Original Article

A Hybrid Model for Short-Term Bacillary Dysentery Prediction in Yichang City, China

Weirong Yan*, Yong Xu1, Xiaobing Yang1, and Yikai Zhou2

Department of Epidemiology and Biostatistics, and
1Department of Environmental Medicine, School of Public Health,
Tongji Medical College of Huazhong University of Science & Technology, Wuhan; and
2Yichang Center for Disease Control and Prevention, Yichang, China

(Received August 7, 2009. Accepted June 9, 2010)

SUMMARY: Bacillary dysentery is still a common and serious public health problem in China. This paper is aimed at developing and evaluating an innovative hybrid model, which combines the seasonal autoregressive integrated moving average (SARIMA) and the generalized regression neural network (GRNN) models, for bacillary dysentery forecasting. Data of monthly bacillary dysentery incidence in Yichang City from 2000–2007 was obtained from Yichang Disease Control and Prevention Center. The SARIMA and SARIMA-GRNN model were developed and validated by dividing the data file into two data sets: data from the past 5 years was used to construct the models, and data from January to June of the 6th year was used to validate them. Simulation and forecasting performance was evaluated and compared between the two models. The hybrid SARIMA-GRNN model was found to outperform the SARIMA model with the lower mean square error, mean absolute error, and mean absolute percentage error in simulation and prediction results. Developing and applying the SARIMA-GRNN hybrid model is an effective decision supportive method for producing reliable forecasts of bacillary dysentery for the study area.

INTRODUCTION

Bacillary dysentery, mainly caused by Shigella bacteria, is endemic throughout the world. Patients who are infected with Shigella often develop diarrhea and/or dysentery with frequent mucoid bloody stools accompanied by fever, abdominal cramps, and tenesmus. Excessive dehydration can be fatal in some severe cases and is thought responsible for 120 million cases of severe dysentery, the majority of which occur in developing countries, especially among children and the elderly (1–3).

In China, although the incidence rate of bacillary dysentery has declined considerably over the past decade together with the rapid socioeconomic development that has occurred in this country, it is still a serious public health problem. It was reported that 312,522 new cases of bacillary dysentery occurred in 2008, ranking number three among the notifiable infectious diseases in the nation (4). Bacillary dysentery can quickly develop into large-scale epidemics in overcrowded areas with poor sanitation as transmission usually occurs via contaminated water and food or through person-to-person contact. Reliable forecasting of the incidence rate and early recognition of the epidemic pattern are significantly important for disease prevention and control.

Statistical models are practical tools in public health surveillance. With better forecasting performance of the model class used in disease analysis, public health officials can enhance the understanding of epidemic patterns in order to prepare for intervention measures in advance. Many models such as generalized regression, Markov chain model, Grey model, and autoregressive moving average (ARMA) class (including AR, ARMA, and ARIMA) models have been applied in infectious disease forecasting research. Among them, ARMA class models have been widely used (5–8). Based on stationary time-series data, linear models can be constructed through curve fitting and parameter estimation. In modeling nonlinear series, it is necessary to turn to other approaches (9). Forecasting based on unitary linear models (such as ARMA class models) does not provide satisfactory incidence data of infectious diseases, which generally exhibit some nonlinear attributes.

Artificial neural networks (ANN) have the advantage of approximating nonlinear functions. This technique has been successful in analyzing series for which the mathematical knowledge of the stochastic process behind the series is either unknown, or difficult to rationalize (10). In recent years, researchers have examined how well the neural network model performed compared with the linear model using data from industry, finance, medical research, and others (11–13).

The occurrence of bacillary dysentery has its own rules, affected by the speed of pathogen variation, susceptible accumulation, and environmental changes. The real-world incidence data of bacillary dysentery exhibits both linear and nonlinear patterns (14), thus a single model is not the best way of forecasting. In the current study, a combined new model which integrated the seasonal time series ARIMA (SARIMA) model and generalized regression neural network (GRNN) model was

*Corresponding author: Mailing address: Department of Epidemiology and Biostatistics, School of Public Health, Tongji Medical College of Huazhong University of Science & Technology, Hangkong Road 13#, Wuhan 430030, China. E-mail: weirongy@gmail.com
constructed, in order to improve the accuracy of prediction by taking advantage of each model’s capabilities. Based on literature searches using the key words ‘statistical model,’ ‘neural networks,’ and ‘forecasting,’ this is the first study to develop and apply a SARIMA-GRNN hybrid model for forecasting the incidence of bacillary dysentery.

MATERIALS AND METHODS

Study area and data collection: Yichang City, located in the west of Hubei Province in China, has a temperate climate and a population of about 4.1 million. Bacillary dysentery is a legally mandated notifiable disease in China. All notifiable cases are based on clinical and laboratory diagnosis in hospitals. Monthly incidence data of bacillary dysentery for the period 2000–2007 were provided by a passive surveillance system at the Yichang Disease Control and Prevention Center. The time series data of the incidence of bacillary dysentery showed a strong seasonality trend, with the higher incidence rate in June–August each year (Fig. 1). A 5-year data set for the past 5 years was used as the training data to construct the forecasting model, while the data for the first 6 months of the 6th year was used to test the forecasting ability of the model. Study was repeated for two sets of data.

Data analysis: The SARIMA and the neural network GRNN models were summarized in the following as the foundation to describe the SARIMA-GRNN model.

Development of SARIMA model: In an ARMA model, the future value of a variable is assumed to be a linear function of several past observations and random errors. It uses autoregressive parameters, moving average parameters and the number of differencing passes to describe a series in which a pattern is repeated over time. The SARIMA model is developed from the ARMA models. Because bacillary dysentery incidence data represents a clear seasonal variation, the SARIMA model may be appropriate for allowing long-term trends and seasonal effects (15–17). In this study, SARIMA models were developed using monthly incidence of bacillary dysentery as the dependent variable and its past variables as independent variable with the control of seasonality.

The general SARIMA model has the following form:

\[ \phi(B)\Phi(B')(1 - B)^d(1 - B)^qX_t = \theta(B)\Theta(B')e_t \]  \[1\]

with

\[ \phi(B) = 1 - \phi_1B - \phi_2B^2 - \cdots - \phi_pB^p \]

\[ \Phi(B') = 1 - \Phi_1B' - \Phi_2B'^2 - \cdots - \Phi_pB'^p \]

\[ \theta(B) = 1 - \theta_1B - \theta_2B^2 - \cdots - \theta_qB^q \]

\[ \Theta(B') = 1 - \Theta_1B' - \Theta_2B'^2 - \cdots - \Theta_qB'^q \]

In the equation, \( B \) 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 \( t \) (\( t = 1, 2, \ldots, k \)).

Six main parameters were selected when fitting the SARIMA model: the order of autoregressive (\( p \)) and seasonal autoregressive (\( P \)), the order of integration (\( d \)) and seasonal integration (\( D \)), and the order of moving average (\( q \)) and seasonal moving average (\( Q \)). The process is called SARIMA \( (p, d, q) \) \( (P, D, Q)s \) \( (s \) is the length of the seasonal period). Autocorrelation function (ACF) and partial autocorrelation function (PACF) were performed to identify the order of autoregressive and moving average of the time series. Akaike information criterion (AIC) and Schwarz Bayesian criterion (SBC) were used to determine the optimal model that most closely fit the data. The SARIMA model was built using a SPSS 13.0 statistical package and a \( P \) value of less than 0.05 was used for statistical significance.

Development of GRNN model: ANN has the characteristics of self-organizing and self-learning processes. Neural networks are “trained” by presenting a set of data together with the outcomes that the trainer wishes the network to learn. The trained neural network can then be evaluated by inputting similar, but previously unseen data. It reduces the model’s reliance on the prior information and removes the need for the operator to correctly specify the precise functional form of the relationship that the model seeks to represent (18–20).

In spite of the wide application of ANN in multivariate nonlinear problems, the main disadvantage of ANNs is that they are a form of black box: the rela-

Fig. 1. Incidence rate of bacillary dysentery (BD) in Yichang City.
tionship that the network finds cannot be expressed easily in mathematical form. The model is implicit, hidden in the network structure and optimized weights, between the nodes (19,21,22).

Among ANNs, GRNN is a universal approximator for smooth functions based on nonlinear regression theory. It is able to solve any smooth function-approximation problem given enough data. The training set of a GRNN network consists of values of inputs \(x\), each with a corresponding value of an output \(y\). This regression method produces the estimated value of \(y\), which minimizes the squared error. GRNN is based on the following formula:

\[
E[y|x] = \frac{1}{\sqrt{2\pi \sigma^2}} \int_{-\infty}^{\infty} yf(x,y)dy \\
= \frac{1}{\sqrt{2\pi \sigma^2}} \int_{-\infty}^{\infty} f(x,y)dy
\]

Where \(y\) is the output of the estimator, \(x\) is the estimator input vector, \(E[y|x]\) is the is the expected value of the output \(y\) with a given input vector \(x\), and \(f(x,y)\) is the joint probability density function (PDF) of \(x\) and \(y\). Choosing an appropriate smooth factor \(\sigma\) is of critical importance for GRNN, because it determines how tightly the network matches its predictions to the data in the training patterns. Generally, it requires a certain amount of experimentation. The development of a GRNN model involves the following four steps: (i) to switch all data into \([0, 1]\) intervals; (ii) identification of the input and output variables of the network; (iii) determination of an appropriate smoothing factor \(\sigma\); and (iv) training the network with the selected smoothing factor \(\delta\). After the four-step model building process, the final established model can be used for prediction purposes. The GRNN model was developed with the neural network toolbox in Matlab 7.0.

Combining SARIMA model with GRNN model: Based on the established SARIMA model, the estimated monthly incidence of bacillary dysentery can be obtained. For each month, the estimated and the actual incidence rates are generally different since the SARIMA model is used to analyze the linear part of the actual data, and the residuals from the linear model will contain the nonlinear relationship (23,24). If the association between the estimated values from SARIMA model and the actual data can be evaluated, the nonlinear component of the latter can be captured.

The structure of the hybrid model is shown in Figure 2. There are two input variables. One is the estimated incidence at time \(t\) based on the SARIMA model, and the other was time \(t\). There is only one output variable \(Y\), the actual incidence at time \(t\). Through the process of GRNN learning and simulating data repeatedly, the relationship between the input and output variables can be determined. The trained neural network can then be evaluated by inputting similar, but previously unseen data. When the estimated incidence from the SARIMA model is obtained, the corresponding forecasting incidence from the hybrid model can also be obtained.

Performance statistic index: In this study, three criteria were used to make comparisons of the forecasting capabilities of the SARIMA and SARIMA-GRNN hybrid model.

One index was the mean square error (MSE):

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{X}_i)^2
\]

The second was the mean absolute error (MAE):

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |X_i - \hat{X}_i|
\]

And the last one was the mean absolute percentage error (MAPE):

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|X_i - \hat{X}_i|}{X_i}
\]

Where \(X_i\) the real incidence at time \(t\), \(\hat{X}_i\) is the estimated incidence at time \(t\), and \(n\) is the number of predictions.

RESULTS

SARIMA model equation: In the SARIMA time series analysis, two 5-year data sets from January 2000 to December 2005 and January 2001 to December 2006 were used to construct models accordingly. The best model generated from the data set was SARIMA (0, 1, 1) \((1, 1, 0)_{12}\). It took the first-order regular difference and the first seasonal difference in order to remove the growth trend and the seasonality characteristics. The estimations of model parameters and their testing results were presented in Table 1.

\[
(1 + 0.409B^{12})(1 - B)(1 - B^{12})X_i = (1 - 0.823B)e_i, \text{VAR}(e_i) = 0.267
\]

\[
(1 + 0.484B^{12})(1 - B)(1 - B^{12})X_i = (1 - 0.912B)e_i, \text{VAR}(e_i) = 0.225
\]

[6] Incidence data from January 2000 to December 2005 as the training dataset.

[7] Incidence data from January 2001 to December 2006 as the training dataset.

SARIMA-GRNN hybrid model construction: As described before, the SARIMA-GRNN hybrid model inte-

![Fig. 2. Topological structure of the hybrid model.](image)

Table 1. SARIMA model coefficients and their testing results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE of coefficient</th>
<th>(t)</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-year data set from January 2000 to December 2005 as the training data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA1</td>
<td>0.823</td>
<td>0.081</td>
<td>10.169</td>
<td>0.000</td>
</tr>
<tr>
<td>SAR1</td>
<td>-0.409</td>
<td>0.121</td>
<td>-3.371</td>
<td>0.001</td>
</tr>
<tr>
<td>5-year data set from January 2001 to December 2006 as the training data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA1</td>
<td>0.912</td>
<td>0.073</td>
<td>12.505</td>
<td>0.000</td>
</tr>
<tr>
<td>SAR1</td>
<td>-0.484</td>
<td>0.118</td>
<td>-4.100</td>
<td>0.000</td>
</tr>
</tbody>
</table>
grated the SARIMA model with the GRNN model. Because the simulation accuracy of the GRNN model was determined by the smoothing factor $\delta$, $\delta$ was experimentally changed to between 0.1 and 0.5 to achieve sufficient accuracy after all data was switched into $[0, 1]$ intervals. When the smoothing factor was equal to 0.1, the hybrid model had the related lowest in-sample (training set) errors (Table 2).

**Comparison between SARIMA and SARIMA-GRNN model:** In this part, the fitting and prediction eff-

<table>
<thead>
<tr>
<th>Error</th>
<th>$\delta = 0.1$</th>
<th>$\delta = 0.2$</th>
<th>$\delta = 0.3$</th>
<th>$\delta = 0.4$</th>
<th>$\delta = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.051</td>
<td>0.126</td>
<td>0.164</td>
<td>0.185</td>
<td>0.209</td>
</tr>
<tr>
<td>MAE</td>
<td>0.177</td>
<td>0.276</td>
<td>0.310</td>
<td>0.326</td>
<td>0.344</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.079</td>
<td>0.124</td>
<td>0.143</td>
<td>0.156</td>
<td>0.173</td>
</tr>
</tbody>
</table>

MSE, mean square error; MAE, mean absolute error; MAPE, mean absolute percentage error.

**Fitting effects of two models:** After taking the first-order regular difference and the first seasonal difference, the first 13 values were lost and the fitted values of the remaining 59 values in the training set were compared. It was shown that the hybrid model performed better than the SARIMA model from Figs. 3A and B and Figs. 4A and B. The combined model had a lower MSE, MAE, and MAPE than the single SARIMA model.

**Prediction results of the two models:** The prediction results for the first 6 months in the 6th year from two models were reported in Tables 3A and B. The hybrid model showed a better performance than the SARIMA model with the lower MSE, MAE, and MAPE.

**DISCUSSION**

Early recognition of epidemic behavior is signifi-
Fig. 4. (A) Incidence of bacillary dysentery (BD) and its estimated value from SARIMA model for the training set. January 2001–December 2006. MSE = 0.223; MAE = 0.363; MAPE = 0.161. (B) Incidence of bacillary dysentery and its estimated value from SARIMA-GRNN model for the training set. January 2001–December 2006. MSE = 0.073; MAE = 0.194; MAPE = 0.082. The smaller MSE, MAE, and MAPE mean the better fitting effects with the applied model.

Cantly important for disease prevention and control. The capacity of statistical models to forecast future disease incidence values exemplifies their usefulness (25). In this paper, we have described and compared two models, the SARIMA and the SARIMA-GRNN hybrid model applied in the forecasting of the epidemic pattern of bacillary dysentery.

With the efforts of Box and Jenkins (16), the ARMA class models have become one of the most popular methods in forecasting research and practice. The SARIMA model that comprises an inherent seasonal function was applied in the current study.

In previous studies, researchers developed multivariate SARIMA models with some meteorological variables, such as temperature, rainfall, air pressure, and humidity as explanatory variables to enable a better predicting performance in infectious disease analysis (15,26–28). Bacillary dysentery is not only influenced by meteorological factors, but also by many other biological, social, and environmental factors. The observed incidence series was produced by the interaction of many known and unknown factors. In the current study, instead of applying a multivariate SARIMA model, we developed a univariate model to identify the linear structure of the observed data. Some studies also showed that the multivariate ARIMA model could not beat its univariate counterpart for forecasting performance (29,30). On the other hand, from a practical point of view, the univariate model is mostly used for its simplicity. For the multivariate ARIMA model, firstly, the selection of explanatory variables and knowledge of its values for the given time periods are required. Further, the lag lengths of the independent values need to be estimated.
Table 3. Prediction results of SARIMA and SARIMA-GRNN models

(A) Observation and forecasted value from January to June 2006 (1/100,000)

<table>
<thead>
<tr>
<th>Date</th>
<th>Observation</th>
<th>SARIMA model</th>
<th>SARIMA-GRNN model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2006</td>
<td>1.093</td>
<td>0.819</td>
<td>1.247</td>
</tr>
<tr>
<td>Feb. 2006</td>
<td>1.169</td>
<td>0.754</td>
<td>1.086</td>
</tr>
<tr>
<td>Mar. 2006</td>
<td>1.471</td>
<td>1.276</td>
<td>1.624</td>
</tr>
<tr>
<td>Apr. 2006</td>
<td>1.855</td>
<td>1.550</td>
<td>1.795</td>
</tr>
<tr>
<td>May 2006</td>
<td>2.414</td>
<td>2.611</td>
<td>2.713</td>
</tr>
</tbody>
</table>

Prediction error
MSE 0.086
MAE 0.283
MAPE 0.180

(B) Observation and forecasted value from January to June 2007 (1/100,000)

<table>
<thead>
<tr>
<th>Date</th>
<th>Observation</th>
<th>SARIMA model</th>
<th>SARIMA-GRNN model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 2007</td>
<td>0.971</td>
<td>1.127</td>
<td>1.151</td>
</tr>
<tr>
<td>Feb. 2007</td>
<td>1.253</td>
<td>0.970</td>
<td>1.120</td>
</tr>
<tr>
<td>Mar. 2007</td>
<td>1.578</td>
<td>1.494</td>
<td>1.550</td>
</tr>
<tr>
<td>Apr. 2007</td>
<td>2.166</td>
<td>1.818</td>
<td>1.813</td>
</tr>
<tr>
<td>May. 2007</td>
<td>2.608</td>
<td>2.648</td>
<td>2.692</td>
</tr>
</tbody>
</table>

Prediction error
MSE 0.066
MAE 0.219
MAPE 0.121

As opposed to the ARMA class models, ANN have been widely accepted as a potentially useful way of modeling complex nonlinear and dynamic systems in medical research (12,13,31). Compared with the widely used back propagation neural network (BPNN) method, GRNN, which is based on nonlinear regression theory for function estimation, has advantages of fast training time, great stability, and relatively easy network parameter settings during the training stage (32). GRNN was selected as a preferred solution to the hybrid model for nonlinear characteristics analysis of the incidence series of bacillary dysentery.

It is almost universally agreed within the forecasting literature that no single method is the best in every situation. By combining different models, different aspects of the underlying patterns of original data may be captured and prediction accuracy improved over individual models (23,33,34). In the current study, the SARIMA model was applied to extract linear information from the actual data. Then the relationship between the observed data and estimations from the SARIMA model, which represented the nonlinear component of the actual data, was learned and analyzed repeatedly by the GRNN model. In addition to the estimations from the SARIMA model, another input variable in the GRNN model is a time indicator. Time is very useful for the modeling of periodic components of a series, as those exhibited by the incidence time series of bacillary dysentery. With the hybrid model, the linear and nonlinear components of the raw data were extracted sufficiently and the overall modeling and forecasting performance were improved.

The limitations of this study should be acknowledged. Firstly, under reporting is inevitable in passive disease surveillance systems such as where we obtained our data for the current study. The notified cases are those with severe symptoms that chose to visit doctors in a hospital. Secondly, weekly data would provide a more accurate result than monthly data, as we used. Thirdly, the study only focused on a medium sized city, Yichang in China. The developed model must be generalized with caution as the situation in other areas may differ. Fourthly, since the ANN model belongs to the black-box type of models, it may limit the model’s ability to extrapolate beyond its training data as well as the implementation of subjective initiatives by operators in ANN analysis.

The results showed that the SARIMA-GRNN model is superior to the single SARIMA model for the prediction of bacillary dysentery in Yichang City. The MSE, MAE, and MAPE were lower with the hybrid model in data fitting and extrapolated forecasting than the single SARIMA model. Constructing the SARIMA-GRNN model is likely to be an effective method for the forecasting of the epidemic pattern of bacillary dysentery. The construction and explanation of the model should be further explored. Future work with this prediction technique will include the expansion of the training set and validation with a larger set of incidence data.

REFERENCES

hourly and daily diffuse fraction using neural network, as compared to linear regression models. Energy, 32, 1513–1523.


