BMJ Open

Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study

Mirxat Alim,[1] Guo-Hua Ye,[1] Peng Guan ,[1] De-Sheng Huang,[2] Bao-Sen Zhou,[1] Wei Wu [1]

To cite: Alim M, Ye G-H, Guan P, *et al*. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. *BMJ Open* 2020;**10**:e039676. doi:10.1136/bmjopen-2020-039676

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-039676).

Received 25 April 2020
Revised 03 November 2020
Accepted 17 November 2020

[1]Department of Epidemiology, China Medical University, Shenyang, China
[2]Department of Mathematics, China Medical University, Shenyang, China

Correspondence to
Dr Wei Wu; wuwei@cmu.edu.cn

ABSTRACT

Objectives Human brucellosis is a public health problem endangering health and property in China. Predicting the trend and the seasonality of human brucellosis is of great significance for its prevention. In this study, a comparison between the autoregressive integrated moving average (ARIMA) model and the eXtreme Gradient Boosting (XGBoost) model was conducted to determine which was more suitable for predicting the occurrence of brucellosis in mainland China.

Design Time-series study.

Setting Mainland China.

Methods Data on human brucellosis in mainland China were provided by the National Health and Family Planning Commission of China. The data were divided into a training set and a test set. The training set was composed of the monthly incidence of human brucellosis in mainland China from January 2008 to June 2018, and the test set was composed of the monthly incidence from July 2018 to June 2019. The mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) were used to evaluate the effects of model fitting and prediction.

Results The number of human brucellosis patients in mainland China increased from 30 002 in 2008 to 40 328 in 2018. There was an increasing trend and obvious seasonal distribution in the original time series. For the training set, the MAE, RSME and MAPE of the ARIMA(0,1,1)×$(0,1,1)_{12}$ model were 338.867, 450.223 and 10.323, respectively, and the MAE, RSME and MAPE of the XGBoost model were 189.332, 262.458 and 4.475, respectively. For the test set, the MAE, RSME and MAPE of the ARIMA(0,1,1)×$(0,1,1)_{12}$ model were 529.406, 586.059 and 17.676, respectively, and the MAE, RSME and MAPE of the XGBoost model were 249.307, 280.645 and 7.643, respectively.

Conclusions The performance of the XGBoost model was better than that of the ARIMA model. The XGBoost model is more suitable for prediction cases of human brucellosis in mainland China.

Strengths and limitations of this study

- ► The occurrence of human brucellosis is usually affected by a variety of uncertain factors. Therefore, it often shows the characteristics of non-linearity. The eXtreme Gradient Boosting (XGBoost) model is a machine learning method that can yield high precision prediction results through its strong self-learning ability for these non-linear data. A regular term is included in the objective function of the XGBoost model, which helps to prevent overfitting, and can control the complexity of the model.
- ► In this study, we compared the performance of the XGBoost model and autoregressive integrated moving average model, the results could give us a reference to select suitable model in human brucellosis prediction.
- ► There are many kinds of prediction models, and we need to keep trying different types of models to predict cases of human brucellosis in mainland China and identify the most suitable model for this prediction.
- ► In this study, we mainly considered the influence of time on cases, which makes our model easier to establish and easier to predict. However, we determined that the time factor may not be sufficient, and we should consider climate factors, policy factors, animal brucellosis and so on in future studies.
- ► At present, we have studied the situation of human brucellosis in mainland China as a whole, but China is a large country areawise, and the meteorological, environmental, economic and medical conditions of different provinces (autonomous regions) are different. To obtain a complete grasp of the situation of human brucellosis in mainland China, it is necessary to study each province (autonomous region) in great depth.

BACKGROUND

Brucellosis is a globally recognised zoonotic disease caused by a variety of Brucella species.[1] Approximately 500 000 people are reported to be infected with brucellosis every year. However, the actual incidence is estimated to be 5 million to 12.5 million per year.[2] There are four species of Brucellosis that cause disease in humans[1]: *Brucella melitensis*, found in goats and sheep, *Brucella* *abortus*, typically found in cattle, *Brucella canis* in dogs and *Brucella suis* in swine.[3] The main route of transmission to humans is through contact with infected animals and the consumption of contaminated food.[4–6] Brucellosis has a wide range of clinical manifestations and often lacks specificity, which could lead to misdiagnosis, patients may show fatigue (67%), fever (64%), arthralgia (63%) and sweating (54%). Fever can also be the only manifestation of some patients, and a small number of patients may only manifest as low fever. If not treated in time, long-term illness could lead to severe debilitating and crippling diseases.[7 8] Brucellosis remains a serious public health problem in low/middle-income countries (including China).[9] It is classified as a class B animal epidemic by the World Organisation for Animal Health.[10] In the past decade, the incidence and area of human brucellosis infection in China have been increasing.[7 11–14] According to relevant research, the prevalence of human brucellosis in the summer and autumn is higher than that in spring and winter.[15] Brucellosis epidemics have been reported in 32 provinces (autonomous regions) in mainland China.[12] It is classified as a class B statutory infectious disease in China[3] and is considered a public health problem endangering health and property in China.[15–17] Early warning is important for controlling or reducing the risk of infectious disease outbreaks, and it is therefore vital to ensure the accuracy of Brucellosis predictions.

The most common prediction model for infectious diseases is the autoregressive integrated moving average (ARIMA) model. It is widely used for many infectious diseases,[18] such as influenza viruses,[19] haemorrhagic fever[20] and malaria.[21] However, the occurrence of infectious diseases is affected by a variety of uncertain factors; it often shows the characteristics of non-linearity, and the linearity often does not accord with practical situations.[22] Machine learning methods can address this problem very well. The eXtreme Gradient Boosting (XGBoost) model is a machine learning method that can yield high precision prediction results through its strong self-learning ability. The XGBoost model has achieved excellent performance in many fields of medical research.[23–26] Currently, no researchers have used the XGBoost model to predict the time series data of human brucellosis.

In this study, we used the ARIMA model and XGBoost model to fit and predict the time series of human brucellosis in mainland China. Furthermore, by comparing the fitting effect and prediction accuracy of the two models, we sought to find a model that is more suitable for predicting the trend of human brucellosis in mainland China.

METHODS

Data sources

Cases of human brucellosis were collected from the website of the National Health and Family Planning Commission of China (http://www.nhc.gov.cn). According to the criteria of WHO, all cases of human brucellosis were diagnosed according to clinical symptoms (sweating, fever, myalgia, fatigue, arthralgia, splenomegaly, hepatomegaly, epididymal orchitis, etc) and confirmed by serological testing or isolation of the organism.[27–29] The monthly data of human brucellosis in mainland China from January 2008 to June 2019 were divided into two parts: a training set (from January 2008 to June 2018) and a test set (from July 2018 to June 2019). The training set was used to build the seasonal ARIMA model and XGBoost model. The test set was used to assess the predictive performance of each model. No missing data existed in this study.

Seasonal ARIMA model

The ARIMA model is widely used in time-series modelling of infectious diseases.[30] In general, the components of infectious diseases include long-term trends, periodicity, seasonality and random fluctuations.[31] The ARIMA model can be expressed as ARIMA(p, d, q), where p is the autoregressive order, d is the difference order, and q is the moving average order. The purpose of the difference operation is to stabilise a non-stationary time series and obtain a stationary series. A seasonal ARIMA model mainly extracts information from the seasonal parts that cannot be processed by the standard ARIMA model. According to its complexity, the seasonal model can be divided into an additive model (simple seasonal model) and a product seasonal model. The mathematical relationship of the simple seasonal model is as follows:

$$X_t = S_t + T_t + I_t \tag{1}$$

where S_t, T_t, and I_t represent seasonal information, trend information and random fluctuation information, respectively. For the seasonal ARIMA model, the non-seasonal part is generally identified first, and then the seasonality is identified. After being processed according to seasonal differences or trend differences, the sequence can be transformed into a stationary sequence, and the sequence can be fitted. An augmented Dickey-Fuller (ADF) test can help to confirm whether the sequence is stable. According to the graphs of the autocorrelation function (ACF) and partial ACF (PACF), the possible values of p, q, P and Q can be determined by the Box-Jenkins order determination method. The conditional sum of squares method is used to fit several ARIMA models, and the model with the best fitting effect is selected.[20] The ARIMA model with the lowest corrected Akaike's information criterion (AICc) is regarded as the best fitting effect model.[32] A Ljung-Box test was used to conduct a white noise test for residual sequences.

XGBoost model

The basic idea of improving machine learning models (boosting) is to combine thousands of prediction models with low accuracy into a model with high accuracy. Under reasonable parameter settings, this often requires a certain number of models to be combined to achieve satisfactory prediction accuracy. If the data set is large or complex, the model may need to be iterated thousands of times or more to achieve satisfactory accuracy;

the XGBoost model can better solve this problem. The XGBoost model was first proposed by Chen Tianqi and Carlos Gestrin in 2011 and has been continuously optimised and perfected in follow-up research by many scientists.[33] XGBoost is an efficient and scalable variable of the gradient boosting machine.[34] The objective function of the XGBoost model algorithm is:

$$Obj_m = \sum_{i=1}^{n} l\left(\left(y_i, y_i^{m-1}\right) + f_m\left(x_i\right)\right) + \Omega\left(f_m\right) \qquad (2)$$

where n represents the sample size, m represents the number of iterations, and f_m represents the error in the m iterations. l represents the cost function, which is used to measure the difference between the label and the prediction in the last step, as well as the output of the new tree, and Ω is the regularisation term that punishes the complexity of the new tree.[35] The cost function is extended to a second-order Taylor expansion, and L1 and L2 regularisation are introduced at the same time to avoid overfitting. We used cross-validation with the XGBoost model to obtain a more reliable model.[36] Many techniques for avoiding overfitting in the XGBoost model can help reduce the degree of overfitting and improve the accuracy of regression prediction. When we established the XGBoost model, we used the lag term of the time series as the input item and let the input lag term predict the univariable time series. Considering the seasonality of the time series, we took 12 time-lagged variables as input features, and the maxlag parameter was set to 12. The parameters that needed to be adjusted were nrounds, nrounds_method, nfold, lambda, seas_method, and trend_method. The goal of our study was to find the optimal parameters to minimise the loss deviation.

Evaluation and comparison of models

Evaluation and comparison of the models was mainly based on the model accuracy. Accuracy refers to the degree to which the predicted results are consistent with the actual results. Therefore, the error can be used to evaluate the accuracy of the model prediction. The smaller the error is, the better the fitting effect. Commonly used evaluation indexes are mean absolute error (MAE), rooted mean squared error (RMSE) and mean absolute percentage (MAPE), which are, respectively, defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_i\right| \qquad (3)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - x_i\right)^2} \qquad (4)$$

$$\mathrm{MAPE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\left|\frac{\mathrm{x_i - x_i}}{\mathrm{x_i}}\right| \times 100\% \qquad (5)$$

where x_i is the actual observed value, x_i is the predicted value, and n is the sequence sample size.

Data analysis

The statistical analysis was performed with R V.3.6.1 software. The TSstudio package and stats package were used to process the time series. The drawing portion was carried out with the ggplot2 package. The ARIMA models were built with the forecast package. The fitting of the XGBoost model was completed by the forecastxgb package. In this study, the statistical significance level was set at 0.05.

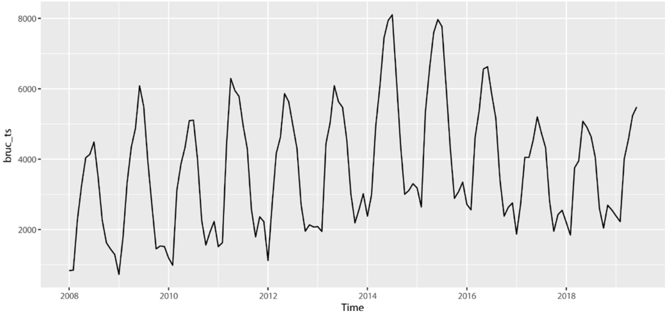

Figure 1 Time series plot for cases of human brucellosis in mainland China from 2008 to 2019.

Patient and public involvement

No patients were involved in this study.

RESULTS

Characteristics of human brucellosis cases in mainland China

From January 2008 to June 2019, there were 512 541 cases of human brucellosis in mainland China. We mapped the time series of human brucellosis cases and searched for trends and seasonality. As shown in figure 1, the time series seems to have seasonality, but it can be seen that the data fluctuate greatly, and the ADF test also verifies that the data are not smooth. Box-Cox transformation was used to make the raw data more stable and demonstrate less variance (figure 2).[37] We then decomposed the data and found that they have a strong seasonal pattern (figure 3). The right side of the diagram contains bars for ease of comparison of the size of each component, the time series of the Box-Cox transform, seasonality, trend and noise components, displayed from top to bottom. The graphs show that there is a large trend change. We performed first-order differencing of the sequence to correct for the trend change and then drew a seasonal chart with the monthly data to better assess changes in the seasonality (figure 4). Human brucellosis in mainland China began to increase rapidly in February, reached its peak in June and July, and began to decline in August until it reached a trough in October, November, December and January. In addition, we drew a diagram of the relationship between

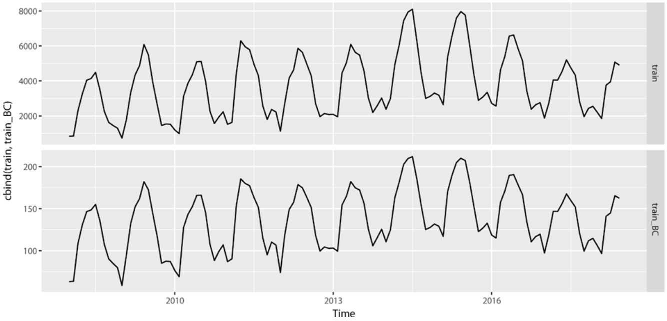

Figure 2 Comparison of the original sequence and Box-Cox transformed sequence.

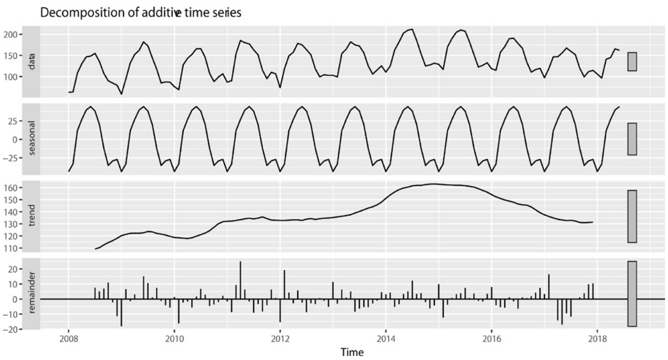

Figure 3 Seasonal decomposition of the Box-Cox transformed human brucellosis cases.

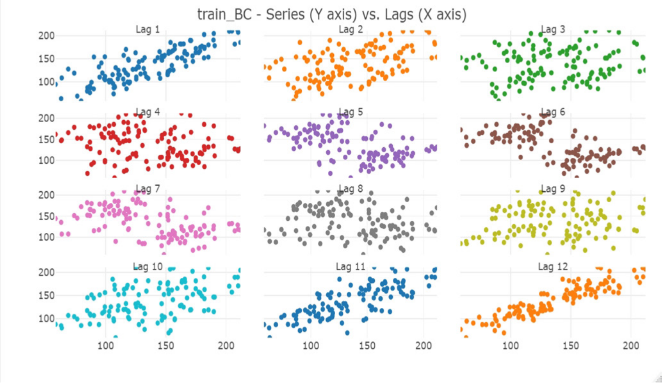

Figure 5 Differencing item correlations of the first 12 lags.

the sequence and the lag sequence (figure 5). To address the instability caused by seasonal factors, we performed a 12-order differencing (seasonal differencing) of the data according to the lag diagram. The above steps were all in preparation for building of the seasonal ARIMA model.

Seasonal ARIMA model

Seasonal differencing and first order differencing stabilised the time series of human brucellosis transformed by Box-Cox (figure 6). The ADF test of the differenced time series showed that it was stationary (ADF test: t=−5.6219, p<0.01). As a result, the two parameters d and D of the seasonal ARIMA model were 1 and 1, respectively.

In the ACF diagram (figure 7, top), there was an obvious peak at lag 1, indicating that the non-seasonal MA may be 1, and an obvious peak at lag 12, indicating that the seasonal MA may be 1. In the PACF diagram (figure 7, bottom), there were obvious spikes at lags 1 and 2, indicating that the non-seasonal AR may be 2, and an obvious peak at lag 12, indicating that the seasonal AR may be 1, so the maximum values of p, q, P, Q are 2, 1, 1 and 1, respectively. We then used the auto.arima function to build the model, listed all the models and selected the model with the lowest AICc value. In the end, the model we obtained was ARIMA(0,1,1)×(0,1,1) $_{12}$, the details of which are shown in table 1. The results of the Ljung-Box test showed that the residual lag of the ARIMA model was white noise (p=0.3981) from 1 to 36 orders. We then drew the residual diagram, the ACF diagram of the residual and the histogram of the residual, which indicated a normal distribution (figure 8). Therefore, the ARIMA(0,1,1)×(0,1,1) $_{12}$ model was significant, and the corresponding fitting diagrams is shown in figure 9.

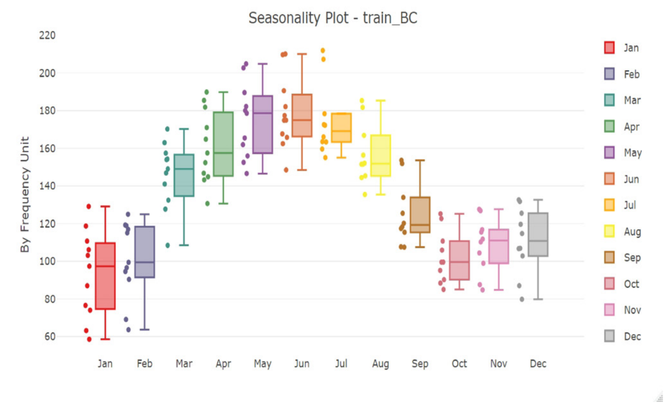

Figure 4 Box chart of monthly cases of brucellosis in mainland China from 2008 to 2019.

XGBoost model

By adjusting the parameters repeatedly, the XGBoost model with the best performance was obtained. The adjusted parameters were: nrounds=10, nrounds_method = ‘cv’, nfold=10, lambda=1, seas_method = ‘dummies’, trend_method = ‘none’. As shown in figure 10, the predicted results of the XGBoost model were in good agreement with the real values.

Comparison of models

Differencing and seasonal differencing were performed when establishing the seasonal ARIMA(0,1,1)×(0,1,1) $_{12}$ model. As a result, we lost the first 13 values in the training set, so only the remaining 125 values were compared. Considering the seasonality of the time series, we took 12 time-lagged variables as the input features for XGBoost; therefore, the remaining 126 values were compared for the XGBoost model. The comparison of the prediction accuracy between the ARIMA model and XGBoost model is shown in table 2. Regardless of the use of the training set or the test set, the MAE value, RSME value and MAPE value of the XGBoost model were much lower than those of the ARIMA model, even though the result for

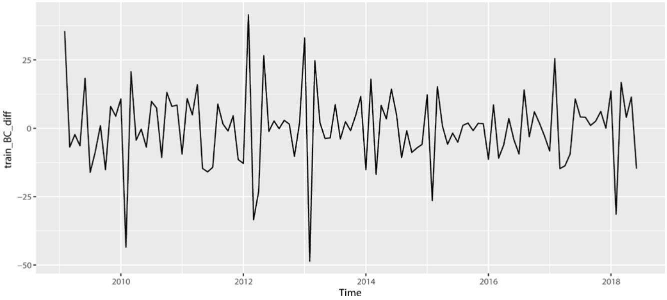

Figure 6 Monthly cases of brucellosis in mainland China after Box-Cox transformation, first-order difference and seasonal difference.

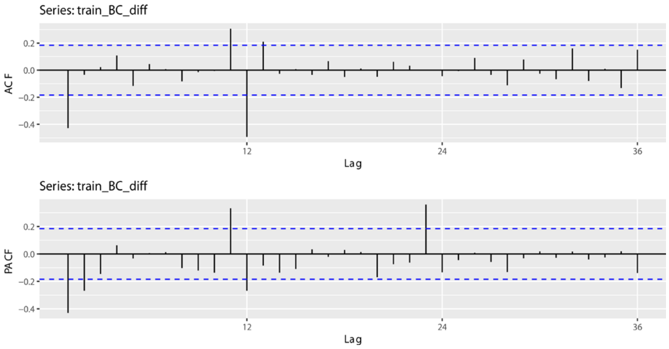

Figure 7 ACF and PACF diagrams for monthly cases of brucellosis in mainland China after Box-Cox transformation, first-order differencing and seasonal differencing. ACF, autocorrelation function; PACF, partial autocorrelation function.

the test set of the XGBoost model is better than that for the training set of the ARIMA model. The MAPE values in the training set and test set of the XGBoost model were 4.475% and 7.643%, respectively, which is excellent according to the literature we consulted.

DISCUSSION

According to our research, although the monthly incidence of brucellosis in mainland China has fluctuated in different months of each year since 2008, the overall incidence continues to increase year by year, showing an obvious upward trend that began to alleviate in 2014; nevertheless, the baseline incidence of the disease is too large to be taken lightly,[3] and increased awareness of continuous prevention and control of the disease is essential.[5] Spring is the popular season, with the annual high incidence of brucellosis occurring soon thereafter in June, July and August, and the lowest incidence occurring in January and December.[30] The incidence of brucellosis can therefore be considered to be seasonal. Through seasonal decomposition, it was easier to determine the seasonality and trend of this disease, which provided a reference for us to analyse, process and stabilise data, establishing a basis for developing a mathematical model. Some researchers found that climatic factors had a certain impact on the spread of brucellosis.[38 39] In this study, we have studied the human brucellosis in mainland China as a whole. It is well known that the land area of China ranks third in the world, and China spans wide latitudes. From north to south, China has cold temperate, middle temperate, warm temperate, subtropical, tropical and other temperature zones. Meteorological factors in mainland China have obvious spatial variability. Therefore, it is not reasonable to use the mean values of meteorological factors to forecast the occurrence of human brucellosis in mainland China. To obtain a complete grasp of the situation of human brucellosis in mainland China, it is necessary to study each province (autonomous region) or city in great depth or use methods such as spatial panel models to solve the problem of spatial variability in the future.

Table 1 Estimated parameters of the seasonal ARIMA(0,1,1)×(0,1,1) $_{12}$ model

Coefficients		ma1	sma1
		−0.4411	−0.8149
	SE	0.0970	0.1111
CIs of coefficients		2.5%	97.5%
	ma1	−0.6313151	−0.2508861
	sma1	−1.0327691	−0.5970962
AICc		842.49	

AIC, Akaike's information criterion; ARIMA, autoregressive integrated moving average.

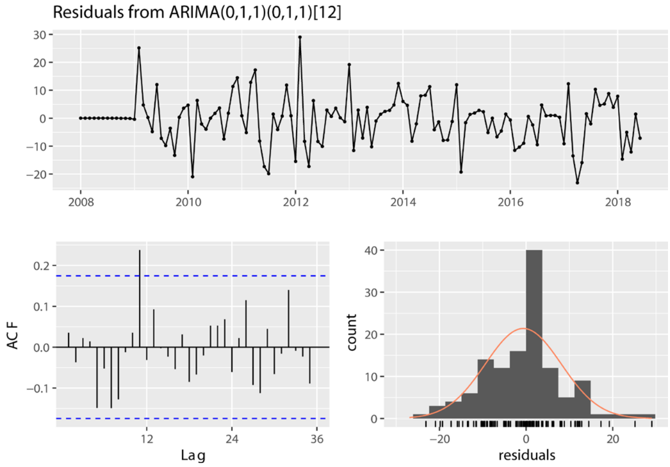

Figure 8 A time plot of the residuals, the corresponding ACF diagram, and a histogram for the ARIMA(0,1,1)×(0,1,1) $_{12}$ model. ACF, autocorrelation function; ARIMA, autoregressive integrated moving average.

The ARIMA model was developed based on a linear regression model, which can reveal the dynamic laws of the data itself and predict future data values. The ARIMA model integrates the original time series variables of

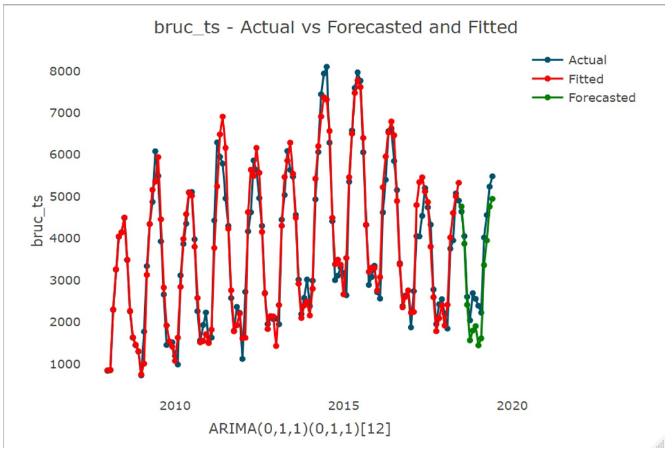

Figure 9 Fit and prediction of the seasonal ARIMA(0,1,1)×(0,1,1) $_{12}$ model. ARIMA, autoregressive integrated moving average.

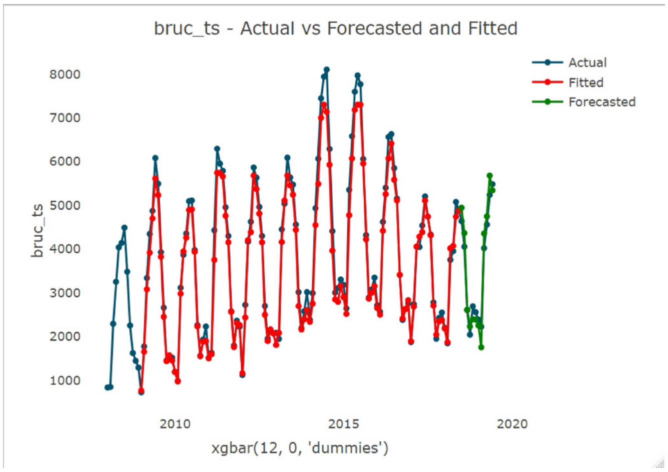

Figure 10 Fit and prediction of the XGBoost model. XGBoost, eXtreme Gradient Boosting.

trend factors, periodic factors and random errors. This model combines the advantages of autoregressive and average moving models, is characterised by not being constrained by data types and strong adaptability and shows good predictability over a short time.[20] Because of its relatively mature time series prediction method and its wide usage, the ARIMA model was compared with the XGBoost model in this study. After Box-Cox transformation and differencing adjustment, the ARIMA model could also analyse non-stationary time series, which also demonstrates the ability of the model to predict diseases. In general, the more differences used, the more data loss occurs. Because differencing was used in this study, 13 months of data were lost before fitting the ARIMA model. It should be noted that when using auto.arima, the parameter lambda needs to be set to auto so that the original data can be Box-Cox transformed automatically when building the model and to ensure the stability of the data. Finally, according to the AICc value, we determined that the model with the best fitting effect was the seasonal $\text{ARIMA}(0,1,1)\times(0,1,1)_{12}$ model.[40,41] Although it was the optimal ARIMA model, the performance of the training set and the test set were quite different. The MAPE value of the training set was 10.323%, while that of the test set was 17.676%, a difference of 7.353. The reason for the large gap is that the ARIMA model is more suitable for fitting linear trends, and the performance for non-linear trends is not ideal. In most cases, it is necessary to predict non-linear data, whether in infectious diseases or other fields. At this time, we required a model that could better fit nonlinear data.[42] Because the fitting of the ARIMA model for non-linear data is non-ideal, it is more suitable as a short-term prediction model. When the prediction accuracy of the ARIMA model is not ideal, it is necessary to constantly collect data, obtain as long a time series as possible, and repeatedly modify and optimise the model. At the same time, a model that can better fit the non-linear data is considered to yield the expected effect.

A lot of machine learning methods such as artificial neural network,[27,38,43] support vector machine[44] and random forest[45] were widely used in brucellosis prediction and achieved good forecasting performance. The XGBoost model, a relatively new model, was first proposed by Chen Tianqi and Carlos Gestrin in 2011. Currently, no researchers have used the XGBoost model to predict the time series data of human brucellosis. We intended to explore the possibility of using this method in human brucellosis prediction. Therefore, we selected the most widely used ARIMA time series model as baseline in this study. From the results, there is no doubt that the performance of the non-linear data in the XGBoost model is very excellent. It has a larger late pruning penalty than the traditional gradient boosting decision tree, which makes the learnt model less prone to overfitting. At the same time, the XGBoost model can be used for cross-verification and can find important feature vectors automatically. In this study of human brucellosis in mainland China, the fitting and prediction effect of the XGBoost model are much better than those of the ARIMA model. It was suggested that the XGBoost model could feasibly predict human brucellosis. In addition, the evaluation index values (MAE, RMSE and MAPE) of the XGBoost model for the training samples and the prediction samples were relatively small. It also shows that the XGBoost model has high precision fitting and prediction effects. Such a highly accurate result depends on the automatic learning ability of the model. Compared with the complexity of the conditions that the ARIMA model needs to meet, the modelling process of the XGBoost model is very simple.

In real infectious disease surveillance, the time series data of infectious disease always show non-stationary characteristics. This is a serious problem for traditional ARIMA model. Data transformation and differencing adjustment are needed to make the time series to be stationary. The ARIMA model is not suitable for processing the type of data that cannot be transformed into a stationary process,

Table 2 Comparison of the fitting and prediction accuracy of the two models

	Training set			Test set		
Model	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
$\text{ARIMA}(0,1,1)\times(0,1,1)_{12}$	338.867	450.223	10.323	529.406	586.059	17.676
XGBoost	189.332	262.458	4.475	249.307	280.645	7.643

ARIMA, autoregressive integrated moving average; MAE, mean absolute error; MAPE, mean absolute percentage error; RMSE, root mean square error; XGBoost, eXtreme Gradient Boosting.

while machine learning methods are not affected by the stationarity of time series. In addition, machine learning methods can achieve better performance for non-linear time series. Therefore, with the emergence of new algorithms, machine learning will have a wider range of applications than traditional ARIMA models in real infectious disease surveillance. The prediction method should be selected according to the characteristics of the time series of the infectious disease under investigation. To better reflect the future trend of the disease and obtain higher prediction accuracy, new data should be constantly collected to re-establish the model. Brucellosis remains a serious infectious human and animal disease that continues to endanger people's health and life, as well as production and the economy, in mainland China today.[4] The purpose of this study was to provide a reference for medical staff in the prevention and control of brucellosis and to make their own contributions to the prevention and control of other infectious diseases in mainland China.

CONCLUSIONS

In this study, we established a seasonal ARIMA model and an XGBoost model for human brucellosis in mainland China and used them to make short-term predictions. The prediction accuracy of the XGBoost model was much better than that of the ARIMA model. The XGBoost model has many advantages in model prediction, such as the lack of a need to preprocess the data, a fast operation speed, complete feature extraction, a good fitting effect and high prediction accuracy.

Contributors MA designed and drafted the manuscript. G-HY and PG participated in the data collection. MA and WW participated in the data analysis. D-SH and B-SZ made critical revision to the manuscript. The final manuscript has been read and approved by all the authors.

Funding This study was supported by the National Natural Science Foundation of China (grant no. 81202254 and 71974199) and Health and Medical Big Data Research Project of China Medical University (grant no. HMB201903105).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data in this manuscript are available.

ORCID iDs
Peng Guan http://orcid.org/0000-0003-0190-7301
Wei Wu http://orcid.org/0000-0001-5535-1682

REFERENCES

1 Chen Y, Ke Y, Wang Y, *et al*. Changes of predominant species/biovars and sequence types of Brucella isolates, inner Mongolia, China. *BMC Infect Dis* 2013;13:514.
2 Hull NC, Schumaker BA. Comparisons of brucellosis between human and veterinary medicine. *Infect Ecol Epidemiol* 2018;8:1500846.
3 Li M-T, Sun G-Q, Zhang W-Y, *et al*. Model-Based evaluation of strategies to control brucellosis in China. *Int J Environ Res Public Health* 2017;14. doi:10.3390/ijerph14030295. [Epub ahead of print: 12 Mar 2017].
4 Ran X, Chen X, Wang M, *et al*. Brucellosis seroprevalence in ovine and caprine flocks in China during 2000-2018: a systematic review and meta-analysis. *BMC Vet Res* 2018;14:393.
5 Zhang N, Zhou H, Huang D-S, *et al*. Brucellosis awareness and knowledge in communities worldwide: a systematic review and meta-analysis of 79 observational studies. *PLoS Negl Trop Dis* 2019;13:e0007366.
6 Dieckhaus KD, Kyebambe PS. Human brucellosis in rural Uganda: clinical manifestations, diagnosis, and comorbidities at Kabale regional referral Hospital, Kabale, Uganda. *Open Forum Infect Dis* 2017;4:ofx237.
7 Zheng R, Xie S, Lu X, *et al*. A systematic review and meta-analysis of epidemiology and clinical manifestations of human brucellosis in China. *Biomed Res Int* 2018;2018:5712920.
8 Shi Y, Gao H, Pappas G, *et al*. Clinical features of 2041 human brucellosis cases in China. *PLoS One* 2018;13:e0205500.
9 Tan Z, Huang Y, Liu G, *et al*. A familial cluster of human brucellosis attributable to contact with imported infected goats in Shuyang, Jiangsu Province, China, 2013. *Am J Trop Med Hyg* 2015;93:757–60.
10 Zhang J, Sun G-Q, Sun X-D, *et al*. Prediction and control of brucellosis transmission of dairy cattle in Zhejiang Province, China. *PLoS One* 2014;9:e108592.
11 Chen Q, Lai S, Yin W, *et al*. Epidemic characteristics, high-risk townships and space-time clusters of human brucellosis in Shanxi Province of China, 2005-2014. *BMC Infect Dis* 2016;16:760.
12 Zhao Z-J, Li Q, Ma L, *et al*. The early diagnostic value of serum neopterin and cartilage oligomeric matrix protein for osteoarticular changes among brucellosis patients at an early period. *J Orthop Surg Res* 2018;13:222.
13 Jia P, Joyner A. Human brucellosis occurrences in inner Mongolia, China: a spatio-temporal distribution and ecological niche modeling approach. *BMC Infect Dis* 2015;15:36.
14 Liang C, Wei W, Liang X, *et al*. Spinal brucellosis in Hulunbuir, China, 2011-2016. *Infect Drug Resist* 2019;12:1565–71.
15 Lou P, Wang L, Zhang X, *et al*. Modelling seasonal brucellosis epidemics in Bayingolin mongol autonomous Prefecture of Xinjiang, China, 2010-2014. *Biomed Res Int* 2016;2016:5103718.
16 Zhang J, Yin F, Zhang T, *et al*. Spatial analysis on human brucellosis incidence in mainland China: 2004-2010. *BMJ Open* 2014;4:e004470.
17 Wang T, Wang X, Tie P, *et al*. Spatio-Temporal cluster and distribution of human brucellosis in Shanxi Province of China between 2011 and 2016. *Sci Rep* 2018;8:16977.
18 Zhang X, Hou F, Qiao Z, *et al*. Temporal and long-term trend analysis of class C notifiable diseases in China from 2009 to 2014. *BMJ Open* 2016;6:e011038.
19 He Z, Tao H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: a nine-year retrospective study. *Int J Infect Dis* 2018;74:61–70.
20 Wang T, Liu J, Zhou Y, *et al*. Prevalence of hemorrhagic fever with renal syndrome in Yiyuan County, China, 2005-2014. *BMC Infect Dis* 2016;16:69.
21 Anwar MY, Lewnard JA, Parikh S, *et al*. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. *Malar J* 2016;15:566.
22 Li Z, Wang Z, Song H, *et al*. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. *Infect Drug Resist* 2019;12:1011–20.
23 Dinh A, Miertschin S, Young A, *et al*. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak* 2019;19:211.
24 Liu L, Yu Y, Fei Z, *et al*. An interpretable boosting model to predict side effects of analgesics for osteoarthritis. *BMC Syst Biol* 2018;12:105.
25 Liu Z, Zhou T, Han X, *et al*. Mathematical models of amino acid panel for assisting diagnosis of children acute leukemia. *J Transl Med* 2019;17:38.
26 Zou LS, Erdos MR, Taylor DL, *et al*. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics* 2018;19:390.

27 Wu W, An S-Y, Guan P, *et al*. Time series analysis of human brucellosis in mainland China by using Elman and Jordan recurrent neural networks. *BMC Infect Dis* 2019;19:414.

28 Ryu S, Soares Magalhães RJ, Chun BC. The impact of expanded brucellosis surveillance in beef cattle on human brucellosis in Korea: an interrupted time-series analysis. *BMC Infect Dis* 2019;19:201.

29 Guan P, Wu W, Huang D. Trends of reported human brucellosis cases in mainland China from 2007 to 2017: an exponential smoothing time series analysis. *Environ Health Prev Med* 2018;23:23.

30 Wang L, Liang C, Wu W, *et al*. Epidemic situation of brucellosis in Jinzhou city of China and prediction using the ARIMA model. *Can J Infect Dis Med Microbiol* 2019;2019:1–9.

31 Wang Y, Xu C, Zhang S, *et al*. Temporal trends analysis of human brucellosis incidence in mainland China from 2004 to 2018. *Sci Rep* 2018;8:15901.

32 Liu Q, Li Z, Ji Y, *et al*. Forecasting the seasonality and trend of pulmonary tuberculosis in Jiangsu Province of China using advanced statistical time-series analyses. *Infect Drug Resist* 2019;12:2311–22.

33 Li W, Yin Y, Quan X, *et al*. Gene expression value prediction based on XGBoost algorithm. *Front Genet* 2019;10:1077.

34 Babajide Mustapha I, Saeed F. Bioactive molecule prediction using extreme gradient boosting. *Molecules* 2016;21. doi:10.3390/molecules21080983. [Epub ahead of print: 28 Jul 2016].

35 Nishio M, Nishizawa M, Sugiyama O, *et al*. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoS One* 2018;13:e0195875.

36 Zeng X, An J, Lin R, *et al*. Prediction of complications after paediatric cardiac surgery. *Eur J Cardiothorac Surg* 2020;57:350–8.

37 Curran-Everett D. Explorations in statistics: the log transformation. *Adv Physiol Educ* 2018;42:343–7.

38 Bagheri H, Tapak L, Karami M, *et al*. Forecasting the monthly incidence rate of brucellosis in West of Iran using time series and data mining from 2010 to 2019. *PLoS One* 2020;15:e0232910.

39 Liu K, Yang Z, Liang W, *et al*. Effect of climatic factors on the seasonal fluctuation of human brucellosis in Yulin, Northern China. *BMC Public Health* 2020;20:506.

40 Nsoesie EO, Mekaru SR, Ramakrishnan N, *et al*. Modeling to predict cases of hantavirus pulmonary syndrome in Chile. *PLoS Negl Trop Dis* 2014;8:e2779.

41 Zeng Q, Li D, Huang G, *et al*. Time series analysis of temporal trends in the pertussis incidence in mainland China from 2005 to 2016. *Sci Rep* 2016;6:32367.

42 Wu W, Guo J, An S, *et al*. Comparison of two hybrid models for forecasting the incidence of hemorrhagic fever with renal syndrome in Jiangsu Province, China. *PLoS One* 2015;10:e0135492.

43 Tapak L, Shirmohammadi-Khorram N, Hamidi O, *et al*. Predicting the frequency of human brucellosis using climatic indices by three data mining techniques of radial basis function, multilayer perceptron and nearest neighbor: a comparative study. *Iran J Epidemiol* 2018;14:153–65.

44 Bagheri H, Tapak L, Karami M, *et al*. Epidemiological features of human brucellosis in Iran (2011-2018) and prediction of brucellosis with data-mining models. *J Res Health Sci* 2019;19:e00462.

45 Shirmohammadi-Khorram N, Tapak L, Hamidi O, *et al*. A comparison of three data mining time series models in prediction of monthly brucellosis surveillance data. *Zoonoses Public Health* 2019;66:759–72.