Epidemiology and Infection

cambridge.org/hyg

Original Paper


Received: 8 January 2021
Revised: 19 June 2021
Accepted: 20 July 2021

Key words: Hybrid model; LS-SVM; SARIMA; varicella

Author for correspondence: Mengliang Ye,
E-mail: yemengliang@cqmu.edu.cn

Introduction

Varicella is a highly infectious contagious disease caused by varicella-zoster virus [1]. Relevant studies have shown [2, 3] that varicella has obvious seasonality, with one or two peaks per year, often breaking out in winter and spring in temperate regions. In Spain, the incidence of varicella peaked from May to July, with a low incidence in October [4]. Giammanco et al. showed that varicella was one of the common childhood diseases [5]. In China, Bao et al. [6], Cao et al. [7] and Bai et al. [8] have described the epidemic situation of varicella in Wuhan, Wuxi and Shenyang, respectively. Their studies have shown that the incidence of varicella has obvious seasonality, and it mostly happens in student groups. According to the literature [9], a total of 3,047,715 cases of varicella were reported from 2016 to 2019, including 30 deaths in China. The annual reported incidence and mortality rates were 5505/100,000 and 0.0005/100,000, respectively. In 2018, the incidence of varicella in Chongqing ranked second in China, with a rate of 120.50/100,000, second only to Jiangsu Province. Chongqing is the largest city and economic centre in Southwest China. In 2018, the permanent resident population of Chongqing was about 31.02 million, and the proportion of children aged 0–14 is about 16.93%. To analyse the characteristics of varicella epidemic and select the appropriate prediction model to forecast the incidence of varicella in Chongqing, so as to provide an important epidemiological basis for the prevention and control of varicella in the future, is the current issue to be discussed.

For the prediction models of varicella, in foreign, Soysal et al. conducted a temporal trend study on the incidence of varicella in Turkey [10]. Giraldo et al. used an infectious disease dynamic model to conduct a preliminary study of varicella [11–13]. Lee et al. discussed the incidence of varicella in South Korean children [14]. In China, there are more descriptive studies on varicella [15–18]. Some scholars used the infectious disease dynamics model to predict the varicella in Changsha [19] and analysed the spatial aggregation of varicella in Jilin province [20], while others used ARIMA model [21] and grey model [22] to predict the incidence of varicella. In general, the SARIMA model can only analyse the linear information, but cannot deal with the non-linear information [23]. However, the least squares support vector machine (LS-SVM) is a kind of support vector model (SVM), which is not only suitable for small samples, but also can solve non-linear information well [24]. Considering the advantages and disadvantages of the prediction methods and the amount of research data, a single prediction model and a combined prediction model were established, respectively, based on the varicella data, and the seasonality of varicella was analysed. By comparing the prediction errors of different models, the best prediction model was selected.
The best prediction model was used for short-term prediction to provide reference information for the prevention and intervention of varicella in Chongqing.

Materials and methods

Materials

The monthly incidence of varicella in Chongqing from January 2014 to December 2018 was studied in this paper, and the monthly incidence of varicella data is primarily gained from the Chongqing CDC.

Methods

SARIMA model

Compared with the ARIMA model, the SARIMA model introduces one more seasonal effect, and the modelling process is similar to the ARIMA model. The SARIMA expression is [25]

\[
\nabla^d \nabla^s x_t = \frac{\Theta(B) \Theta_s(B)}{\Phi(B) \Phi_s(B)} e_t
\]

\[\Theta(B) = 1 - \theta_1 B - \cdots - \theta_p B^p\]

\[\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p\]

\[\Theta_s(B) = 1 - \theta_1 B^s - \cdots - \theta_Q B^{Qs}\]

\[\Phi_s(B) = 1 - \phi_1 B^s - \cdots - \phi_p B^{ps}\]

\(B\) is the backward shift operator, \(e_t\) is the estimated residual at time \(t\) with zero mean and constant variance and \(x_t\) denotes the observed value at time \(t\) \((t = 1, 2 \ldots k)\), \(s\) is the length of the seasonal period, \(p, P, d, D, q\) and \(Q\) are the autoregressive order, seasonal autoregressive order, number of difference, number of seasonal difference, moving average order and seasonal moving average order, respectively.

SARIMA model modelling steps

First, judge the stationarity of the sequence, and make the sequence stable through appropriate methods. Second, according to the tailing and truncation of the autocorrelation coefficient and partial autocorrelation coefficient, determine the four main parameter values of the model \((p, q, P, Q)\). Then, residual and parameter tests were carried out for the model. Compare the AIC and BIC values between the models, and choose the optimal model with the smallest two index values. Finally, the optimal model was used for prediction.

Hybrid model

The difference between the optimal SARIMA model-fitting value \(\hat{y}_i\) and the actual value \(y_i\) constitutes the residual sequence \(e_i = \hat{y}_i - y_i\), and normalise the residual sequence [26], then, fitting the LS-SVM model with the residual as the sample. Assuming a training set \((x_i, y_i), x \in R, y \in R, i = 1, 2, \cdots, l\) of \(l\) data, \(x_i\) is the input data, \(y_i\) is the output data, and the objective optimisation function of the LS-SVM algorithm is:

\[
\min J(\omega, \epsilon) = \frac{1}{2} \omega^T \omega + \frac{1}{2} \gamma \sum_{i=1}^{l} e_i^2
\]

s.t. \(y_i = \omega^T \phi(x_i) + b + cM_i, i = 1, 2, \cdots, l\)

### Table 1. Distribution of varicella by sex, age and occupation in Chongqing from 2014 to 2018

<table>
<thead>
<tr>
<th>Variable</th>
<th>2014 (n = 11,684) (%)</th>
<th>2015 (n = 18,269) (%)</th>
<th>2016 (n = 19,022) (%)</th>
<th>2017 (n = 26,243) (%)</th>
<th>2018 (n = 37,055) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>6172 (52.82)</td>
<td>9709 (53.14)</td>
<td>10,062 (52.90)</td>
<td>13,742 (52.36)</td>
<td>19,212 (51.85)</td>
</tr>
<tr>
<td>Female</td>
<td>5512 (47.18)</td>
<td>8560 (46.86)</td>
<td>8960 (47.10)</td>
<td>12,501 (47.64)</td>
<td>17,843 (48.15)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–9</td>
<td>7057 (60.40)</td>
<td>11,098 (60.7)</td>
<td>10,585 (55.65)</td>
<td>14,392 (54.84)</td>
<td>20,143 (54.36)</td>
</tr>
<tr>
<td>10–19</td>
<td>3283 (28.10)</td>
<td>5240 (28.68)</td>
<td>6161 (32.39)</td>
<td>8812 (33.58)</td>
<td>12,975 (35.02)</td>
</tr>
<tr>
<td>20–29</td>
<td>1021 (8.74)</td>
<td>1444 (7.90)</td>
<td>1626 (8.55)</td>
<td>2003 (7.63)</td>
<td>2427 (6.55)</td>
</tr>
<tr>
<td>30–39</td>
<td>288 (2.46)</td>
<td>417 (2.28)</td>
<td>581 (3.05)</td>
<td>926 (3.53)</td>
<td>1370 (3.70)</td>
</tr>
<tr>
<td>40–49</td>
<td>23 (0.20)</td>
<td>54 (0.30)</td>
<td>48 (0.25)</td>
<td>84 (0.32)</td>
<td>106 (0.29)</td>
</tr>
<tr>
<td>≥50</td>
<td>12 (0.10)</td>
<td>16 (0.09)</td>
<td>21 (0.11)</td>
<td>26 (0.10)</td>
<td>34 (0.09)</td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workers</td>
<td>197 (1.69)</td>
<td>277 (1.52)</td>
<td>295 (1.55)</td>
<td>360 (1.37)</td>
<td>486 (1.32)</td>
</tr>
<tr>
<td>Scattered children</td>
<td>1226 (10.49)</td>
<td>1901 (10.41)</td>
<td>1963 (10.32)</td>
<td>2455 (9.35)</td>
<td>3108 (8.39)</td>
</tr>
<tr>
<td>Children in kindergarten</td>
<td>2397 (20.52)</td>
<td>3842 (21.03)</td>
<td>3406 (17.91)</td>
<td>4562 (17.38)</td>
<td>6946 (18.75)</td>
</tr>
<tr>
<td>Students</td>
<td>6729 (57.59)</td>
<td>10,613 (58.09)</td>
<td>11,400 (59.93)</td>
<td>16,309 (62.15)</td>
<td>23,149 (62.47)</td>
</tr>
<tr>
<td>Farmers</td>
<td>227 (1.94)</td>
<td>339 (1.86)</td>
<td>354 (1.86)</td>
<td>453 (1.73)</td>
<td>612 (1.65)</td>
</tr>
<tr>
<td>Teachers</td>
<td>56 (0.48)</td>
<td>103 (0.56)</td>
<td>104 (0.55)</td>
<td>163 (0.62)</td>
<td>247 (0.67)</td>
</tr>
<tr>
<td>Others</td>
<td>852 (7.29)</td>
<td>1194 (6.54)</td>
<td>1500 (7.89)</td>
<td>1941 (7.40)</td>
<td>2505 (6.76)</td>
</tr>
</tbody>
</table>

https://doi.org/10.1017/S0950268821001813 Published online by Cambridge University Press
In the formula, \( \phi(\cdot):R^n \rightarrow R^{nh} \) is the kernel space mapping function; \( e_i \) is the error variable; \( \gamma \) is the adjustment parameter factor.

Sample data normalisation formula:

\[
x^* = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}
\]

Anti-normalisation formula:

\[
x' = (x_{\text{max}} - x_{\text{min}}) x^* + x_{\text{min}}
\]

where \( x_i \) is sample data, \( x_{\text{max}}, x_{\text{min}} \) are the maximum and minimum values of the sample data, respectively, \( x^* \) is the normalised data, \( x \) is the predicted value, \( x' \) is the anti-normalisation value.

The root mean square error (RMSE) and mean absolute percentage error (MAPE) were used to compare the fitting effect.

The RMSE and MAPE calculation formulas are [25]:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{x}_t)^2}
\]

\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{x_t - \hat{x}_t}{x_t} \right|
\]

In the above equation, \( x_t \) is the actual incidence value, \( \hat{x}_t \) is the estimated incidence value, \( n \) is the amount of months for

Fig. 1. Trend, seasonal and residual components derived from ‘STL’ decomposition of monthly varicella incidence for Chongqing during 2014–2018.

Table 2. The seasonal index after the decomposition of ‘STL’ function

<table>
<thead>
<tr>
<th>Month</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>December</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasonal index</td>
<td>1.1495</td>
<td>0.4185</td>
<td>0.5781</td>
<td>1.0232</td>
<td>1.8454</td>
<td>2.0374</td>
<td>0.9039</td>
<td>0.3800</td>
<td>0.4907</td>
<td>1.3517</td>
<td>2.0686</td>
<td>1.9826</td>
</tr>
</tbody>
</table>
Fig. 2. Reported monthly incidence of varicella from January 2014 to June 2018.

Fig. 3. Sequence diagram after a one-step difference and seasonal difference with a period of 12.
forecasting. The lower the RMSE value and MAPE value, the better the data fitting effect.

**Results**

**Descriptive analyses**

Table 1 shows that this study reported 112,273 varicella cases in the past 5 years (2014–2018), in Chongqing, including 58,897 males and 53,376 females, and a male-to-female ratio of 1.1034:1. Varicella mostly occurs within the ages of 0–9 years ($n = 63,275$), what is more, the age group of 0–9 accounted for 56.36% of all reported cases. The highest percentage of varicella cases was found in students, which amount to 60.74% ($n = 68,200$), followed by children in kindergarten and scattered children.

**SARIMA model construction**

This study used the ‘STL’ function to decompose the sequence, Figure 1 shows that the sequence has obvious seasonality, and the incidence rate presents an upward trend over time. Table 2 shows that the peak incidence of varicella was from April to June and October to December in Chongqing, and the seasonal index was >1. According to the time series diagram (Fig. 2), the monthly incidence of varicella presented a non-stationary state. After the difference processing of the original sequence, the data presented a stationary state (Fig. 3), and the unit root test showed that the sequence was stationary ($P < 0.05$). From the autocorrelation and partial autocorrelation graphs of the sequence (Fig. 4), the autocorrelation coefficient and partial autocorrelation coefficient showed tailing. Considering that the value of $p$, $q$, $P$ and $Q$ does not exceed 2, we verify the four parameters from 0 to 2, respectively. Only six models passed the residual test and parameter test, the six models were SARIMA$(1, 1, 1) \times (1, 1, 0)_{12}$, SARIMA$(2, 1, 2) \times (1, 1, 1)_{12}$, SARIMA$(1, 1, 1) \times (1, 1, 1)_{12}$, SARIMA$(2, 1, 1) \times (1, 1, 1)_{12}$, SARIMA$(2, 1, 2) \times (1, 1, 0)_{12}$, SARIMA$(1, 1, 1) \times (0, 1, 1)_{12}$, respectively. By comparing the AIC, BIC values and two error indicators of the six models in Table 3, SARIMA$(2, 1, 1) \times (1, 1, 1)_{12}$ model is finally selected as the best model in this paper.
Table 4 shows the estimated, standard errors and significance values of model parameters, all the parameter tests were statistically significant. In addition, the $P$-values of LB statistics at order 6 and 12 of delay were 0.9091 and 0.6901, respectively. The white noise test of residuals was significant that indicates the fitted SARIMA(2, 1, 1) × (1, 1, 1)12 model was sufficient. The model equation is given as

$$\nabla^2 x_t = \frac{(1 - 0.5510B)(1 - 0.9997B^2)}{(1 - 0.5554B + 0.3933B^2)(1 + 0.4038B^{12})} e_t, \quad e_t 
\sim N(0, 0.01676)$$

SARIMA(2, 1, 1) × (1, 1, 1)12 model was used to forecast the incidence of varicella. Table 5 shows the value of prediction; RMSE and MAPE values are 0.7843 and 0.0654, respectively. The actual value of incidence and fitted incidence of SARIMA model monthly is shown in Figure 5. As shown in Figure 5 and Table 5, the tendency and epidemics from predicted incidence are very close to the actual value of incidence and epidemic circumstance of varicella.

**Hybrid model construction**

First, we took the residual sequence of SARIMA(2, 1, 1) × (1, 1, 1)12 model from January 2014 to June 2018 as the training set, the residual from July 2018 to December 2018 as the test set, and normalise the training samples. Then, we choose RBF kernel function for the LS-SVM kernel function, take different values for the embedding dimension $m$ and the time delay $\tau$, compare the prediction errors, and finally determine that the prediction error is the smallest when $m$ is 3 and $\tau$ is 12. That is, using the incidence of the same period in the first 3 years to predict the incidence of the same period in the fourth year, after 50 times, iterative parameter values tend to be stable. Then, sample reconstruction was performed, and the optimal parameters $\gamma$ and $\sigma$ were solved by genetic algorithm with the values of 8.8540 and 110.8799, respectively, so as to establish the optimal combination model. Finally, the residual was predicted and the inverse normalisation was carried out to obtain the predicted residual value (Table 6); the predicted value of the monthly incidence of varicella obtained by the combination model was $y^* = \hat{y}_t + \hat{\epsilon}_t$ (Table 5).

**Model comparison**

First, compare the fitting effects of the two models. It can be seen from Figure 6 that the fitting value of the mixed model is between...
the actual value and a single model. Second, a comparison of the prediction effects of the two models, from Table 5 and Figure 7, shows that the mixed model has a slightly smaller value of RMSE and MAPE, and the predicted value of the mixed model is closer to the actual value. Thus it can be seen that the best prediction model is the mixed model.

Discussion

The descriptive analysis of varicella shows that the ratio of male and female is approximately equal, and the high incidence of varicella occurs in students, children in kindergarten and scattered children, so the incidence of varicella can be effectively controlled in this age group. The decomposing of the sequence by the ‘STL’ function not only shows the trend and seasonal changes of the varicella incidence sequence, but also calculates the seasonal index of each month, and can intuitively understand its seasonality. In this paper, we can conclude that the peak incidence of varicella in Chongqing was from April to June and October to December, and the periods from February to March and August to September were two low stages of the disease, which is consistent with relevant studies [27–29]. The trough period may be related to the students’ winter and summer vacations. During the winter and summer vacations, children’s exposure opportunities are significantly reduced. Therefore, it is necessary to strengthen the intervention measures to avoid infection during the high incidence of varicella.

SARIMA model is suitable for the complex interaction among the sequential seasonal effects, long-term trends and random fluctuations. This model is one of the time series analysis models commonly used in the prediction of infectious diseases, such as tuberculosis [30], hand-foot-mouth disease [31], conjunctivitis [32], mumps [33], influenza [34] and other infectious diseases. We use the SARIMA model to perform linear fitting on the varicella series. By comparing the AIC, BIC values and combining the RMSE and MAPE values, SARIMA(2, 1, 1) × (1, 1, 1)₁₂ is the best model, and the RMSE and MAPE values of this model are 0.7843 and 0.0654, respectively. It can be seen from the fitting diagram (Fig. 5) that there was a very good match between the observed values and the fitted values, the 95% CI of the forecast value contain all of the real observed data, and SARIMA(2, 1, 1) × (1, 1, 1)₁₂ model can extract the deterministic information in the sequence.
Considering that infectious diseases will be affected by external factors and internal factors of the human body, with irregular changes and non-linear dynamic characteristics, the combined model of SARIMA and LS-SVM combines linear analysis with non-linear analysis.

The SVM has greater potential and better performance in practical applications [35–37]. LS-SVM uses the second square of the error as the loss function, and replaces the inequality constraints with equations to simplify the SVM algorithm, reducing the complexity of the algorithm; furthermore, LS-SVM maintains the advantages of the standard SVM. At present, SVM has been gradually introduced into the field of infectious diseases, such as bacillary dysentery [38], hepatitis B [25], hand-foot-mouth disease [39] and so on. In this study, we chose the RBF kernel function as the kernel function of the LS-SVM model; compared with other kernel functions, the parameters are easier to choose, the space complexity changed little and it was easy to implement.

As can be seen from Table 5, the predicted RMSE value of SARIMA model is 0.7843 and MAPE value is 0.0654, while the predicted RMSE value of the mixed model is 0.7525 and MAPE value is 0.0647. Compared with the single SARIMA model, the mixed model has the advantage of treating the non-linear part of the residual error. In addition, Figure 6 shows that both the single model and the mixed model can well reflect the trend, peak and change trend of the actual varicella incidence. However, the fitting value and predicted value of the mixed model are between the actual value and the single model (Figs 6 and 7), indicating that the prediction effect of the mixed model is better. The mixed model can not only describe the periodicity and seasonal variation of varicella incidence in Chongqing, but also fit the non-linear part well.

In conclusion, although the prediction effect of the model is relatively good, prevention and control work should be done as soon as possible for the high incidence of varicella, strengthen daily disinfection in public places, and large-scale vaccination and other prevention and control measures should also be taken. In order to improve the accuracy of the prediction model, it is necessary to keep updating the data in the future analysis, so that the model can be optimised continuously and reflect the law and development trend of the data.

Conclusions

Based upon the results of this study, applying the hybrid models to forecast the incidence of varicella is feasible. The fitted value and predicted value in the mixed model have the same trend as the actual value of varicella incidence, and the curve is relatively close. It suggests that a hybrid model can be used to predict the incidence of varicella. The short-term prediction of varicella is very effective, which is helpful for the evaluation of prevention or control measures. Meanwhile, we can adopt timely and effective countermeasures for the epidemic peak that may occur.
Acknowledgements. The authors express their thanks to the Chongqing Municipal Center for Disease Control and Prevention for the disease data as well as the help from teachers of Chongqing Medical University.

Author contributions. H.Q. and H.Z. contributed equally to this paper. Conceptualisation, M.Y.; Methodology, H.Q. and H.Z.; Software, H.Q.; Validation, R.O.; Formal analysis, H.Q.; Investigation, H.Q. and H.Z.; Resources, H.Z.; Data curation, Q.C. and Q.W.; Writing – original draft preparation, H.Q.; Writing – review and editing, H.Q., R.O. and M.Y.; Visualisation, Q.C. and Q.W.; Supervision, M.Y. and R.O.; Project administration, M.Y. All authors read and approved the final manuscript.


Conflict of interest. None.

Data availability statement. The incidence of varicella data are gained from the Chongqing Center of Disease and Control; it is confidential data and cannot be uploaded to your organisation. The incidence is equal to the number of new cases of a disease in a population during a period divided by the number of people exposed during the same period.

References


