Article

A Combined Model of SARIMA and Prophet Models in Forecasting AIDS Incidence in Henan Province, China

Zixiao Luo, Xiaocan Jia [ID], Junzhe Bao, Zhijuan Song, Huili Zhu, Mengying Liu, Yongli Yang * and Xuezhong Shi *

Department of Epidemiology and Biostatistics, College of Public Health, Zhengzhou University,
Zhengzhou 450001, China; lzxiao0206@163.com (Z.L.); jxc@zzu.edu.cn (X.J.); baojz@zzu.edu.cn (J.B.);
zjsong178@163.com (Z.S.); zhuhuili39@163.com (H.Z.); 18063131171@163.com (M.L.)
* Correspondence: ylyang377@zzu.edu.cn (Y.Y.); xzshi@zzu.edu.cn (X.S.)

Abstract: Acquired immune deficiency syndrome (AIDS) is a serious public health problem. This study aims to establish a combined model of seasonal autoregressive integrated moving average (SARIMA) and Prophet models based on an L1-norm to predict the incidence of AIDS in Henan province, China. The monthly incidences of AIDS in Henan province from 2012 to 2020 were obtained from the Health Commission of Henan Province. A SARIMA model, a Prophet model, and two combined models were adopted to fit the monthly incidence of AIDS using the data from January 2012 to December 2019. The data from January 2020 to December 2020 was used to verify. The mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were used to compare the prediction effect among the models. The results showed that the monthly incidence fluctuated from 0.05 to 0.50 per 100,000 individuals, and the monthly incidence of AIDS had a certain periodicity in Henan province. In addition, the prediction effect of the Prophet model was better than SARIMA model, the combined model was better than the single models, and the combined model based on the L1-norm had the best effect values (MSE = 0.0056, MAE = 0.0553, MAPE = 43.5337). This indicated that, compared with the L2-norm, the L1-norm improved the prediction accuracy of the combined model. The combined model of SARIMA and Prophet based on the L1-norm is a suitable method to predict the incidence of AIDS in Henan. Our findings can provide theoretical evidence for the government to formulate policies regarding AIDS prevention.

Keywords: AIDS; SARIMA model; Prophet model; L1-norm

1. Introduction

Acquired immune deficiency syndrome (AIDS) is caused by the human immune deficiency virus (HIV). According to the latest data of the Joint United Nations Programme on HIV/AIDS, approximately 37.6 million people were living with HIV and AIDS worldwide at the end of 2020 [1]. By the end of 2018, an estimated 1,250,000 people were living with HIV in China, and 80,000 new infections were reported in 2018 [2]. Henan is one of the provinces hit hardest by HIV and AIDS in China [3,4]. In the past, AIDS patients mostly appeared in paid blood donors in central China, especially in Henan province. Plasma marketing in Henan province in 1994–1995 lead to the rapid emergence of a large number of AIDS cases in the early 2000s [5]. Henan province ranks first in terms of population. Although China has responded with the prevention and treatment of AIDS, there is still a certain number of undetected HIV infections and continuous spread in society every year, which has led to a serious AIDS epidemic in Henan province [6]. The number of AIDS patients in Henan province ranked first in China in 2015, and 65,896 people were living with HIV and AIDS in Henan province in October 2020 [7]. At present, the medical field has not developed specific drugs to cure AIDS or an effective vaccine to prevent AIDS. Due to the latter, developing scientific and effective AIDS prevention and control strategies has become a top priority to curb the AIDS epidemic. Therefore, it is important to establish a

mathematical model of AIDS epidemic prediction so as to understand the process of the AIDS epidemic, explore its epidemic characteristics and develop law, and seek the optimal strategy for its prevention and control [8].

Over the past few years, mathematical models have been used to successfully predict the incidence of HIV and AIDS [9,10]. At present, the most common method used to predict disease incidences is the autoregressive integrated moving average (ARIMA) model, which is been widely used for the prediction of timeseries data, such as the incidence of human brucellosis [11] and COVID-19 [12]. However, ARIMA models have one major limitation of pre-assumed linearity [13]. In most cases, nonlinear structures are adopted during timeseries analyses, as adequate results cannot be obtained from linear models. On the contrary, a Prophet model does not need to specify a detailed model, and it simultaneously simulates multiple seasons through a generalized additive model. It adopts a Bayesian-based curve-fitting method to smooth and forecast timeseries data, and it has a fast and robust fitting process for large outliers, missing values, and dramatic changes [14]. To date, Prophet model has been used in many disciplines, including infectious diseases such as COVID-19 [15,16], but few studies have applied it to AIDS. In the present study, we have built both Seasonal Autoregressive Integrated Moving Average (SARIMA) and Prophet models because time series data for AIDS had both linear and non-linear characteristics.

It is universally agreed that combining different models can increase the chance of capturing various data features and improve prediction accuracy [17,18]. More recently, combined forecasting models have been extensively applied in various fields with high predictive performance, including air quality [19] and influenza [20]. However, the parameter estimation methods of proposed models have mainly been based on the minimum L2-norm of the prediction error vector, that is, the combination prediction model is based on the minimum sum of the squares of the prediction error. However, the prediction error becomes enlarged or reduced after the prediction error is squared [21]. Considering the defect of the L2-norm, it is necessary to introduce the prediction error vector L1-norm index, which uses the sum of the absolute values of the prediction errors. Its robustness is better than prediction residual error sum of squares, especially when there are abnormal, extreme values in the data [22,23]. Wang et al. proposed a combined model to predict air pollutant concentration. Based on the L1-norm, the model performed a weighted combination of the prediction results of three single models (extreme learning machine, Elman neural network, and support vector machine), and the results showed that the proposed combined model had a stable prediction performance [24]. Therefore, we propose a combined model of the SARIMA and Prophet models to predict AIDS incidence based on the L1-norm. We compare the prediction effects to explore which is the most precise model for AIDS incidence prediction in Henan province. The results provide reference information for AIDS prevention and intervention in Henan province.

2. Methods

2.1. Data Sources

The incidence data of AIDS in Henan from January 2012 to December 2020 were obtained from the Health Commission of Henan Province (http://wsjkw.henan.gov.cn/, accessed on 1 April 2021). The monthly incidences of AIDS in Henan province from January 2012 to December 2019 were set as the training set, and the incidences of AIDS from January to December 2020 were set as the test set. A SARIMA model, a Prophet model, and two combined models were adopted to fit the monthly incidence of AIDS by using the data from the training set. The forecasting performances of the four fitted models were verified by using the data from the test set. The technical roadmap is shown in Figure 1.
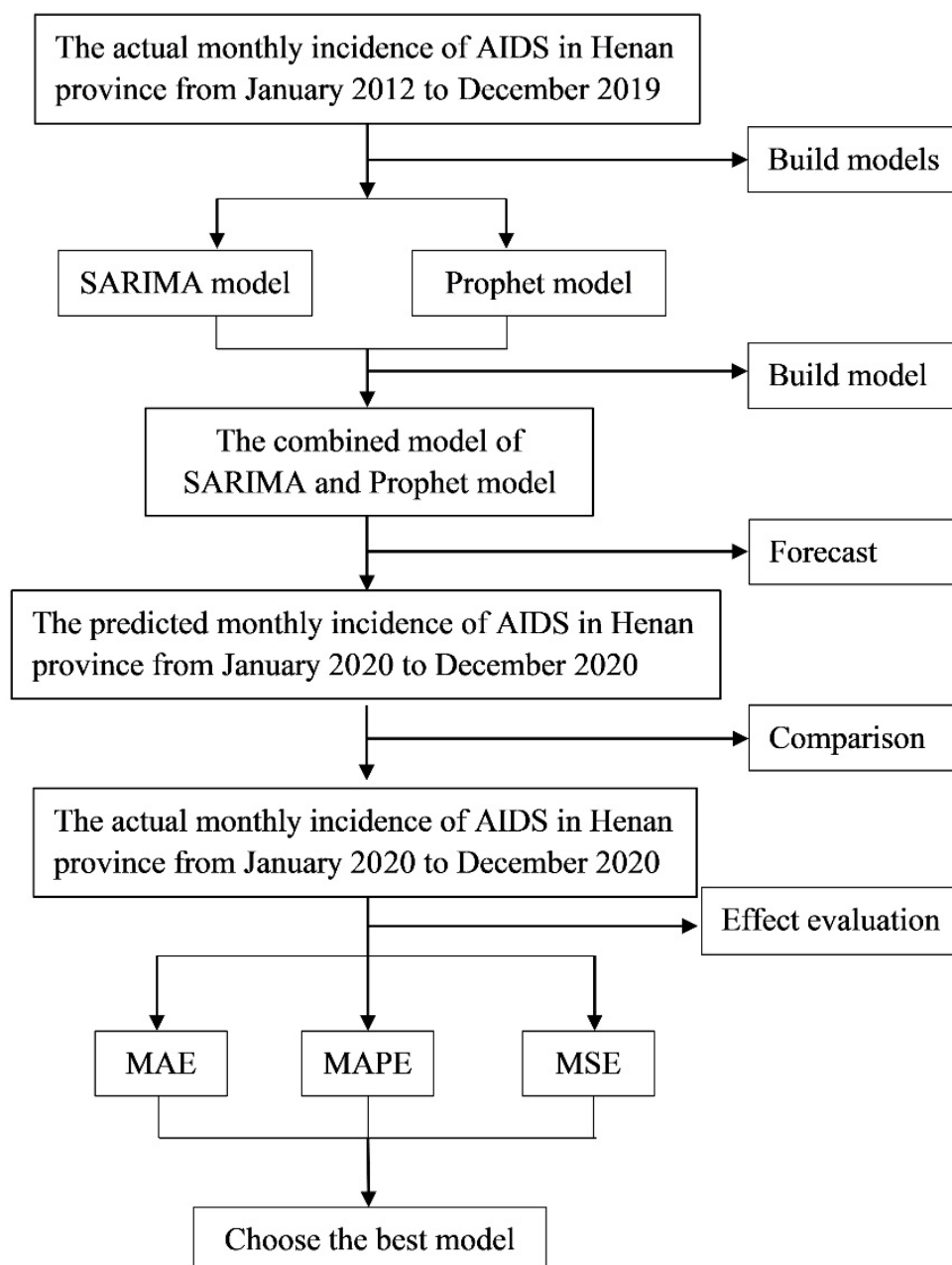
Figure 1. Technical roadmap.

2.2. SARIMA Model

ARIMA model consists of three parts: autoregression (p), the degree of difference (d), and the order of moving average (q) [25]. For seasonal trends, a SARIMA model combines nonseasonal and seasonal components and can be specified as SARIMA (p, d, q) $\times$ (P, D, Q)s, where p, d, and q refer to the orders of the nonseasonal autoregressive (AR), nonseasonal differencing, and nonseasonal moving average (MA) parts of the model. P, D, and Q refer to the orders of the seasonal AR, seasonal differencing, and seasonal MA parts of the model, and the subscripted letter "s" is the length of the seasonal period [26]. In this study, the incidence of AIDS varied in the annual cycle, so s = 12.

A timeseries modeling approach involves the following three steps: model identification, parameter estimation, and diagnostic checking. Firstly, if it is necessary, appropriate differencing of the series is performed to achieve stationarity and normality. We used an augmented Dickey–Fuller (ADF) unit root test to estimate whether the timeseries was stationary or not. Secondly, the temporal correlation structure of the transformed data

is identified by examining its autocorrelation (ACF) and partial autocorrelation (PACF) functions. In addition, the values of p, P and q, Q are finally determined by considering that smaller Akaike information criterion (AIC) and Bayesian information criterion (BIC) values correspond to a higher prediction accuracy. Finally, in order to test the normality of the SARIMA residuals, a white noise test is conducted in the residual series [27].

2.3. Prophet Model

A Prophet model is a data prediction tool for the timeseries of Facebook. It was introduced by Taylor and Letham and it is available in packages such as Python and R [14]. It can almost automatically predict the future trend of a timeseries. In addition, the model deals with the case of outliers in a time series and it also deals with some missing values. A Prophet model adopts a curve-fitting method based on Bayes to smooth and forecast timeseries data, so the results that need to be predicted can be obtained faster. In general, timeseries prediction or data analysis can use this algorithm to predict the trend of future timeseries. The formulation of a Prophet model is similar to a generalized additive model, including trend, seasonality, and holidays:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

where $g(t)$ is the trend function representing nonperiodic changes in timeseries values, and $s(t)$ represents periodic changes (for example, weekly and annual seasonality). $h(t)$ represents the effects of holidays that occur on potentially irregular schedules over one or more days. ε_t is the error term and was assumed to be normally distributed in this study.

For a trend model, it involves fitting a piecewise linear curve or a nonlinear saturating growth model. This sort of growth is typically modeled using a logistic growth model, which, in its most basic form, is:

$$g(t) = \frac{C}{1 + \exp(-k(t - m))}$$

where C indicates the carrying capacity, k is the growth rate, and m represents an offset parameter. Both the carrying capacity and rate of growth are not constant. By altering the parameter rate, the flexibility of the model can be controlled [28].

In this study, the change points were automatically selected, the number of change points was set as 25, and the carrying capacity of the logistic growth model was set as 8.5. We set the interval width as 0.85 and the number of uncertainty samples as 1000.

2.4. The Combined Model Based on L1-Norm

Supposing the observed values of an index are $\{x_t, t = 1, 2, \ldots, N\}$, there are m feasible single prediction methods to forecast it, where x_{it} is the predicted value of ith method at time t, $i = 1, 2, \ldots, m$, $t = 1, 2, \ldots, N$. When $\{l_i, i = 1, 2, \ldots, m\}$ is the weighting coefficient of m single prediction in the combined prediction model, it satisfies normality and non-negativity:

$$\sum_{i=1}^{m} l_i = 1, \ l_i \geq 0, \ i = 1, 2, \ldots, m \tag{1}$$

We first considered $\hat{x}_t$, applying weighted geometric mean:

$$\hat{x}_t = \prod_{i=1}^{m} x_{it}^{l_i}, t = 1, 2, \ldots, N \tag{2}$$

$$\ln \hat{x}_t = \sum_{i=1}^{m} l_i ln x_{it}, t = 1, 2, \ldots, N \tag{3}$$

where $\hat{x}$ is the weighted geometric average of x_t.

Supposing e_t is the logarithmic error between the combined predicted value and the corresponding actual value at time t yields the following:

$$e_{it} = ln x_t - ln x_{it} \tag{4}$$

$$e_t = lnx_t - ln\hat{x}_t = \sum_{i=1}^{m} l_i(lnx_t - lnx_{it}) = \sum_{i=1}^{m} l_i e_{it} \qquad (5)$$

where e_{it} is the logarithmic error between the actual value at time t and the corresponding predicted value of the ith single model, $i = 1, 2, \ldots, m$, $t = 1, 2, \ldots, N$.

F is the logarithmic error between the combined prediction model and the actual value of the index based on the L1-norm, and F_i is defined as the logarithmic error between the predicted value and the corresponding actual value of the ith single model, yielding:

$$F = \sum_{t=1}^{N} |e_t| = \sum_{t=1}^{N} \left| \sum_{i=1}^{m} l_i e_{it} \right| \qquad (6)$$

$$F_i = \sum_{t=1}^{N} |e_{it}|, i = 1, 2, \ldots, m \qquad (7)$$

where F is the function of the weighting coefficient vector $L = (l_1, l_2, \ldots, l_m)^T$ of various prediction methods, which can be denoted as $F(L)$.

In the ideal case, if there is no prediction error, $F(L) = 0$. However, predicted error is inevitable. The smaller $F(L)$ is, the closer the combined prediction value is to the actual value of the index, and the more accurate and effective the combined model. Therefore, model (1) was expressed as:

$$\min F(L) = \sum_{t=1}^{N} \left| \sum_{i=1}^{m} l_i e_{it} \right|$$
$$\text{s t} \begin{cases} \sum_{i=1}^{m} l_i = 1, \\ l_i \geq 0, \ i = 1, 2, \ldots, m \end{cases}$$

In this study, $m = 2$, $N = 12$, l_1 was the weighting coefficient of the SARIMA model and l_2 was the weighting coefficient of the Prophet model.

2.5. Model Evaluation

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used to screen parameters of the SARIMA model. The SARIMA model with the minimum AIC and BIC was the most suitable one. The models were estimated using the mean square error (MSE), mean absolute percentage error (MAPE), and mean absolute error (MAE), and the model with the smallest values of these indices was identified as optimal [10].

2.6. Data Processing and Analysis

R software (version 3.6.2, R Foundation for Statistical Computing, Vienna, Austria) was adopted to develop the SARIMA model and the Prophet model. In addition, LINGO software (version 15.0, Lindo System Inc., Chicago, IL, USA) was adopted to create the combined model. The significant level was 0.05.

3. Results

3.1. Trends of AIDS in Henan Province

The timeseries data covered 108 months, from January 2012 to December 2020. As shown in Figure 2, the monthly incidence fluctuated from 0.05 to 0.50 per 100,000 individuals, and the monthly incidence of AIDS in Henan province had a certain periodicity. The peaks of the disease occurred mainly in December, while the low peaks occurred mainly in January or February. There was a sudden decline in the monthly incidence of AIDS in January 2020 and a record low in February 2020.
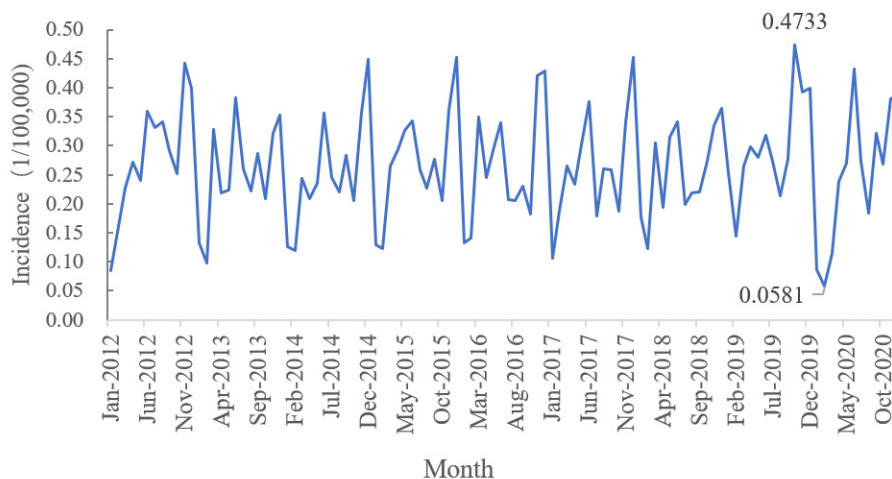
Figure 2. The timeseries diagram of monthly incidence of AIDS in Henan province from 2012−2020.

3.2. SARIMA Models

For SARIMA, the ADF test indicated that the original series was stationary (Dickey–Fuller = − 3.796, p < 0.05), which did not need trend differencing ACF and PACF graphs were used to estimate the parameter ranges of p, P and q, Q. We found that the plots of the original series displayed slow decay at the seasonal lags (Figure 3a). Therefore, at lag-12 (subtracting the observations after a lag of 12 periods) differencing was used to remove seasonality features. The sequence of one-order seasonal difference was stable (Figure 3b). Then, some candidate SARIMA models were assessed to forecast future values based on the previously observed values (Table 1). Further, of all the tested models shown in Table 1, the SARIMA(1,0,1)(0,1,1)[12] model was found to best fit the data, which had the lowest Akaike information criterion (AIC = −253.67) and Bayesian information criterion (BIC = −244.00). This model also passed the Ljung–Box Q Test (p = 0.420); the testing results of the estimated parameters were all statistically significant (p < 0.05).
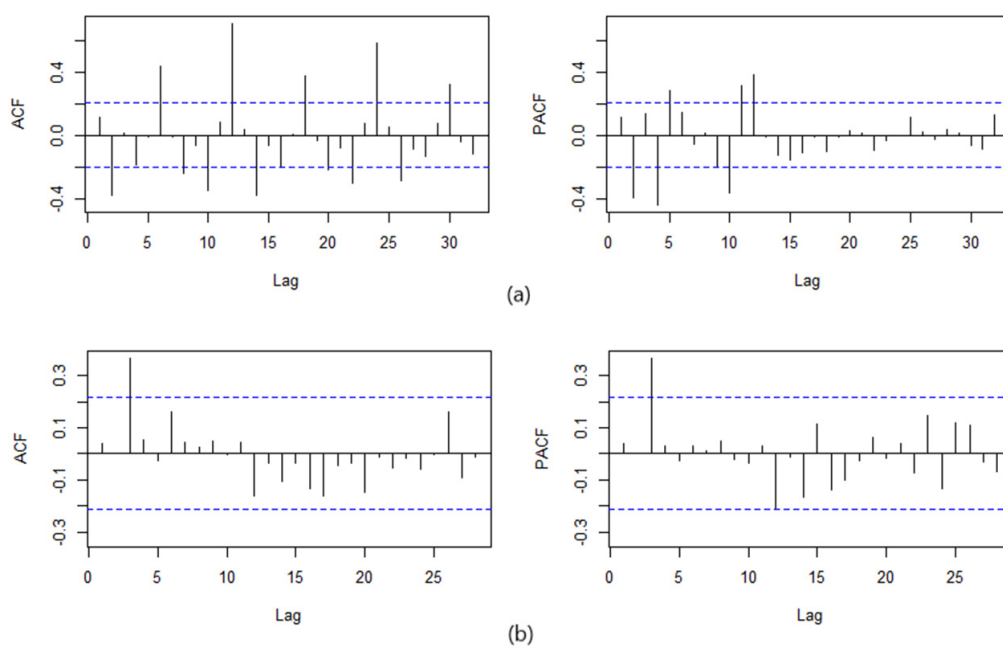

Figure 3. The ACF and PACF plots of monthly incidence of AIDS in Henan province from 2012−2019: (**a**) plots before seasonal difference and (**b**) plots of one-order seasonal difference.

Table 1. Comparison of candidate SARIMA models.

Model	Estimate	*p*-Value	Ljung–Box Q Test			AIC	BIC	RMSE	MAPE
			Statistics	DF	*p*-Value				
SARIMA(0,1,1)(0,1,0)[12]			19.42	17	0.305	−240.68	−235.87	0.050	14.008
MA1	−0.868	<0.001							
SARIMA(0,1,1)(1,1,2)[12]			14.96	15	0.454	−240.51	−228.48	0.046	12.249
MA1	−0.882	<0.001							
SAR1	−0.780	<0.001							
SMA1	0.567	0.046							
SMA2	−0.433	0.044							
SARIMA(1,0,1)(0,1,0)[12]			19.40	16	0.249	−249.89	−242.63	0.048	12.411
AR1	−0.777	<0.001							
MA1	1.000	<0.001							
SARIMA(1,0,1)(0,1,1)[12]			15.45	15	0.420	−253.67	−244.00	0.045	11.888
AR1	−0.751	<0.001							
MA1	1.000	<0.001							
SMA1	−0.398	0.019							
SARIMA(1,0,1)(1,1,0)[12]			16.85	15	0.328	−252.53	−242.85	0.046	12.178
AR1	−0.746	<0.001							
MA1	1.000	<0.001							
SAR1	−0.297	0.025							
SARIMA(1,0,1)(1,1,1)[12]			13.26	14	0.506	−252.07	−239.98	0.045	11.754
AR1	−0.759	<0.001							
MA1	1.000	<0.001							
SAR1	0.289	0.477							
SMA1	−0.688	0.084							
SARIMA(2,0,2)(0,1,0)[12]			18.18	16	0.313	−247.64	−235.55	0.047	12.780
AR1	−1.435	<0.001							
AR2	−0.925	<0.001							
MA1	1.581	<0.001							
MA2	1.000	<0.001							
SARIMA(2,0,2)(0,1,1)[12]			16.67	15	0.339	−251.68	−237.17	0.045	12.266
AR1	−1.067	<0.001							
AR2	0.950	<0.001							
MA1	1.907	<0.001							
MA2	0.830	<0.001							
SMA1	−0.500	<0.001							
SARIMA(3,0,0)(0,1,0)[12]			15.10	17	0.588	−256.66	−246.98	0.045	12.564
AR1	0.062	0.536							
AR2	−0.018	0.859							
AR3	0.469	<0.001							

AIC: Akaike information criterion; BIC: Bayesian information criterion; RMSE: root mean squared error; MAPE: mean absolute percent error; DF: degree of freedom.

Finally, we predicted the monthly incidence of AIDS from January to December in 2020 with the SARIMA(1,0,1)(0,1,1)[12] model. The results are shown in Table 2.

Table 2. Comparison of predicted values and actual values from January to December 2020 (per 100,000 population).

Time	Actual Value	Predicted Value			
		SARIMA(1,0,1)(0,1,1)[12]	Prophet Model	Combined Model Based on L2-Norm	Combined Model Based on L1-Norm
January-2020	0.087	0.235	0.168	0.205	0.199
February-2020	0.058	0.126	0.142	0.133	0.135
March-2020	0.114	0.289	0.305	0.296	0.298
April-2020	0.238	0.255	0.239	0.248	0.246
May-2020	0.270	0.298	0.262	0.282	0.279
June-2020	0.432	0.326	0.369	0.345	0.349
July-2020	0.273	0.247	0.250	0.249	0.249
August-2020	0.183	0.217	0.274	0.243	0.248
September-2020	0.322	0.262	0.273	0.267	0.268
October-2020	0.269	0.379	0.201	0.299	0.283
November-2020	0.382	0.375	0.409	0.390	0.393
December-2020	0.387	0.397	0.423	0.409	0.411

Int. J. Environ. Res. Public Health **2022**, *19*, 5910

8 of 12

3.3. Prophet Model

The prophet model was automatically fitted with the incidence rates in Henan province from January 2012 to December 2019, and then the AIDS incidence rates from January to December 2020 were predicted. The predicted results are shown in Table 2, and the fitting prediction curve is shown in Figure 4. The results showed that the monthly incidence of AIDS was seasonal in Henan province.
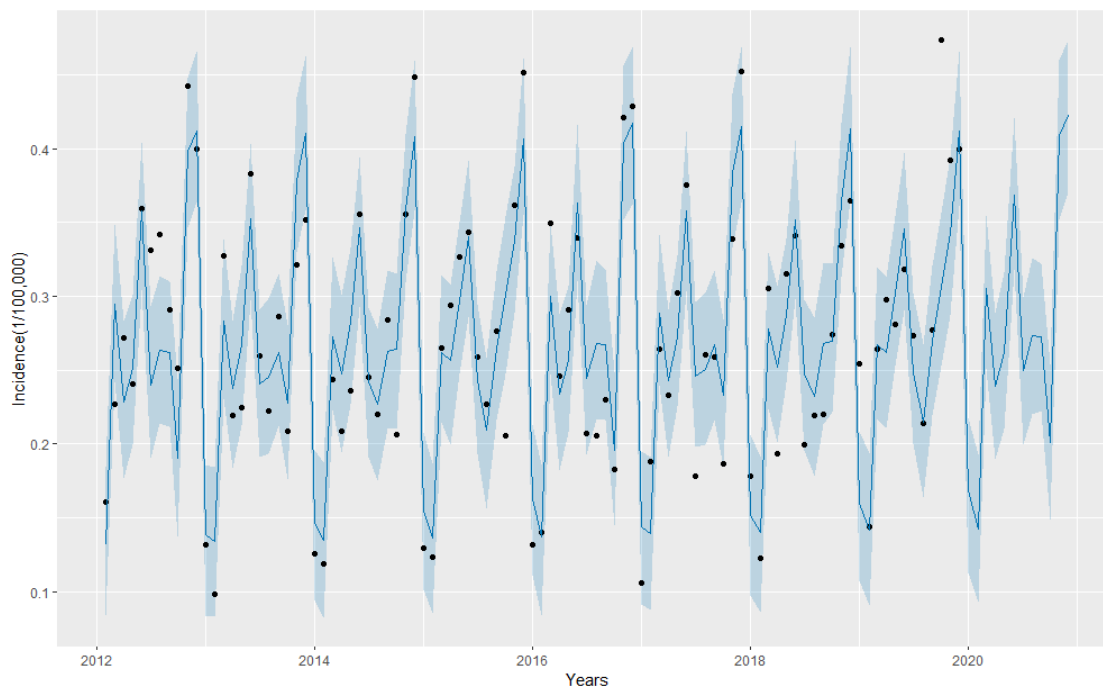

Figure 4. Predicted new infection rate of AIDS from January 2012 to December 2020 by Prophet model. The black dots represent the observed values, the blue line represents the fitted or predicted data of Prophet model, and the shadow area represents 95% confidence intervals.

The decomposed components of the monthly incidence of AIDS included the effect of trend and the yearly seasonality (Figure 5). An increasing trend in the reported incidence of AIDS was observed from 2012 to 2020. For yearly seasonality, an apparent local maximum appeared in September, and an apparent local minimum appeared in November.
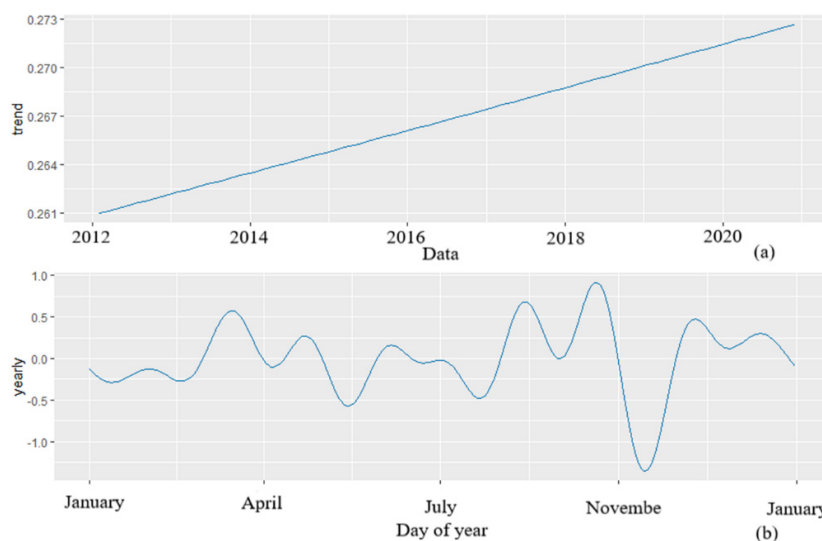

Figure 5. The decomposed components of the monthly incidence of AIDS in Henan province, China: (**a**) the effect of trend and (**b**) the effect of the yearly seasonality.

3.4. The Combined Model

For the combined model based on the L2-norm, we obtained $l_1 = 0.548$ and $l_2 = 0.452$. For the combined model based on the L1-norm, we obtained $l_1 = 0.4587$ and $l_2 = 0.5417$. The prediction values of the combined models are shown in Table 3.

Table 3. Effect evaluation of models.

Model	MSE	MAE	MAPE
SARIMA(1,0,1)(0,1,1)[12]	0.0073	0.0657	47.8470
Prophet Model	0.0060	0.0602	44.8336
Combined Model based on L2-norm	0.0060	0.057	44.1950
Combined Model based on L1-norm	0.0056	0.0553	43.5337

MSE: the mean square error; MAE: mean absolute error; MAPE: mean absolute percentage error.

According to model (1), the minimum log error of the combined model based on the L1-norm was $F(l_1, l_2) = 3.5479$, while the logarithmic error of the SARIMA model was $F_1 = 4.0063$ and that of the Prophet model was $F_2 = 3.8331$. The combinatorial model based on the L1-norm was the optimal combinatorial model, as it resulted in $F(l_1, l_2) < \min\{F_1, F_2\}$.

3.5. Model Evaluation

Compared with the other models, the effect values of the combined model based on L1 norm were all lower (Table 3). It showed that the prediction effect of the combined model based on L1 was the best.

4. Discussion

In this study, we focused on monthly incidence data for AIDS from 2012 to 2020 in Henan province, China. The results showed that the incidence of AIDS in Henan province was relatively stable. We built four models for analyzing the timeseries: a SARIMA model, a Prophet model, and two combined models of SARIMA and Prophet. The predictive abilities of the four models were compared and it was discovered that the combined model based on the L1-norm proposed in this study had the best model effect values and was superior to the other models in predicting effect. These findings indicated the potential value of the combined model based on the L1-norm in forecasting short-term AIDS incidence in Henan province. The model can provide reference for AIDS prevention and intervention in Henan province.

The Henan provincial government has done an excellent job in the prevention and treatment of AIDS during a transfer of AIDS carriers to a new AIDS population in Henan province. It warned us that the AIDS epidemic cannot be ignored. The Henan government has not slackened on AIDS in recent years. To implement China's 13th Five-Year Action Plan for HIV/AIDS Containment and Prevention, the Henan University AIDS Prevention and Control Alliance was established in 2018 [7]. In the face of the impact of COVID-19, we still need to pay attention to the prevention and control of AIDS [29,30]. It requires us to go further in AIDS prevention, treatment, and research and to raise public awareness of the growing threat posed by infectious diseases. In addition, AIDS is not a seasonal infection. Still, the findings suggested some seasonality in the AIDS incidence data reported through the Center for Disease Control and Prevention (CDC): the rate of new infections is low in January and February and this can be attributed to the influence of the annual Spring Festival, which falls in late January or early February. During the Spring Festival, national and provincial CDCs and most hospitals and clinical laboratories operate with limited capacity, and people's willingness to seek medical treatment falls, resulting in low AIDS infection records. Some other studies have also supported the latter phenomenon [31].

Different models have their own merits and faults. Based on historical AIDS incidence data, our study used timeseries analysis methods to establish SARIMA(1,0,1)(0,1,1)[12], a Prophet model, and two combined models for forecasting the monthly incidence of AIDS in Henan province. We found that the prediction effect of the Prophet model was better

than the SARIMA model. We know that the SARIMA model combines the advantages of the two methods of regression analysis and moving average, which can analyze seasonal timeseries [32]. A previous study found that the most appropriate ARIMA models for HIV incidence in 2015 and 2016 in Guangxi, China, were ARIMA(1,1,2)(0,1,2)[12] and ARIMA(2,1,0)(1,1,2)[12], respectively [8]. A limitation of the SARIMA model is that it is easy to overfit with poor generalization ability when processing daily data, and it is more suitable for linear models [13]. However, the monthly incidences of AIDS had both linear and nonlinear characteristics, and there was a missing value, which meant some information may not have been captured by the SARIMA model. Adopting a generalized additive model formulation, the Prophet model is fast in its fitting procedure and robust for large outliers, missing values, and dramatic changes. The Prophet model automatically fitted the data, and the interpolation of missing values was not required [14], so it produced better results. This may explain why the performances of the SARIMA model were not as accurate as those of the Prophet model, which was consistent with a previous study [33].

Previously, Adesoye Idowu Abioye et al. [34] used an ARIMA and Prophet model to fit and predict the COVID-19 cases in Nigeria from September to October 2020. On this basis, we built combined models to predict the incidence of AIDS in Henan province. The results showed that the prediction effects of the combined models were better than those of the two single models. Firstly, the combined models could combine the characteristics of a single model to capture more data information, and they were more suitable for fitting data with both linear and nonlinear characteristics, such as AIDS incidence. Secondly, the reasonable distribution of the weight of a single model may result in an ensemble with more accurate and lower variance [17]. To date, many related studies have shown that a combined model is better than a single model. For example, Grzegorz Dudek [35] believed that a hybrid residual dilated long short-term memory and exponential smoothing model was more competitive. Li [36] found that a combined model of modified linear ARIMA and modified nonlinear ARIMA improved the single models. Furthermore, we found that, compared with the L2-norm, the L1-norm improved the prediction accuracy of the combined model. The combined model based on the L2-norm was based on the minimum sum of the squares of the prediction error. However, the prediction error was enlarged or reduced after the prediction error was squared (that is, if the absolute value of the prediction absolute error was greater than 1, it was larger after it was squared; if the absolute value of the predicted relative error was less than 1, it was smaller after being squared). The combined model based on the L1-norm used the sum of the absolute values of the prediction errors. Based on the L1-norm, the model was sparse and regularized. The L1-norm decreased very quickly for small weights and slowly for large weights [24]. Overall, the combined forecasting model of SARIMA and Prophet models based on the L1-norm was an appropriate way for predicting the incidence of AIDS in Henan province. With the help of the combined model, it is reasonable for the government to allocate health resources to control the epidemic efficiently.

This study had several limitations. First, since the analyzed data was the monthly incidence data of AIDS, the decomposition components of the Prophet model only included the trend effect and yearly effect, which could not reflect the weekly effect and short-term holiday effect. Second, the timeseries prediction model was adopted without considering factors affecting the incidence of AIDS, such as age, gender, social status, epidemic variation, and humanity. Stratified analysis is essential for a better understanding of AIDS epidemiology. In the future, we can explore other suitable models (such as support vector regression, exponential smoothing models, and machine learning) for predicting AIDS in combination with epidemiological data and socio-economic determinants [37,38].

5. Conclusions

The combined forecasting model of the SARIMA and Prophet models based on the L1-norm was an appropriate way for predicting the incidence of AIDS in Henan province. The results showed that the combined model was suggested for use in AIDS surveillance,

which provide evidence for the government to formulate policies by providing estimates on AIDS incidence trends in Henan, China.

References

1. Joint United Nations Programme on HIV/AIDS. Available online: https://www.unaids.org/en (accessed on 1 April 2021).
2. Cao, W.; Hsieh, E.; Li, T. Optimizing Treatment for Adults with HIV/AIDS in China: Successes over Two Decades and Remaining Challenges. *Curr. HIV AIDS Rep.* **2020**, *17*, 26–34. [CrossRef] [PubMed]
3. Beach, M.V. "Blood heads" and AIDS haunt China's countryside. *Lancet* **2001**, *357*, 49. [CrossRef]
4. Wu, Z.; Liu, Z.; Detels, R. HIV-1 infection in commercial plasma donors in China. *Lancet* **1995**, *346*, 61–62. [CrossRef]
5. Zhang, F.J.; Pan, J.; Yu, L.; Wen, Y.; Zhao, Y. Current progress of China's free ART program. *Cell Res.* **2005**, *15*, 877–882. [CrossRef]
6. Qiao, Y.C.; Xu, Y.; Jiang, D.X.; Wang, X.; Wang, F.; Yang, J.; Wei, Y.S. Epidemiological analyses of regional and age differences of HIV/AIDS prevalence in China, 2004–2016. *Int. J. Infect. Dis.* **2019**, *81*, 215–220. [CrossRef]
7. Health Commission of Henan Province. Available online: http://wsjkw.henan.gov.cn/ (accessed on 1 April 2021).
8. Zang, X.; Krebs, E.; Wang, L.; Marshall, B.D.; Granich, R.; Schackman, B.R.; Montaner, J.S.; Nosyk, B. Structural Design and Data Requirements for Simulation Modelling in HIV/AIDS: A Narrative Review. *PharmacoEconomics* **2019**, *37*, 1219–1239. [CrossRef]
9. Wang, G.; Wei, W.; Jiang, J.; Ning, C.; Chen, H.; Huang, J.; Liang, B.; Zang, N.; Liao, Y.; Chen, R.; 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. [CrossRef]
10. 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*, 143. [CrossRef]
11. Cao, L.T.; Liu, H.H.; Li, J.; Yin, X.D.; Duan, Y.; Wang, J. Relationship of meteorological factors and human brucellosis in Hebei province, China. *Sci. Total Environ.* **2020**, *703*, 135491. [CrossRef]
12. Malki, Z.; Atlam, E.S.; Ewis, A.; Dagnew, G.; Alzighaibi, A.R.; Elmarhomy, G.; Elhosseini, M.A.; Hassanien, A.E.; Gad, I. ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Comput. Appl.* **2020**, *33*, 2929–2948. [CrossRef]
13. Elsheikh, A.H.; Saba, A.I.; Abd Elaziz, M.; Lu, S.; Shanmugan, S.; Muthuramalingam, T.; Kumar, R.; Mosleh, A.O.; Essa, F.A.; Shehabeldeen, T.A. Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia. *Process Saf. Environ. Prot. Trans. Inst. Chem. Eng. Part B* **2021**, *149*, 223–233. [CrossRef] [PubMed]
14. Taylor, S.J.; Letham, B. Forecasting at Scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]
15. Talkhi, N.; Akhavan Fatemi, N.; Ataei, Z.; Jabbari Nooghabi, M. Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods. *Biomed. Signal Process. Control* **2021**, *66*, 102494. [CrossRef] [PubMed]
16. Tulshyan, V.; Sharma, D.; Mittal, M. An Eye on the Future of COVID-19: Prediction of Likely Positive Cases and Fatality in India over a 30-Day Horizon Using the Prophet Model. *Disaster Med. Public Health Prep.* **2020**, *18*, 1–7. [CrossRef] [PubMed]
17. Reich, N.G.; McGowan, C.J.; Yamana, T.K.; Tushar, A.; Ray, E.L.; Osthus, D.; Kandula, S.; Brooks, L.C.; Crawford-Crudell, W.; Gibson, G.C.; et al. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S. *PLoS Comput. Biol.* **2019**, *15*, e1007486. [CrossRef]
18. English, T.M. Stacked generalization and simulated evolution. *Bio Syst.* **1996**, *39*, 3–18. [CrossRef]
19. Zhu, S.; Yang, L.; Wang, W.; Liu, X.; Lu, M.; Shen, X. Optimal-combined model for air quality index forecasting: 5 cities in North China. *Environ. Pollut.* **2018**, *243*, 842–850. [CrossRef]
20. Yamana, T.K.; Kandula, S.; Shaman, J. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. *PLoS Comput. Biol.* **2017**, *13*, e1005801. [CrossRef]

Int. J. Environ. Res. Public Health **2022**, *19*, 5910

12 of 12

21. Kwak, N. Principal component analysis based on l1-norm maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1672–1680. [CrossRef]

22. Wang, C.; Ye, Q.; Luo, P.; Ye, N.; Fu, L. Robust capped L1-norm twin support vector machine. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **2019**, *114*, 47–59. [CrossRef]

23. Wu, D.; Shang, M.; Luo, X.; Wang, Z. An L_1-and-L_2-Norm-Oriented Latent Factor Model for Recommender Systems. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *22*, 1–14. [CrossRef] [PubMed]

24. Wang, B.; Jiang, Q.; Jiang, P. A combined forecasting structure based on the L(1) norm: Application to the air quality. *J. Environ. Manag.* **2019**, *246*, 299–313. [CrossRef] [PubMed]

25. Tobias, A.; Díaz, J.; Saez, M.; Alberdi, J.C. Use of poisson regression and box-jenkins models to evaluate the short-term effects of environmental noise levels on daily emergency admissions in Madrid, Spain. *Eur. J. Epidemiol.* **2001**, *17*, 765–771. [CrossRef] [PubMed]

26. Bas, M.D.; Ortiz, J.; Ballesteros, L.; Martorell, S. Evaluation of a multiple linear regression model and SARIMA model in forecasting (7)Be air concentrations. *Chemosphere* **2017**, *177*, 326–333. [CrossRef] [PubMed]

27. Xia, Y.; Liao, C.; Wu, D.; Liu, Y. Dynamic Analysis and Prediction of Food Nitrogen Footprint of Urban and Rural Residents in Shanghai. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1760. [CrossRef] [PubMed]

28. Devaraj, J.; Elavarasan, R.M.; Pugazhendhi, R.; Shafiullah, G.M.; Ganesan, S.; Jeysree, A.K.; Khan, I.A.; Hossain, E. Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant? *Results Phys.* **2021**, *21*, 103817. [CrossRef]

29. Brown, L.B.; Spinelli, M.A.; Gandhi, M. The interplay between HIV and COVID-19: Summary of the data and responses to date. *Curr. Opin. HIV AIDS* **2021**, *16*, 63–73. [CrossRef]

30. Anonymous. How to stop COVID-19 fuelling a resurgence of AIDS, malaria and tuberculosis. *Nature* **2020**, *584*, 169. [CrossRef]

31. Xu, B.; Li, J.; Wang, M. Epidemiological and time series analysis on the incidence and death of AIDS and HIV in China. *BMC Public Health* **2020**, *20*, 1906. [CrossRef]

32. Mao, Q.; Zhang, K.; Yan, W.; Cheng, C. Forecasting the incidence of tuberculosis in China using the seasonal auto-regressive integrated moving average (SARIMA) model. *J. Infect. Public Health* **2018**, *11*, 707–712. [CrossRef]

33. Lu, J.; Meyer, S. Forecasting Flu Activity in the United States: Benchmarking an Endemic-Epidemic Beta Model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1381. [CrossRef] [PubMed]

34. Adesoye Idowu Abioye, M.D.U.; Peter, O.J.; Edogbanya, H.O.; Oguntolu, F.A.; Kayode, O.; Amadiegwu, S. Forecasting of COVID-19 pandemic in Nigeria using real statistical data. *Commun. Math. Biol. Neurosci.* **2021**, *2021*, 2052–2541. [CrossRef]

35. Dudek, G.; Pelka, P.; Smyl, S. A Hybrid Residual Dilated LSTM and Exponential Smoothing Model for Midterm Electric Load Forecasting. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *8*, 1–13. [CrossRef] [PubMed]

36. Shuyu, L.; Xuan, Y.L.R. Forecasting coal consumption in India by 2030: Using linear modified linear (MGM-ARIMA) and linear modified nonlinear (BP-ARIMA) combined models. *Sustainability* **2019**, *11*, 695. [CrossRef]

37. Oshinubi, K.; Rachdi, M.; Demongeot, J. Modeling of COVID-19 Pandemic vis-à-vis Some Socio-Economic Factors. *Front. Appl. Math. Stat.* **2022**, *7*, 786983. [CrossRef]

38. Oshinubi, K.; Rachdi, M.; Demongeot, J. Analysis of Reproduction Number R0 of COVID-19 Using Current Health Expenditure as Gross Domestic Product Percentage (CHE/GDP) across Countries. *Healthcare* **2021**, *9*, 1247. [CrossRef]