PLOS ONE

RESEARCH ARTICLE

The research of ARIMA, GM(1,1), and LSTM models for prediction of TB cases in China

Daren Zhao[1☯], Huiwu Zhang[1☯], Qing Cao[2], Zhiyi Wang[3], Sizhang He[4], Minghua Zhou[5], Ruihua Zhang[6]*

1 Department of Medical Administration, Sichuan Provincial Orthopedics Hospital, Chengdu, Sichuan, P.R. China, **2** Department of Medical Administration, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, Sichuan, P.R. China, **3** Department of Medical Administration, Sichuan Cancer Hospital & Institute, Chengdu, Sichuan, P.R. China, **4** Department of Information and Statistics, The Affiliated Hospital of Southwest Medical University, Luzhou, Sichuan, P.R. China, **5** Department of Medical Administration, Luzhou People's Hospital, Luzhou, Sichuan, P.R. China, **6** School of Management, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, P.R. China

☯ These authors contributed equally to this work.
* cdzhangrh@126.com

Citation: Zhao D, Zhang H, Cao Q, Wang Z, He S, Zhou M, et al. (2022) The research of ARIMA, GM (1,1), and LSTM models for prediction of TB cases in China. PLoS ONE 17(2): e0262734. https://doi.org/10.1371/journal.pone.0262734

Editor: Esteban Tlelo-Cuautle, Instituto Nacional de Astrofisica Optica y Electronica, MEXICO

Received: July 30, 2021

Accepted: January 4, 2022

Published: February 23, 2022

Data Availability Statement: All data TB cases were taken from National Health Commission of the People's Republic of China (http://www.nhc.gov.cn). Anyone meeting the requirements can gain access to them. The data were relatively uninvolved in detailed patient personal information. The authors confirm they did not have any special access privileges that other would not have.

Funding: We state that the study and the paper were financially supported by the Hospital Management Institute, the National Health Commission of the People's Republic of China

Abstract

Background and objective

Tuberculosis (Tuberculosis, TB) is a public health problem in China, which not only endangers the population's health but also affects economic and social development. It requires an accurate prediction analysis to help to make policymakers with early warning and provide effective precautionary measures. In this study, ARIMA, GM(1,1), and LSTM models were constructed and compared, respectively. The results showed that the LSTM was the optimal model, which can be achieved satisfactory performance for TB cases predictions in mainland China.

Methods

The data of tuberculosis cases in mainland China were extracted from the National Health Commission of the People's Republic of China website. According to the TB data characteristics and the sample requirements, we created the ARIMA, GM(1,1), and LSTM models, which can make predictions for the prevalence trend of TB. The mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) were applied to evaluate the effects of model fitting predicting accuracy.

Results

There were 3,021,995 tuberculosis cases in mainland China from January 2018 to December 2020. And the overall TB cases in mainland China take on a downtrend trend. We established ARIMA, GM(1,1), and LSTM models, respectively. The optimal ARIMA model is the ARIMA (0,1,0) × (0,1,0)12. The equation for GM(1,1) model was $X(k+1) = -10057053.55e^{(-0.01k)} + 10153178.55$ the Mean square deviation ratio C value was 0.49, and the Small probability of error P was 0.94. LSTM model consists of an input layer, a hidden layer and an output layer,

[grant no. YLZLXZ-2021-005]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no conflicts of interest.

the parameters of epochs, learning rating are 60, 0.01, respectively. The MAE, RMSE, and MAPE values of LSTM model were smaller than that of GM(1,1) and ARIMA models.

Conclusions

Our findings showed that the LSTM model was the optimal model, which has a higher accuracy performance than that of ARIMA and GM (1,1) models. Its prediction results can act as a predictive tool for TB prevention measures in mainland China.

Introduction

Tuberculosis, an infectious disease caused by Mycobacterium tuberculosis, is still a major global public health problem [1]. According to the GLOBAL TUBERCULOSIS REPORT 2020 released by the World Health Organization, approximately 10 million people are reported to be infected with TB, among the HIV-negative people, with an estimated 1.2 million people died of TB, and among the HIV-positive people, 208,000 people died of TB [2]. It is classified as a class B infectious disease, and the morbidity and mortality of TB have always been among the top two in the Class A and B infectious diseases in mainland China currently.

If not treated in time, long-term illness could lead to laying a huge economic burden for patients and exerting a stronger influence on social development. There are 30 high TB burden countries account for almost 90% of those who fall sick with TB each year [2]. China has the second largest burden of TB in the world with huge health and economic losses [3].

According to the data from the China Health Statistics Yearbook, the TB mortality rate of China shows an ascendant trend although the TB incidence has been declining for the past few years. Therefore, the prevention and control of TB is still the focus of current research. Meanwhile, a few TB patients have experienced interruptions to their treatment schedules because of the COVID-19 pandemic threatens. Consequently, it is important to realize the early warning in infectious disease surveillance. It is of great significance to create an accurate prediction model for the morbidity of TB and then predict the future epidemic situation, which can provide a basis for scientific guidance on its control and prevention [4].

There are several mathematical methods for infectious disease prediction at present. Through a literature review, the infectious disease prediction model is mainly classified into two categories [5, 6]: the traditional mathematical forecasting model and the machine-learning based forecasting models. The traditional mathematical forecasting models consist of Autoregressive Integrated Moving Average model (ARIMA) [7], Exponential Smoothing model [8], Regression forecast model [9], Grey Markov forecasting model [10], and GM (1,1) model [11] and so on, and the machine-learning based forecasting models, such as BP artificial neural network model [12], Multivariate Adaptive Regression Splines (MARS) [13], Support Vector Machine (SVM) [14], and Long Short-Term Memory (LSTM) [15], etc.

However, different models are suitable for different data characteristics [8]. So there is no doubt that according to the data characteristics and the sample requirements to construct the optimized prediction model, which was a precondition for obtaining accurate prediction performance. Recent studies show that the traditional mathematical forecasting models have performed well in infectious disease prediction. Zheng et al. [4] revealed that ARIMA models are an important tool for infectious disease prediction, and prediction results can help set public health planning by the government in Xinjiang, China. Ilie et al. [16] demonstrated that ARIMA time-series models have been successfully applied in the overall prevalence of

COVID-19 in Romania. Wang et al. [17] used ARIMA and GM(1,1) models to predict hepatitis B in China, and the ARIMA model achieved better results than GM(1,1) model. Guo et al. [18] used GM(1,1) and novel SMGM(1,1) models to predict dysentery and gonorrhea in China. Despite the traditional mathematical forecasting models have better ability in infectious disease prediction, these models can not extract nonlinear relationships in time series [15]. However, the machine-learning based forecasting models have good performance in handling nonlinear relationships in time series without any limitations [15].

Yet to date, no researchers have applied the ARIMA, GM(1,1), and LSTM models to predict TB cases in mainland China. Under such a background, the present researcher tried to use the ARIMA, GM(1,1), and LSTM models for prediction of TB cases in mainland China. The TB cases data were obtained from the National Health Commission of the People's Republic of China website, based on the data characteristics and the sample requirements, we created the ARIMA, GM(1,1), and LSTM models, respectively. In order to achieve more accurate prediction results, three models were compared by MAE, RMSE and MAPE, thus, we find the optimized prediction model to predict the epidemic trend of TB cases from January to December 2021 in mainland China.

Methods

Data source

All data TB cases were taken from the National Health Commission of the People's Republic of China (http://www.nhc.gov.cn/), and all TB cases had been laboratory confirmed. In China, TB is classified as a class B infectious disease, and the hospital physicians must report every confirmed TB case information to the local health authority within 24 hours by national Network report of infectious disease [4], the final report to the National Center for Disease Control and Prevention. The data of TB cases only include in mainland China.

We collected monthly TB cases data from January 2018 to December 2020 with 36 samples in our study. The data can also be divided into a training set, a validation set, and a test set. The training set and the validation set were the same, that was made up of the TB cases from January 2018 to December 2020, which was used to build and compare the ARIMA, GM(1,1) and LSTM models respectively, and the test set constituted the TB cases from January to December in 2021, which was used to assess the performance of prediction results in the future trend of TB cases in mainland China.

Model descriptions

ARIMA model. ARIMA model was proposed by Box and Jenkins in the early 1970s, so it is also called Box-Jenkins model and Box-Jenkins method [19]. It is the most commonly used prediction techniques in the evaluation and monitoring epidemiological surveillance [20–26]. The ARIMA model consists of auto regressive (AR) model, moving average (MA) model, seasonal auto regressive integrated moving average (SARIMA) model and etc [17, 27]. If the data of research showed evidence of seasonal tendency, the seasonal auto regressive integrated moving average (SARIMA) model should be used [28].

In general, the ARIMA model can be explained as ARIMA (p,d,q),×(P,D,Q), where p represents the order of auto-regression, d represents the degree of trend difference, q represents the order of moving average, and the P represents the seasonal auto regression lag, D represents the degree of seasonal difference, Q represents the seasonal moving average, s represents the length of the cyclical pattern [29].

There are several steps to construct the ARIMA model, which mainly contains four steps: sequence stationary, model identification, estimation and diagnosis, and model prediction and evaluation.

The first step is sequence stationary. If the sequence shows non-stationary time series, the stationary time series should be transformed by differenceing processes [30]. The original sequence diagram could help suggesting whether the sequence is stationary or not. In order to achieve the stationary time series, differences and Log transformation can be processed by statistical software.

The second step is model identification. The preliminary judgment and estimation of ARIMA model parameters are found to be dependent on the auto-correlation function (ACF) and partial auto-correlation function (PACF) graphs. And then, the candidate ARIMA model parameters can be determined by both skills and experiences, observing the auto-correlation function (ACF) and partial auto-correlation function (PACF) graphs.

The third step is estimation and diagnosis. The candidate ARIMA models evaluated by the diagnostic checking of residuals with Ljung-Box (Q) test [31], which requires the residual error must be random (significant level $p>0.05$). If the Q -statistics is less than 0.8, the tentative model is inadequate [4]. In other words, the model should be fitted again. Furthermore, the optimal model was determined by the lowest the Bayesian information criterion of Schwarz (BIC) values and its residual was white noise (significant level $p>0.05$).

The forth step is model prediction and evaluation. The optimal model was applied to predict TB cases from January 2018 to December 2020. And the prediction power were evaluated by comparing the predicted values with actual values [17].

GM(1,1) model. GM(1,1) model was proposed by Deng J. L, and it is described a novel mathematical prediction system that can be used to process data which some information is known, unknown and/or incomplete [32]. The steps of the GM(1,1) model are [33–37]:

Assuming that the original sequence shown as:

$$X^{(0)} = \left[x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), \ldots x^{(0)}(n)\right] \tag{1}$$

Performing an accumulation to generate a new sequence:

$$X^{(1)} = \left[x^{(1)}(1), x^{(1)}(2), x^{(1)}(3), \ldots x^{(1)}(n)\right] \tag{2}$$

Solving adjacent neighbor means by the following formula:

$$Z^{(1)}(t) = \frac{1}{2}\left[x^{(1)}(t) + x^{(1)}(t+1)\right], t = 1, 2, 3 \ldots n \tag{3}$$

Establishing first-order linear differential equations by the following formula:

$$\frac{dx^{(1)}(t)}{dt} + ax^{(1)}(t) = b \tag{4}$$

Assuming that $\hat{a} = (a, b)^T$ then identification a can be calculated by the following formula:

$$\hat{a} = (a, b)^T = (B^T B)^{-1} B^T Y \tag{5}$$

$$Y = \left[x^{(0)}(2), x^{(0)}(3), \ldots x^{(0)}(n)\right]^T \tag{6}$$

$$B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix} \tag{7}$$

The GM(1,1) model can be expressed as:

$$\hat{x}^{(1)}(k+1) = \frac{b}{a} + \left[x^{(0)}(1) - \frac{b}{a}\right] e^{-ak} \qquad k = 1, 2, 3, \ldots n,$$

where *a* is the development coefficient, *b* is the gray effect.

Model test evaluated by the posterior error test:

$$C = \frac{S_e}{S_x} \tag{8}$$

$$P = P\{\varepsilon(k) - \bar{\varepsilon} < 0.6745\ S_x\} \tag{9}$$

And S_e represents the standard deviation of the residual sequence, S_x represents the standard deviation of the original sequence and P represents the Small probability of error. The accuracy of the model is shown in Table 1, and the model applicable scope shown as Table 2.

LSTM model. LSTM, called Long Short-Term Memory, has been improved by the Recurrent Neural Network (RNN) [38]. It was first proposed by Hochreiter and Schmidhuber in 1997 and has received the widespread application in many fields [39, 40]. The LSTM unit includes an Input Gate, a Forget Gate, and an Output Gate (Fig 1) [41]. It selectively allows information to pass through Gate structure, so as to update or retain historical information.

Table 1. The accuracy of gray GM (1,1) model.

Prediction accuracy grade	Mean square deviation ratio C	Small probability of error P
Level 1(Excellent)	≤0.35	≥0.95
Level 2 (Qualified)	≤0.50	≥0.80
Level 3 (Barely qualified)	≤0.65	≥0.70
Level 4 (Unqualified)	>0.65	<0.70

https://doi.org/10.1371/journal.pone.0262734.t001

Table 2. The applicable scope of the GM (1,1) model.

Developing Coefficient a	Prediction Length
-a≤0.3	Medium- and long-term prediction
0.3<-a<0.5	Short-term prediction
0.5<-a<1.0	Modified model to predict
1.0<-a	Not suitable for grey prediction model

https://doi.org/10.1371/journal.pone.0262734.t002

The LSTM model can be expressed as [42, 43]:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{10}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{11}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{12}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{13}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{14}$$

Here, f_t, i_t, and o_t represent the Forget Gate, Input Gate and Output Gate, respectively. Besides, σ represents sigmoid function, C_t represents the cell state update value at time t, $\tilde{C}_t$ represents the candidate state value of the input cell at time t. W_f, W_i, W_C represent the weight of Forget Gate, Input Gate and Output Gate.

Performance measures

The fitting and prediction accuracy can be employed by the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) [22]. The smaller the error is, the better the fitting effect [19]. In this paper, MAE, RMSE, and MAPE are applicable to evaluating the fitting and prediction accuracy of the three models, respectively, which are

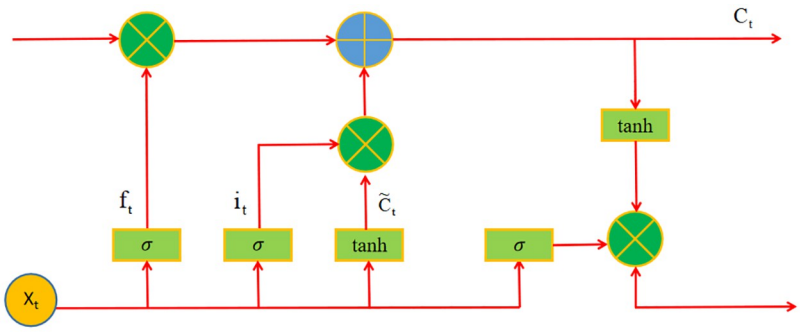

Fig 1. Plot of the LSTM unit structure.

https://doi.org/10.1371/journal.pone.0262734.g001

expressed as:

$$MAF = \frac{\sum_{t=1}^{n} |X_t - \hat{X}_t|}{n} \quad (15)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (X_t - \hat{X}_t)^2}{n}} \quad (16)$$

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{X_t - \hat{X}_t}{X_t} \right| \times 100}{n} \quad (17)$$

Where the $\hat{X}_t$ is the predicted value, X_t is the actual value and n is the sequence sample size [19].

Data processing and analysis

The data of this study was recorded with EXCEL 2010, and the ARIMA was performed using SPSS23.0 software, and Matlab 2020b was adopted to construct the GM(1,1) and LSTM models. Significant level is 0.05.

Results

Trends of TB cases in mainland China

There were 3,021,995 TB cases in mainland China from January 2018 to December 2020. As is shown in Fig 2, the overall TB cases data in mainland China take on a downtrend trend. In a year, the monthly TB cases data presents lowest level in January and February, whereas March and April, in contrast, has the highest level. Simultaneously, it may be observed that the monthly TB cases data has roughly seasonal fluctuations.

ARIMA model

In our study, the SARIMA was selected due to that the data of TB cases has roughly seasonal fluctuations. The basic optical requirement of the ARIMA model is stationary data. As is shown in Fig 3, the original sequence shows non-stationary time series, so the trend difference and the seasonal difference were processed in order to eliminate data instabilities. After a first-order difference and a first-order seasonal difference (Fig 4), the time series were stationary, thus the parameters *d* and *D* were 1, respectively. To confirm the estimation of ARIMA model parameters of *q*, *p*, *Q*, and *P*, the auto-correlation function (ACF) and partial auto-correlation function (PACF) graphs were performed.

Both ACF and PACF graphs showed that the sequence presented high seasonal characteristics with a circle of 12, indicating the parameter of *s* was 12 (s = 12). The ACF graph showed that after a first-order difference (Fig 5), the values of auto-correlation coefficients did not exceed ± 2 times the estimated standard deviation range, and the peak was at lag 0, which was initially identified as the parameters of *q* was 0, p was 0 or 1. Similarly, the PACF graph was represented after a first-order seasonal difference (Fig 6), the partial correlation coefficient did not exceed the range of ± 2 times the estimated standard deviation, and it is initially identified that *P* was 0, Q was 0, or 1.

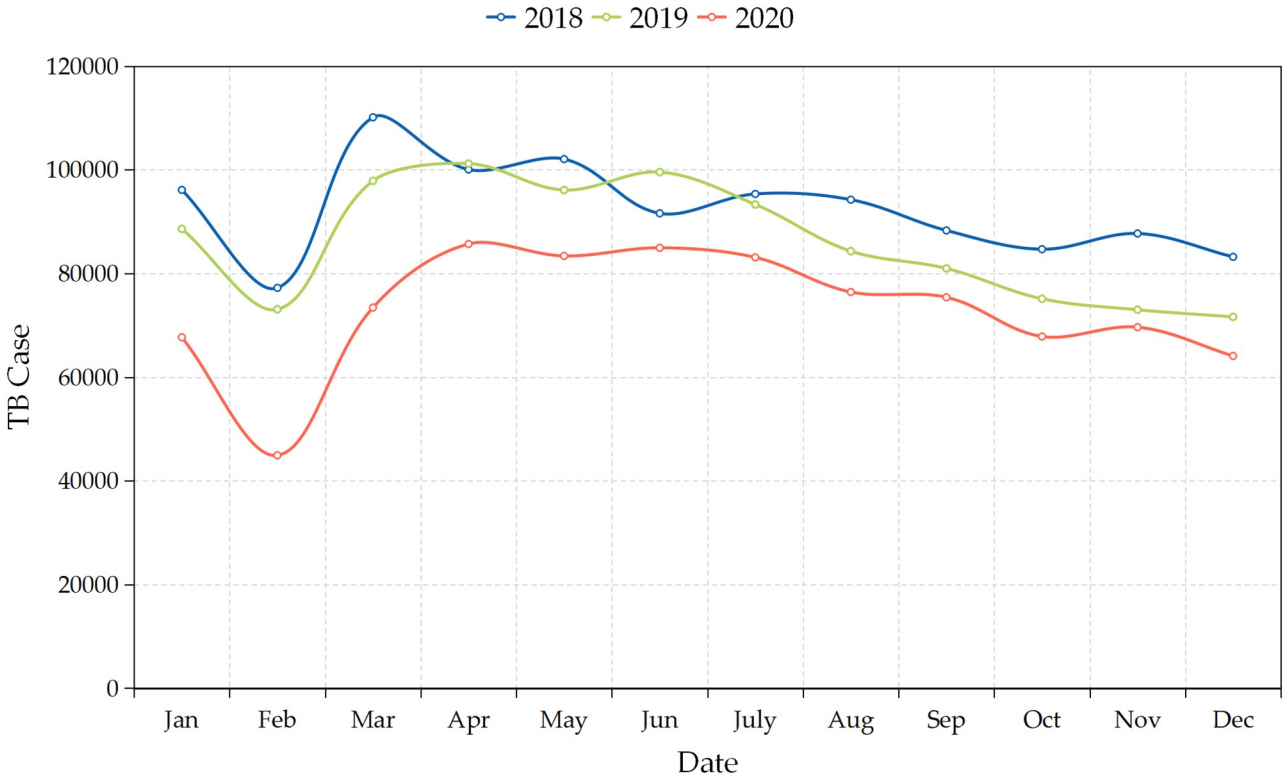

Fig 2. Time series monthly reported cases of TB in mainland China from January 2018 to December 2020.

https://doi.org/10.1371/journal.pone.0262734.g002

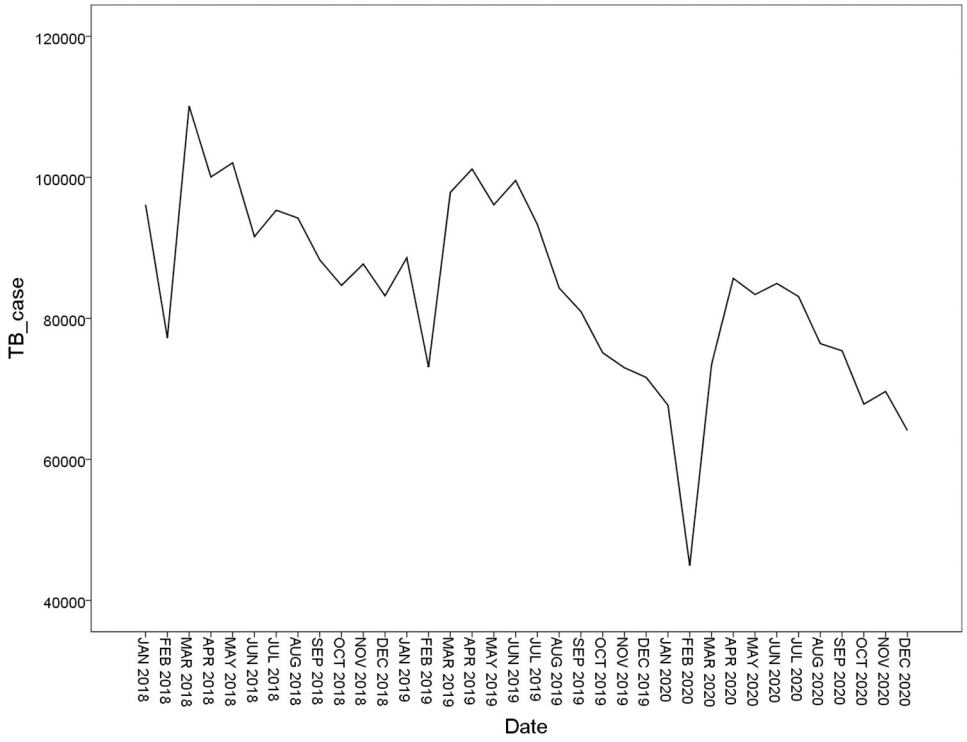

Fig 3. Plot of time series of original sequence.

https://doi.org/10.1371/journal.pone.0262734.g003

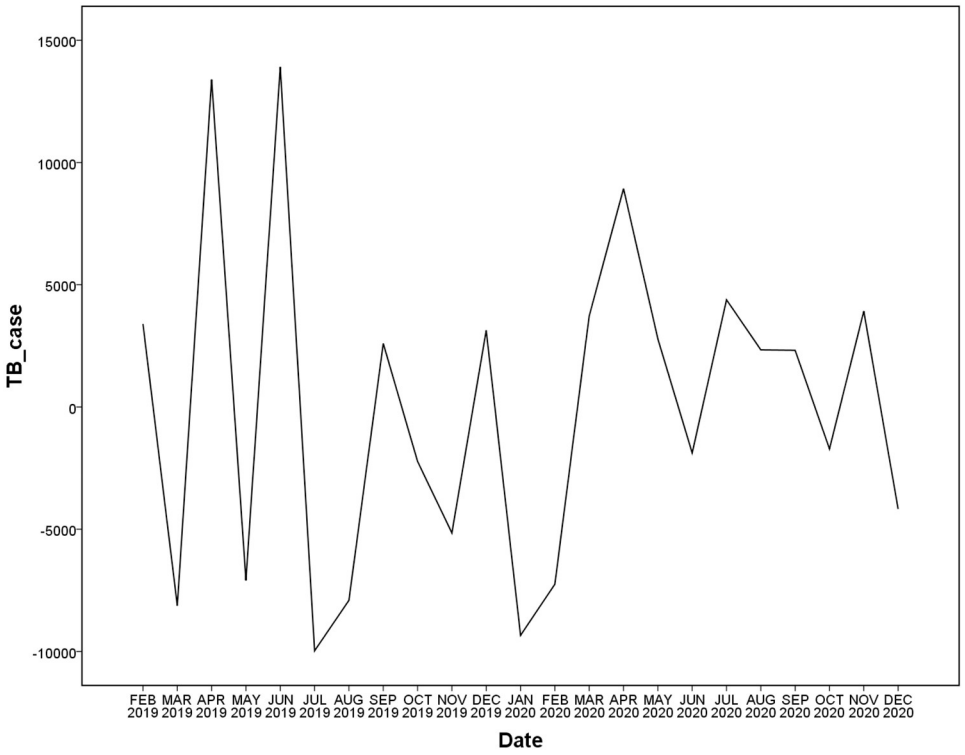

Fig 4. Plot of time series of after a first-order difference and a first-order seasonal difference.

https://doi.org/10.1371/journal.pone.0262734.g004

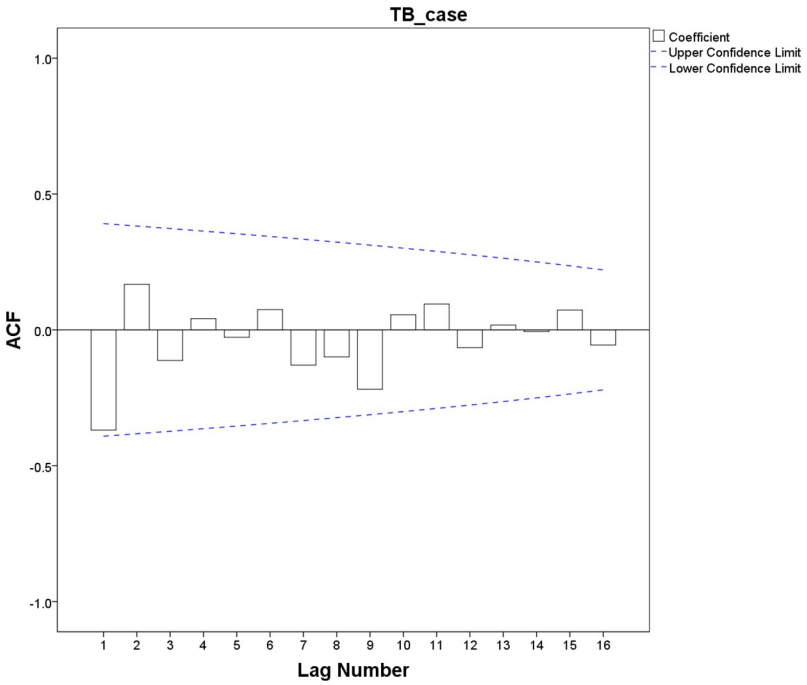

Fig 5. Plot of ACF after a differenced TB cases time series.

https://doi.org/10.1371/journal.pone.0262734.g005

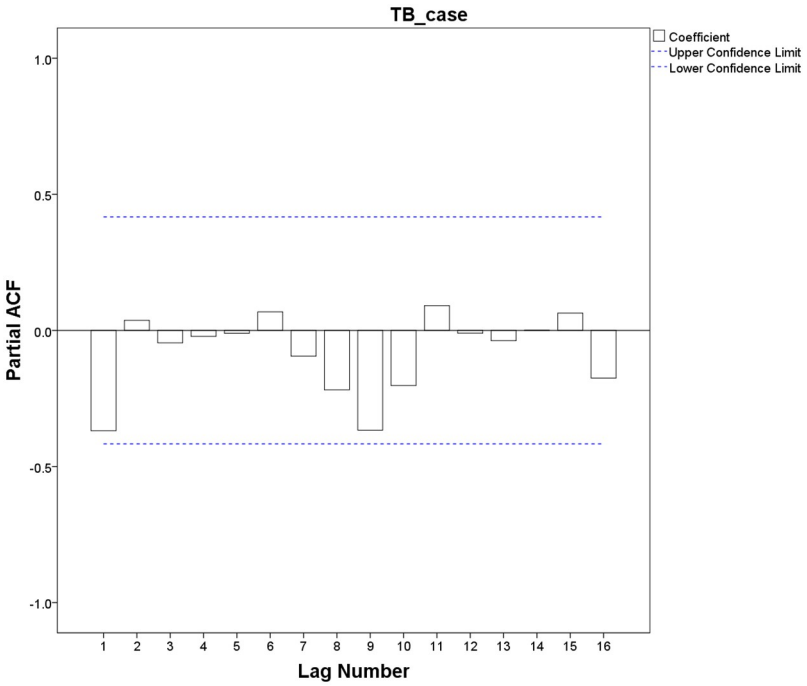

Fig 6. Plot of PACF after a differenced TB cases time series.

https://doi.org/10.1371/journal.pone.0262734.g006

Further, the candidate ARIMA models were based on both the preliminary parameters values and the modeling estimation and diagnosis, which were selected as ARIMA $(0,1,0) \times (0,1,1)_{12}$, ARIMA $(0,1,0) \times (0,1,0)_{12}$, ARIMA $(1,1,0) \times (0,1,0)12$ and ARIMA $(1,1,0) \times (0,1,1)_{12}$. The results are shown in Table 3.

By the diagnostic checking of residuals with the Ljung-Box (Q) test, we find that the Q-statistics of four models were all bigger than 0.8 and the *p*-values were all bigger than 0.05 (Table 3), indicating that these models are all adequate and the white noise sequence. To obtain the optimal model, the lowest BIC values were observed in four models. As a result, the ARIMA $(0,1,0) \times (0,1,0)_{12}$ was the optimal model, whose BIC value was the lowest (17.79) and its residual was white noise sequence, the Q-statistics was 12.62 and *p*-value was 0.81 (Table 3).

GM(1,1) model

The results show that the parameters of model a was 0.01 and u were 99740, the equation for the GM(1,1) model as follows: $X(k+1) = -10057053.55e^{(-0.01k)}+10153178.55$, the Mean square deviation ratio C value was 0.49, and the Small probability of error P was 0.94.

Table 3. Parameter estimation of candidate ARIMA models.

ARIMA Modes	R^2	BIC	Ljung-Box (Q) test	
			Q-Statistics	*p*-value
ARIMA $(0,1,0) \times (0,1,1)_{12}$	0.73	17.97	12.48	0.77
ARIMA $(0,1,0) \times (0,1,0)_{12}$	0.73	17.79	12.62	0.81
ARIMA $(1,1,0) \times (0,1,0)_{12}$	0.77	17.82	9.60	0.92
ARIMA $(1,1,0) \times (0,1,1)_{12}$	0.77	18.00	9.39	0.89

https://doi.org/10.1371/journal.pone.0262734.t003

According to the results in (Table 1), it can ensure that the GM(1,1) model we established is Level 2 (Qualified), and it can be adopted in extrapolated prediction. Besides, the GM(1,1) model parameter of a was 0.01, and $-a \leq 0.3$, which demonstrated the Medium- and long-term prediction can be predicted.

LSTM model

In this section, we build an LSTM model, which consists of three parts, that is, an input layer, a hidden layer, and an output layer. In the LSTM modeling process, the TB cases data from January 2018 to December 2020 were divide into two parts, 2/3 percent of which is the training set and the rest (1/3) is the test set. We set epochs, learning rating with 60, 0.01, respectively. The loss function uses mean_squared_error, and the optimizer uses Adam. The Look_back is set 12 to find the optimal situation of the current network structure. The results are shown in Table 4.

Model comparison

The only 23 values were compared with ARIMA, GM(1,1), and LSTM models, on account of its differencing and seasonal differencing of the ARIMA model, which caused the first 13 values to be lost in the validation set.

In this section, the various statistical tools, such as MAE, RMSE, and MAPE, which used to evaluate the fitting and predicting accuracy. Both the actual values and the predicted values were put them into the formula (15), (16), (17), and then results of MAE, RMSE and MAPE were by calculation as shown in Table 5.

Table 4. The actual values and the prediction results of the three models.

Date	actual values	ARIMA	GM	LSTM
19-Feb	73096	69696	87380	73620
19-Mar	97866	105996	86526	97580
19-Apr	101191	87796	85680	100244
19-May	96106	103200	84842	95929
19-Jun	99555	85646	84013	96138
19-Jul	93318	103290	83192	91464
19-Aug	84304	92212	82379	85603
19-Sep	80973	78374	81573	81252
19-Oct	75123	77351	80776	75207
19-Nov	73000	78152	79986	72188
19-Dec	71631	68496	79204	71270
20-Jan	67682	77023	78430	68097
20-Feb	44933	52181	77663	45219
20-Mar	73427	69703	76904	74241
20-Apr	85684	76752	76152	86788
20-May	83385	80599	75408	83530
20-Jun	84952	86834	74671	86706
20-Jul	83101	78715	73941	82831
20-Aug	76423	74087	73218	72042
20-Sep	75409	73092	72502	66545
20-Oct	67843	69559	71794	60984
20-Nov	69640	65720	71092	53988
20-Dec	64097	68271	70397	53131

https://doi.org/10.1371/journal.pone.0262734.t004

Table 5. The values of MAE and MAPE and RMSE of the three models.

Models	MAE	MAPE	RMSE
ARIMA $(0,1,0) \times (0,1,0)_{12}$	5638.43	0.0706	22599.46
GM (1,1)	8805.39	0.1210	37452.98
LSTM	2676.08	0.0368	16344.92

https://doi.org/10.1371/journal.pone.0262734.t005

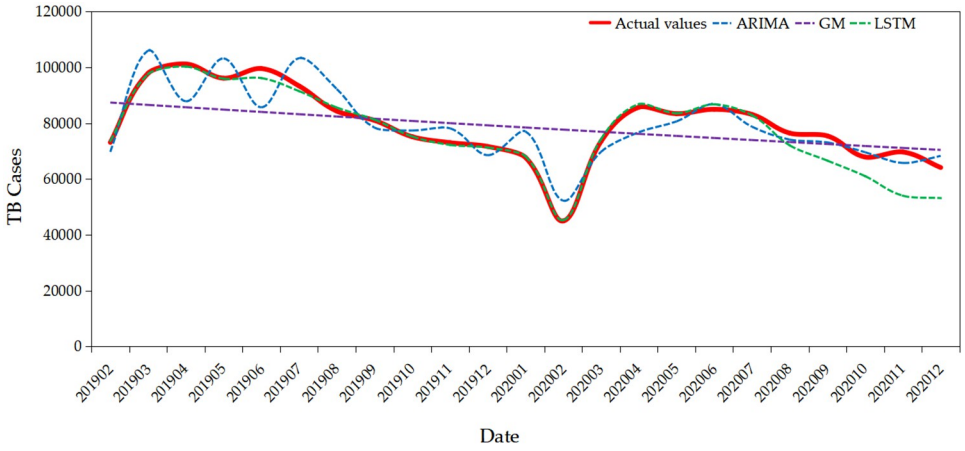

Fig 7. Plot of comparison of predicted values and actual values of three models.

https://doi.org/10.1371/journal.pone.0262734.g007

We could see that the MAE, RMSE, and MAPE of LSTM model values were smaller than that of GM(1,1) and ARIMA $(0,1,0) \times (0,1,0)_{12}$ model. Additionally, the Fig 7 showed that the predicted value obtained by LSTM can better fit the actual value change trend. Therefore, the LSTM model was the most optimal model, which was more suitable to predict the future trend of TB cases of the mainland in China.

Prediction

The LSTM model was applied to predict the trend of TB cases in mainland China from January to December in 2021. Predictions results is shown in Table 6.

Discussion

Although TB can be treated effectively, the morbidity and mortality of TB have always been maintained a relative growth trend [4, 44]. It is still a major global public health problem, which poses a serious danger to human health and the development of the society and

Table 6. Prediction of TB cases in mainland China from January to December in 2021.

Date	Predicted values	Date	Predicted values
21-Jan	60148	21-Jul	75565
21-Feb	37398	21-Aug	68887
21-Mar	65892	21-Sep	67873
21-Apr	78149	21-Oct	60306
21-May	75850	21-Nov	62103
21-Jun	77416	21-Dec	56560

https://doi.org/10.1371/journal.pone.0262734.t006

economy in the world. However, based on a large-scale population and with the rapid development of society and economy in China, the prevention and control of TB is still an important public issue and with difficult challenges. So it still regarded as an important topic in the field of public health in mainland China.

Scientific prediction and analysis of the morbidity and mortality of TB can provide suggestions in public health planning and set the proper policy to adopt effective interventions for this infectious disease [45]. In the field of Disease Surveillance, prediction of the morbidity and mortality infection disease is one of the most important works with higher priorities in public health in China [32]. Upon the above research background, the purpose of this study is to explore the optimized prediction model to predict the epidemic trend of TB cases and propose its prevention and control in mainland China.

Studies have shown that each type of infectious disease prediction model has distinct advantages and disadvantages [46]. It is therefore critical that according to the data characteristics and the sample requirements to create the most suitable prediction model for research objectives. ARIMA is applicable to time series with characteristics of seasonality and periodicity [47]. The GM(1,1) has no special requirements for research data and is also applied in small sample sizes with uncertain time series predictions [17]. While the LSTM model is more suitable for time series with missing values and where there may be a lag of unknown duration [42]. Yet, according to characteristics of the data the sample requirements to construct prediction models is also a precondition for obtaining accurate research results. In our study, TB cases data with 36 samples from January 2018 to December 2020 in mainland China presents characteristics of the seasonality and periodicity, which are fully meet the requirements of ARIMA, GM(1,1), and LSTM models. Therefore, in terms of prediction technique we selected is correct and reasonable.

Further analysis, with the advantage of its structured modeling basis and acceptable forecasting performance, the ARIMA model is the most commonly used technique in time series prediction analysis [4]. Besides, the ARIMA model also can be taken various influencing factors of infection as well as the complicated interactions into consideration in the modeling process [47]. In this study, the seasonal auto regressive integrated moving average (SARIMA) model was applied to predict TB cases, due to the fact that the research data showed evidence of seasonal tendency.

GM(1,1) model was proposed by Deng J. L in 1982, which is widely used in the field of population [48], economic [49], environment [50], power industry [51], medicine [11], etc. This theory is that a system can be defined with a color that represents the amount of clear information about that system [11]. If the information is entirely unknown, it is called a black system. If on the contrary, it is called a white system. Moreover, if some information is known, unknown and/or incomplete, it is called a grey system. In reality, every system can be viewed as a grey system because that its information is known, unknown and/or incomplete. Provided that there are four sets of data, the model can be fitted. So in this paper, the prediction of TB cases can be seen as a grey system, and the monthly TB cases data from January 2018 to December 2020 with 36 samples can be regarded as known information that seeks its changing laws to predict the future state of TB cases.

LSTM model has been widely used to predict infectious disease incidence, such as HIV [15], hepatitis E [52], Dengue fever [53], hand, foot and mouth disease (HFMD) [54], COVID-19 [39], and so on. It is focused on resolving the issues of vanishing gradient [52]. LSTM model has a strong nonlinear mapping capability, which is applicable to process complex issues with long-term dependencies [55]. It can train satisfactory network models by learning from a given sample of data and is suitable for cases where sequence laws are very long time lags of unknown size [15]. Therefore, the LSTM model has a strong capability to address various prediction issues, especially infectious disease prediction.

Model comparison results showed that the MAE, RMSE, and MAPE of LSTM model values were smaller than that of GM(1,1) and ARIMA models. We can deduce that the fitting and predicting performance of the LSTM model was better than that of the ARIMA and GM(1,1) models. The reasons are as follows: first, GM(1,1) and ARIMA models have their disadvantages, for example, they can not extract nonlinear relationships in time series. Second, unlike the other machine-learning based forecasting models, the LSTM model overcomes the shortcomings in vanishing gradient during the training process. Third, the LSTM model is more tolerant to the data and is less vulnerable to model misspecification issues than other time series prediction models. Meanwhile, compared with the GM(1,1) and ARIMA models, it is a more effective deep learning model for continuous data. Thus, the LSTM model has good performance in prediction for infectious disease.

As a result, the LSTM model was applied to predict the TB cases in mainland China from January to December in 2021. The study showed that the predicted values of TB cases present fluctuating and the overall trend is decreasing. However, in practice, the morbidity of TB is the comprehensive effect of manifold causes. The COVID-19 pandemic could affect TB control programs worldwide due to the impairing TB diagnosis and treatment [56, 57], China is no exception. Moreover, although the prevention and control of the COVID-19 pandemic is more effective in China, it also can present great challenges for the medical service system and medical care capacity. Therefore, it must be noted that we still can not ignore the occurrence of an outbreak of TB cases in mainland China in the future.

To the best of our knowledge, this is the first study to construct the ARIMA, GM(1,1), and LSTM models for the prediction of TB cases in mainland China at present. The results showed that the accuracy performance of the LSTM model was higher than that of ARIMA and GM (1,1) models, and the LSTM model can give a credible predictions for TB control and prevention.

However, there are several limitations in this study. One of the limitations is that although the LSTM model has a strong nonlinear mapping capability, the social, cultural, economic, and other factors can not be taken into account in the modeling process [47]. Another limitation is that the prediction results of the LSTM model can act as predictive tools for TB control and prevention, but it should not be used as an exclusive policy-making reference frame due to unexpected emergencies, for example, the COVID-19 pandemic. Since COVID-19 occurred at the end of 2019 in China, missing reports on TB cases are possible to occur in 2020. It results in some inaccuracies in predictions of TB cases to some extent. Therefore, in further work, we will take the influencing factors of TB incidence into the prediction model and continually update data of TB cases so that we can achieve a more suitable and accurate model for its control and prediction.

Conclusion

In this study, we collected the TB cases from January 2018 to December 2020 in mainland China. Based on the data characteristics and the sample requirements, the ARIMA, GM(1,1), and LSTM models were constructed and compared. The fitting and predicting performance of the LSTM model was better than that of the ARIMA and GM (1,1) models. The results of this study can provide policy advice by the government in mainland China.

Supporting information

S1 File. The monthly TB cases data from January 2018 to December 2020.
(DOC)

S1 Data.
(XLS)

Author Contributions

Conceptualization: Daren Zhao, Huiwu Zhang, Ruihua Zhang.

Data curation: Daren Zhao, Huiwu Zhang.

Formal analysis: Daren Zhao, Qing Cao, Zhiyi Wang.

Writing – original draft: Daren Zhao, Huiwu Zhang, Qing Cao, Zhiyi Wang, Sizhang He, Minghua Zhou.

Writing – review & editing: Daren Zhao, Huiwu Zhang, Ruihua Zhang.

References

1. Li Z, Wang Z, Song H, Liu Q, He B, Shi P, et al. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infect Drug Resist. 2019 Apr 29; 12:1011–1020. https://doi.org/10.2147/IDR.S190418 PMID: 31118707
2. WHO. Global tuberculosis report; 2020. [cited 30.07.2021] http://www.who.int/tb/publications/global_report/en/.
3. Zhang G, Huang S, Duan Q, Shu W, Hou Y, Zhu S, et al. Application of a hybrid model for predicting the incidence of tuberculosis in Hubei, China. PLoS ONE. 2013; 8(11):e80969. https://doi.org/10.1371/journal.pone.0080969 PMID: 24223232
4. Zheng YL, Zhang LP, Zhang XL, Wang K, Zheng YJ. Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China. PLoS ONE. 2015; 10(3):e0116832. https://doi.org/10.1371/journal.pone.0116832 PMID: 25760345
5. Zhai M, Li W, Tie P, Wang X, Xie T, Ren H, et al. Research on the predictive effect of a combined model of ARIMA and neural networks on human brucellosis in Shanxi Province, China: a time series predictive analysis. BMC Infect Dis. 2021; 21(1):280. https://doi.org/10.1186/s12879-021-05973-4 PMID: 33740904
6. Alzahrani SI, Aljamaan IA, Al-Fakih EA. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health. 2020; 13(7):914–919. https://doi.org/10.1016/j.jiph.2020.06.001 PMID: 32546438
7. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief. 2020;26; 29:105340. https://doi.org/10.1016/j.dib.2020.105340 PMID: 32181302
8. Zhang YQ, Li XX, Li WB, Jiang JG, Zhang GL, Zhuang Y, et al. Analysis and predication of tuberculosis registration rates in Henan Province, China: an exponential smoothing model study. Infect Dis Poverty. 2020; 9(1):123. https://doi.org/10.1186/s40249-020-00742-y PMID: 32867846
9. Rath S, Tripathy A, Tripathy AR. Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes Metab Syndr. 2020; 14(5):1467–1474. https://doi.org/10.1016/j.dsx.2020.07.045 PMID: 32771920
10. Wang Y, Shen Z, Jiang Y. Analyzing maternal mortality rate in rural China by Grey-Markov model. Medicine (Baltimore). 2019; 98(6):e14384. https://doi.org/10.1097/MD.0000000000014384 PMID: 30732175
11. Yang X, Zou J, Kong D, Jiang G. The analysis of GM (1, 1) grey model to predict the incidence trend of typhoid and paratyphoid fevers in Wuhan City, China. Medicine (Baltimore). 2018; 97(34):e11787. https://doi.org/10.1097/MD.0000000000011787 PMID: 30142765
12. Li Z, Li Y. A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. BMC Med Inform Decis Mak. 2020; 20(1):143. https://doi.org/10.1186/s12911-020-01157-3 PMID: 32616052
13. Lu R, Duan T, Wang M, Liu H, Feng S, Gong X, et al. The application of multivariate adaptive regression splines in exploring the influencing factors and predicting the prevalence of HbA1c improvement. Ann Palliat Med. 2021; 10(2):1296–1303. https://doi.org/10.21037/apm-19-406 PMID: 33040556

14. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics. 2018; 15(1):41–51. https://doi.org/10.21873/cgp.20063 PMID: 29275361

15. Wang G, Wei W, Jiang J, Ning C, Chen H, Huang J, et al. Application of a long short-term memory neural network: a burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China. Epidemiol Infect. 2019; 147:e194. https://doi.org/10.1017/S095026881900075X PMID: 31364559

16. Ilie OD, Ciobica A, Doroftei B. Testing the Accuracy of the ARIMA Models in Forecasting the Spreading of COVID-19 and the Associated Mortality Rate. Medicina (Kaunas). 2020; 56(11):566. https://doi.org/10.3390/medicina56110566 PMID: 33121072

17. Wang YW, Shen ZZ, Jiang Y. Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China. PLoS ONE. 2018; 13(9):e0201987. https://doi.org/10.1371/journal.pone.0201987 PMID: 30180159

18. Guo X, Liu S, Wu L, Tang L. Application of a novel grey self-memory coupling model to forecast the incidence rates of two notifiable diseases in China: dysentery and gonorrhea. PLoS ONE. 2014; 9(12): e115664. https://doi.org/10.1371/journal.pone.0115664 PMID: 25546054

19. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ. 2020; 729:138817. https://doi.org/10.1016/j.scitotenv.2020.138817 PMID: 32360907

20. Alim M, Ye GH, Guan P, Huang DS, Zhou BS, Wu W. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. BMJ Open. 2020; 10(12):e039676. https://doi.org/10.1136/bmjopen-2020-039676 PMID: 33293308

21. Wu W, An SY, Guan P, Huang DS, Zhou BS. Time series analysis of human brucellosis in mainland China by using Elman and Jordan recurrent neural networks. BMC Infect Dis. 2019; 19(1):414. https://doi.org/10.1186/s12879-019-4028-x PMID: 31088391

22. Zheng Y, Zhang L, Wang C, Wang K, Guo G, Zhang X, et al. Predictive analysis of the number of human brucellosis cases in Xinjiang, China. Sci Rep. 2021; 11(1):11513. https://doi.org/10.1038/s41598-021-91176-5 PMID: 34075198

23. Singh S, Parmar KS, Kumar J, Makkhan SJS. Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19. Chaos Solitons Fractals. 2020; 135:109866. https://doi.org/10.1016/j.chaos.2020.109866 PMID: 32395038

24. Singh S, Parmar KS, Makkhan SJS, Kaur J, Peshoria S, Kumar J. Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries. Chaos Solitons Fractals. 2020; 139:110086. https://doi.org/10.1016/j.chaos.2020.110086 PMID: 32834622

25. Dansana D, Kumar R, Das Adhikari J, Mohapatra M, Sharma R, Priyadarshini I, et al. Global Forecasting Confirmed and Fatal Cases of COVID-19 Outbreak Using Autoregressive Integrated Moving Average Model. Front Public Health. 2020; 8:580327. https://doi.org/10.3389/fpubh.2020.580327 PMID: 33194982

26. Zhang X, Yu Y, Xiong F, Luo L. Prediction of Daily Blood Sampling Room Visits Based on ARIMA and SES Model. Comput Math Methods Med. 2020; 2020:1720134. https://doi.org/10.1155/2020/1720134 PMID: 32963583

27. Ramezanian M, Haghdoost AA, Mehrolhassani MH, Abolhallaje M, Dehnavieh R, Najafi B, et al. Forecasting health expenditures in Iran using the ARIMA model (2016–2020). Med J Islam Repub Iran. 2019; 33:25. https://doi.org/10.34171/mjiri.33.25 PMID: 31380315

28. Liu L, Luan RS, Yin F, Zhu XP, Lü Q. Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model. Epidemiol Infect. 2016; 144(1):144–51. https://doi.org/10.1017/S0950268815001144 PMID: 26027606

29. Wei W, Jiang J, Liang H, Gao L, Liang B, Huang J, et al. Application of a Combined Model with Autoregressive Integrated Moving Average (ARIMA) and Generalized Regression Neural Network (GRNN) in Forecasting Hepatitis Incidence in Heng County, China. PLoS ONE. 2016; 11(6):e0156768. https://doi.org/10.1371/journal.pone.0156768 PMID: 27258555

30. ArunKumar KE, Kalaga DV, Sai Kumar CM, Chilkoor G, Kawaji M, Brenza TM. Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). Appl Soft Comput. 2021; 103:107161. https://doi.org/10.1016/j.asoc.2021.107161 PMID: 33584158

31. Esmaeilzadeh N, Shakeri M, Esmaeilzadeh M, Rahmanian V. ARIMA models to forecasting the SARS-COV-2 in the Islamic Republic of Iran. Asian Pac J Trop Med 2020; 13(11): 521–524. https://doi.org/10.4103/1995-7645.291407

32. Wang Y, Wei F, Sun C, Li Q. The Research of Improved Grey GM (1, 1) Model to Predict the Postprandial Glucose in Type 2 Diabetes. Biomed Res Int. 2016; 2016:6837052. https://doi.org/10.1155/2016/6837052 PMID: 27314034

33. Zhang P, Ma X, She K. A novel power-driven fractional accumulated grey model and its application in forecasting wind energy consumption of China. PLoS ONE. 2019; 14(12):e0225362. https://doi.org/10.1371/journal.pone.0225362 PMID: 31805165

34. Gao J, Li J, Wang M. Time series analysis of cumulative incidences of typhoid and paratyphoid fevers in China using both Grey and SARIMA models. PLoS ONE. 2020; 15(10):e0241217. https://doi.org/10.1371/journal.pone.0241217 PMID: 33112899

35. Zhang L, Zheng Y, Wang K, Zhang X, Zheng Y. An optimized Nash nonlinear grey Bernoulli model based on particle swarm optimization and its application in prediction for the incidence of Hepatitis B in Xinjiang, China. Comput Biol Med. 2014; 49:67–73. https://doi.org/10.1016/j.compbiomed.2014.02.008 PMID: 24747730

36. Wu H, Zeng B, Zhou M. Forecasting the Water Demand in Chongqing, China Using a Grey Prediction Model and Recommendations for the Sustainable Development of Urban Water Consumption. Int J Environ Res Public Health. 2017 Nov 15; 14(11):1386. https://doi.org/10.3390/ijerph14111386 PMID: 29140266

37. Hu YC. A genetic-algorithm-based remnant grey prediction model for energy demand forecasting. PLoS ONE. 2017; 12(10):e0185478. https://doi.org/10.1371/journal.pone.0185478 PMID: 28981548

38. Kaya K, Gündüz Öğüdücü Ş. Deep Flexible Sequential (DFS) Model for Air Pollution Forecasting. Sci Rep. 2020; 10(1):3346. https://doi.org/10.1038/s41598-020-60102-6 PMID: 32098977

39. Liu F, Wang J, Liu J, Li Y, Liu D, Tong J, et al. Predicting and analyzing the COVID-19 epidemic in China: Based on SEIRD, LSTM and GWR models. PLoS ONE. 2020; 15(8):e0238280. https://doi.org/10.1371/journal.pone.0238280 PMID: 32853285

40. Ge W, Huh JW, Park YR, Lee JH, Kim YH, Turchin A. An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. AMIA Annu Symp Proc. 2018; 2018:460–469. PMID: 30815086

41. Munir HS, Ren S, Mustafa M, Siddique CN, Qayyum S. Attention based GRU-LSTM for software defect prediction. PLoS ONE. 2021; 16(3):e0247444. https://doi.org/10.1371/journal.pone.0247444 PMID: 33661985

42. Lin YW, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLoS ONE. 2019; 14(7):e0218942. https://doi.org/10.1371/journal.pone.0218942 PMID: 31283759

43. Ma R, Zheng X, Wang P, Liu H, Zhang C. The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method. Sci Rep. 2021; 11(1):17421. https://doi.org/10.1038/s41598-021-97037-5 PMID: 34465820

44. Cheon SA, Cho HH, Kim J, Lee J, Kim HJ, Park TJ. Recent tuberculosis diagnosis toward the end TB strategy. J Microbiol Methods. 2016; 123:51–61. https://doi.org/10.1016/j.mimet.2016.02.007 PMID: 26853124

45. Liu Q, Li Z, Ji Y, Martinez L, Zia UH, Javaid A, 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; 12:2311–2322. https://doi.org/10.2147/IDR.S207809 PMID: 31440067

46. Feng W, Xiuran W. Design and Implementation of a New Nonlinear Combination Forecasting Model Based on RVM and Neural Network. Energy Procedia. 2011; 11(10):3714–3719.

47. Wang L, Liang C, Wu W, Wu S, Yang J, Lu X, et al. Epidemic Situation of Brucellosis in Jinzhou City of China and Prediction Using the ARIMA Model. Can J Infect Dis Med Microbiol. 2019; 2019:1429462. https://doi.org/10.1155/2019/1429462 PMID: 31312278

48. Lu C, Hao Y, Wang X. World population projections using metabolic GM (1,1) model. IEEE International Conference on Grey Systems and Intelligent Services. 2007. pp. 453–457.

49. Wang B, Ge Y. Predicting the influence of Guangfo Metro on the economic level of Foshan City Based on the GM(1,1) model. IOP Conference Series Earth and Environmental Science. 2021; 634 (1):012013. https://doi.org/10.1088/1755-1315/634/1/012013

50. Luo X, Duan H, He L. A Novel Riccati Equation Grey Model And Its Application In Forecasting Clean Energy. Energy (Oxf). 2020; 205:118085. https://doi.org/10.1016/j.energy.2020.118085 PMID: 32546893

51. Wang JZ, Ma XL, Wu J, Dong Y. Optimization models based on GM (1,1) and seasonal fluctuation for electricity demand forecasting. INTERNATIONAL JOURNAL OF ELECTRICAL POWER & ENERGY SYSTEMS.2012; 43(1):109–117. https://doi.org/10.1016/j.ijepes.2012.04.027

52. Guo Y, Feng Y, Qu F, Zhang L, Yan B, Lv J. Prediction of hepatitis E using machine learning models. PLoS ONE. 2020; 15(9):e0237750. https://doi.org/10.1371/journal.pone.0237750 PMID: 32941452

53. Navarro Valencia V, Díaz Y, Pascale JM, Boni MF, Sanchez-Galan JE. Assessing the Effect of Climate Variables on the Incidence of Dengue Cases in the Metropolitan Region of Panama City. Int J Environ Res Public Health. 2021; 18(22):12108. https://doi.org/10.3390/ijerph182212108 PMID: 34831862

54. Zhang R, Guo Z, Meng Y, Wang S, Li S, Niu R, et al. Comparison of ARIMA and LSTM in Forecasting the Incidence of HFMD Combined and Uncombined with Exogenous Meteorological Variables in Ningbo, China. Int J Environ Res Public Health. 2021; 18(11):6174. https://doi.org/10.3390/ijerph18116174 PMID: 34200378

55. Zhang X, Zhang Q, Zhang G, Nie Z, Gui Z, Que H. A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition. Int J Environ Res Public Health. 2018; 15(5):1032. https://doi.org/10.3390/ijerph15051032 PMID: 29883381

56. Togun T, Kampmann B, Stoker NG, Lipman M. Anticipating the impact of the COVID-19 pandemic on TB patients and TB control programmes. Ann Clin Microbiol Antimicrob. 2020; 19(1):21. https://doi.org/10.1186/s12941-020-00363-1 PMID: 32446305

57. Mousquer GT, Peres A, Fiegenbaum M. Pathology of TB/COVID-19 Co-Infection: The phantom menace. Tuberculosis (Edinb). 2021; 126:102020. https://doi.org/10.1016/j.tube.2020.102020 PMID: 33246269