86) (533) 2282667, Fax: (+86) (533) 2283253, E-mail: zibocdc@163.com

differencing (D), and seasonal moving average parameters (Q). The length of the seasonal period is represented by s . For example, the incidence of infectious disease varies in the annual cycle, so $s = 12$ in the present study.

We used the Box-Jenkins strategy to construct models. The ARIMA model procedure consists of 3 iterative steps (6,22,24): identification, estimation, and diagnostic checking. Prior to fitting the ARIMA model, an appropriate difference of the series is usually performed to make the series stationary. Identification is the process of determining seasonal and non-seasonal orders using the autocorrelation functions (ACF) and partial autocorrelation functions (PACF) of the transformed data. Parameters in the ARIMA model(s) are estimated with the conditional least squares method after the identification step. At the diagnosis stage, the adequacy of the established model for the series is verified by employing white noise tests to check whether the residuals are independent and normally distributed. It is possible that several ARIMA models may be identified and the selection of an optimum model is necessary. Such selection of models is usually based on the Akaike information criterion (AIC) and Schwarz Bayesian criterion (SBC). Smaller AIC and SBC statistics indicate the better fitting model (14).

Finally, the fitted ARIMA model was used for short-term forecasting of the monthly HFRS incidence between January and December 2014. In addition, the predictive validity of the model was evaluated by using the root mean square error (RMSE) and the mean absolute percentage error (MAPE). Lower RMSE and MAPE values indicate a better fit of the data (6). All analyses were performed using SAS 9.2 with a significant level of $p < 0.05$.

RESULTS

Model identification: Time series data for HFRS covering 2004–2013 in Zibo were used as the training set and monthly data for 2014 were used as the test set (Fig. 1). The sequence diagrams and seasonal characteristics of HFRS incidence indicated that the data series had a seasonal cycle every 12 months. The white noise hypothesis is rejected very strongly, which is expected because the series is non-stationary. On the basis of these characteristics, we eliminated the effect of seasonal trends by taking one-order trend difference and one-order seasonal difference. The transformed series showed far less dispersion than the original series (Fig. 2). Plausible models, i.e., the ARIMA (3,1,1) $\times$ (2,1,0)₁₂, ARIMA (1,1,0) $\times$ (1,1,1)₁₂, ARIMA (3,1,0) $\times$ (2,1,1)₁₂, and ARIMA (3,1,1) $\times$ (2,1,1)₁₂, were identified on the basis of ACF and PACF (Figs. 3 and 4), and were used for further analysis.

Parameter estimation and model testing: On the basis of parameter estimation and goodness of fit test statistics (Tables 1 and 2), we confirmed that the best model was ARIMA (3,1,1) $\times$ (2,1,1)₁₂. The model diagnostic tests showed that all of the parameter estimates were significant and the residual series was white noise at different lags. Thus, there is no point in trying more complex models.

Forecast analysis: We used ARIMA model (3,1,1) $\times$ (2,1,1)₁₂ and time series data for 2004–2013 as the training set; the data from January–December 2014 were used as the test set (Fig. 5 and Table 3). The predicted data and the actual data for 2014 are shown in Table 3. The RMSE value was 0.045 and the MAPE value was 11.50%, so the predicted values are well matched to the actual values.

Fig. 1. Time series of monthly HFRS incidence in Zibo City from 2004 to 2013 (per 100,000 population).

ARIMA Model for Predicting HFRS Incidence

Fig. 2. One-order difference of monthly HFRS incidence in Zibo City from 2004 to 2013.

Lag	Covariance	Correlation	Autocorrelations																	Std Error				
			-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6		7	8	9	1
0	0.030469	1.00000												*****										0
1	-0.0063227	-.20751									****			.										0.096674
2	-0.0047291	-.15521									.***			.										0.100750
3	-0.0050345	-.16523									.***			.										0.102961
4	0.0035073	0.11511									.	**		.										0.105410
5	-0.0054564	-.17908									****			.										0.106578
6	0.0052436	0.17209									.	***		.										0.109354
7	0.0022407	0.07354									.	*		.										0.111857
8	-0.0034815	-.11426									. **			.										0.112308
9	0.00029517	0.00969									.			.										0.113389
10	0.0031844	0.10451									.	**		.										0.113397
11	-0.0000798	-.00262									.			.										0.114293
12	-0.0068854	-.22598									*****			.										0.114294
13	0.0067474	0.22145									.	****		.										0.118396
14	-0.0025564	-.08390									. **			.										0.122206
15	0.0023438	0.07692									.	**		.										0.122743
16	-0.0002785	-.00914									.			.										0.123192
17	-0.0005343	-.01754									.			.										0.123199
18	-0.0082079	-.26938									*****			.										0.123222
19	0.0067046	0.22004									.	****		.										0.128608
20	0.0020443	0.06709									.	*		.										0.132080
21	0.00054490	0.01788									.			.										0.132398
22	-0.0038429	-.12612									. ***			.										0.132421
23	0.0014657	0.04810									.	*		.										0.133539
24	-0.0015810	-.05189									. *			.										0.133700

Fig. 3. Autocorrelation function of monthly HFRS incidence. Dotted line, 95% confidence intervals; black star, autocorrelation coefficient.

Partial Autocorrelations

Fig. 4. Partial autocorrelation function of monthly HFRS incidence. Dotted line, 95% confidence intervals; black star, partial autocorrelation coefficient.

Table 1. Parameter estimation for plausible ARIMA models

Parameter	ARIMA (3,1,1) × (2,1,0) ₁₂			ARIMA (1,1,0) × (1,1,1) ₁₂			ARIMA (3,1,0) × (2,1,1) ₁₂			ARIMA (3,1,1) × (2,1,1) ₁₂		
	B	t	P	B	t	P	B	t	P	B	t	P
AR1,1	-0.296	-2.89	0.005	-0.809	-2.72	0.008	-0.291	-2.85	0.005	-0.273	-2.81	0.006
AR2,1	-0.252	-2.52	0.013							-0.308	-3.11	0.002
MA1,1	0.264	2.55	0.013	-0.728	-2.10	0.038	0.265	2.58	0.011	0.342	3.46	0.001
MA2,1				0.261	2.64	0.010	0.290	3.02	0.003	0.242	2.46	0.016

AR1,1, autoregressive parameter; AR2,1, seasonal autoregressive parameter; MA1,1, moving average parameter; MA2,1, seasonal moving average parameter.

Table 2. Goodness of fit statistics for plausible ARIMA models

Model	AIC	SBC
ARIMA (3,1,1) × (2,1,0) ₁₂	-79.98	-71.96
ARIMA (1,1,0) × (1,1,1) ₁₂	-72.94	-64.93
ARIMA (3,1,0) × (2,1,1) ₁₂	-81.11	-73.09
ARIMA (3,1,1) × (2,1,1) ₁₂	-88.26	-77.57

AIC, Akaike information criterion; SBC, Schwarz Bayesian criterion.

DISCUSSION

In this study, we applied multiplicative seasonal ARIMA (p, d, q) × (P, D, Q)_s models to analyze the surveillance data of HFRS in Zibo, China. According to

the results above, the ARIMA (3,1,1) × (2,1,1)₁₂ model is reliable with a high validity, which can be used to predict the next year's HFRS incidence in Zibo. The forecast results suggest that the HFRS incidence in Zibo will experience a slight growth in the next year. A rise in the number of HFRS cases may also result from an increase in the number and size of natural foci (25), as well as climate change, especially the increase of mean temperature (7,26). Therefore, knowledge of HFRS forecasts is necessary to prompt health departments to strengthen surveillance systems and reallocate resources in anticipation of increasing HFRS incidence.

Epidemiological surveillance of communicable diseases is one of the most traditional health-related activities. Time-series analysis of infection incidence is extremely useful in developing hypotheses to explain and anticipate the dynamics of the observed phenomena

Fig. 5. Fitted, predicted, and actual incidence of HFRS in Zibo City from 2004 to 2014. Black star, actual HFRS incidence; solid line, ARIMA (3,1,1) $\times$ (2,1,1)₁₂ model fitted curve; dashed lines, 95% confidence intervals of fitted values.

Table 3. Comparison of predicted HFRS incidence and actual values for Zibo, during January–December 2014

Incidence	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Actual value	0.152	0.369	0.239	0.261	0.456	0.152	0.218	0.130	0.217	0.239	0.374	0.239
Predictive value	0.154	0.305	0.246	0.202	0.354	0.149	0.217	0.111	0.197	0.307	0.346	0.217

and, subsequently, in the establishment of a quality control system and reallocation of resources (27). The ARIMA model has its advantages in time-series analyses. The secular trend, seasonal variation, and autocorrelation can all be easily controlled by difference, autoregression, moving average, and seasonal functions without performing complicated transformations or using extra surrogate variables (6). Once a satisfactory model has been obtained, it can be used to forecast the expected number of cases for a given number of future time intervals (28). Earnest et al. found that an ARIMA model was easier to fit in terms of the parameters and ran more quickly in forecasting notifiable infectious diseases through time-series models (29).

In summary, we further confirmed the consensus that an ARIMA model is a useful tool in monitoring and predicting changing trends in HFRS. Based on the prediction results, the government can invest more health resources during high-risk periods and decrease them during low-risk periods to improve the cost-effectiveness of interventions and scheduling of resources. To control and prevent HFRS, a comprehensive preventive strategy, including public health education and promotion, rodent control, surveillance, and vaccination, should be implemented in Zibo.

Conflict of interest None to declare.

REFERENCES

- Schmaljohn CS, Dalrymple JM. Analysis of Hantaan virus RNA: evidence for a new genus of bunyaviridae. *Virology*. 1983;131:482-91.
- Liu YX, Feng D, Zhang Q, et al. Key differentiating features between scrub typhus and hemorrhagic fever with renal syndrome in northern China. *Am J Trop Med Hyg*. 2007;76:801-5.
- Cui F, Wang T, Wang L, et al. Spatial analysis of hemorrhagic fever with renal syndrome in Zibo City, China, 2009–2012. *PLoS One*. 2013;8:e67490.
- Fang LQ, Zhao WJ, de Vlas SJ, et al. Spatiotemporal dynamics of hemorrhagic fever with renal syndrome, Beijing, People's Republic of China. *Emerg Infect Dis*. 2009;15:2043-5.
- Zhang WY, Fang LQ, Jiang JF, et al. Predicting the risk of hantavirus infection in Beijing, People's Republic of China. *Am J Trop Med Hyg*. 2009;80:678-83.
- Liu X, Jiang B, Bi P, et al. Prevalence of haemorrhagic fever with renal syndrome in mainland China: analysis of National Surveillance Data, 2004–2009. *Epidemiol Infect*. 2012;140:851-7.
- Fang LQ, Wang XJ, Liang S, et al. Spatiotemporal trends and climatic factors of hemorrhagic fever with renal syndrome epidemic in Shandong Province, China. *PLoS Negl Trop Dis*. 2010;4:e789.
- Zhang YZ, Zou Y, Fu ZF, et al. Hantavirus infections in humans and animals, China. *Emerg Infect Dis*. 2010;16:1195-203.
- Song G. Epidemiological progresses of hemorrhagic fever with renal syndrome in China. *Chin Med J (Engl)*. 1999;112:472-7.
- Clement JP. Hantavirus. *Antiviral Res*. 2003;57:121-7.
- Yan L, Fang LQ, Huang HG, et al. Landscape elements and Hantaan virus-related hemorrhagic fever with renal syndrome,

- People's Republic of China. *Emerg Infect Dis.* 2007;13:1301-6.
12. Zuo SQ, Fang LQ, Zhan L, et al. Geo-spatial hotspots of hemorrhagic fever with renal syndrome and genetic characterization of Seoul variants in Beijing, China. *PLoS Negl Trop Dis.* 2011;5:e945.
 13. Simmons JH, Riley LK. Hantaviruses: an overview. *Comp Med.* 2002;52:97-110.
 14. Li Q, Guo NN, Han ZY, et al. Application of an autoregressive integrated moving average model for predicting the incidence of hemorrhagic fever with renal syndrome. *Am J Trop Med Hyg.* 2012;87:364-70.
 15. Liu Q, Liu X, Jiang B, et al. Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. *BMC Infect Dis.* 2011;11:218.
 16. Yu HK, Kim NY, Kim SS, et al. Forecasting the number of human immunodeficiency virus infections in the Korean population using the autoregressive integrated moving average model. *Osong Public Health Res Perspect.* 2013;4:358-62.
 17. Reichert TA, Simonsen L, Sharma A, et al. Influenza and the winter increase in mortality in the United States, 1959-1999. *Am J Epidemiol.* 2004;160:492-502.
 18. Sumi A, Kamo K, Ohtomo N, et al. Time series analysis of incidence data of influenza in Japan. *J Epidemiol.* 2011;21:21-9.
 19. Gaudart J, Toure O, Dessay N, et al. Modelling malaria incidence with environmental dependency in a locality of Sudanese savannah area, Mali. *Malar J.* 2009;8:61.
 20. Luz PM, Mendes BV, Codeco CT, et al. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *Am J Trop Med Hyg.* 2008;79:933-9.
 21. Yi J, Du CT, Wang RH, et al. [Applications of multiple seasonal autoregressive integrated moving average (ARIMA) model on predictive incidence of tuberculosis]. *Zhonghua Yu Fang Yi Xue Za Zhi.* 2007;41:118-21. Chinese.
 22. Zhang X, Zhang T, Young AA, et al. Applications and comparisons of four time series models in epidemiological surveillance data. *PLoS One.* 2014;9:e88075.
 23. Lin H, Lu L, Tian L, et al. Spatial and temporal distribution of falciparum malaria in China. *Malar J.* 2009;8:130.
 24. Sato RC. Disease management with ARIMA model in time series. *Einstein (Sao Paulo).* 2013;11:128-31.
 25. Li Q, Zhao W, Wei Y, et al. Analysis of incidence and related factors of hemorrhagic fever with renal syndrome in Hebei province, China. *PLoS One.* 2014;9:e101348.
 26. Bi P, Wu X, Zhang F, et al. Seasonal rainfall variability, the incidence of hemorrhagic fever with renal syndrome, and prediction of the disease in low-lying areas of China. *Am J Epidemiol.* 1998;148:276-81.
 27. 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. *Am J Epidemiol.* 1994;140:943-55.
 28. Allard R. Use of time-series analysis in infectious disease surveillance. *Bull World Health Organ.* 1998;76:327-33.
 29. Earnest A, Tan SB, Wilder-Smith A, et al. Comparing statistical models to predict dengue fever notifications. *Comput Math Methods Med.* 2012;2012:758674.