Research Article

Application of a Hybrid Method Combining Grey Model and Back Propagation Artificial Neural Networks to Forecast Hepatitis B in China

Ruijing Gan, Xiaojun Chen, Yu Yan, and Daizheng Huang

School of Preclinical Medicine, Guangxi Medical University, No. 22, Shuangyong Road, Nanning, Guangxi 530021, China

Correspondence should be addressed to Daizheng Huang; huangdaizheng@gmail.com

Received 20 September 2014; Revised 22 January 2015; Accepted 22 January 2015

Academic Editor: Chung-Min Liao

Copyright © 2015 Ruijing Gan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Accurate incidence forecasting of infectious disease provides potentially valuable insights in its own right. It is critical for early prevention and may contribute to health services management and syndrome surveillance. This study investigates the use of a hybrid algorithm combining grey model (GM) and back propagation artificial neural networks (BP-ANN) to forecast hepatitis B in China based on the yearly numbers of hepatitis B and to evaluate the method's feasibility. The results showed that the proposal method has advantages over GM (1, 1) and GM (2, 1) in all the evaluation indexes.

1. Introduction

Hepatitis B is a vaccine preventable disease caused by the hepatitis B virus (HBV) that can induce potentially fatal liver damage. It has infected approximately 2 billion people worldwide, which represents one-third of the world population. Each year around the world, HBV infection is responsible for about one million deaths due to liver failure and cirrhosis and more than 75% of the hepatocellular carcinomas worldwide develop from HBV infection [1-3]. HBV is most prevalent in China, South East Asia, sub-Saharan Africa, and the Amazon basin of South America where health care resources are most limited [4]. In the Chinese population of 1.3 billion individuals, there are estimated to be 93 million HBV carriers. Each year, 300,000 deaths are attributed to chronic hepatitis B, including deaths associated with liver cirrhosis and hepatocellular carcinoma (HCC) [5]. Therefore, it is critical for early prevention of hepatitis B and an accurate forecasting which would enable public health officials to evaluate intervention strategies and make educated decisions.

Mathematical and computational models have gained in importance in the public-health domain, especially in infectious disease epidemiology, by providing rationales and quantitative analysis to support decision-making and policymaking processes in recent years. And many researchers advocate the use of these models as predictive tools [6–12].

The accurate forecasting of hepatitis B can be obtained by analyzing the sufficient historical data. However, in China and perhaps some other developing countries, the current public health surveillance system does not collect detailed essential epidemiological information as they are often difficult to obtain. The forecasted of hepatitis B will be inaccurate only by the limited data. Therefore, it is significant to make the limited data-processing.

The grey systems theory chiefly including the theory of grey system analysis, modeling, prediction, decisionmaking, and control is established by Deng, which focuses on uncertainty problems with small samples, discrete data and incomplete information that are difficult for probability, and fuzzy mathematics to handle. Grey prediction is an important embranchment of grey systems theory, which makes scientific, quantitative forecasts about the future states of grey systems. The precise prediction of system can be performed by generating and extracting the useful information from the small samples and the partially known information [13–15].

Artificial neural networks (ANN) are complex and flexible nonlinear systems with properties not found in other modeling systems. It allows a method of forecasting with understanding of the relationship among variables and in particular nonlinear relationships. ANN function by initially learning a known set of data from a given problem with a known solution (training) and then the networks, inspired by the analytical processes of the human brain, are able to reconstruct the imprecise rules. Once a model is trained, the forecasted outputs can be generated from novel records [16-19].

The aim of this study is to investigate the use of a hybrid method combining grey model (GM) and back propagation artificial neural networks (BP-ANN) to forecast hepatitis B in China based on the yearly numbers of hepatitis B from the years 2002 to 2012 and to evaluate the method's performances of prediction.

2. Materials and Methods

2.1. Data Sources. The incidence data of hepatitis B are collected from the Ministry of Health of the People's Republic of China from the years 2002 to 2012, which are opening government statistics data [20].

2.2. Methods. The proposed method is established based on the grey systems theory and BP-ANN theory. MATLAB software version 2011b is used for the statistical analysis.

The incidence data are considered as the original time series $X = (x_0, x_1, x_2, ..., x_n)$, where *n* is the length of the time series.

Through grey generations or the effect of sequence operators to weaken the randomness, grey prediction models are designed to excavate the hidden laws; through the interchange between difference equations and differential equations, a practical jump of using discrete data sequences to establish continuous dynamic differential equations is materialized. Here, GM (1, 1) is the main and basic model of grey predictions, that is, a single variable first order grey model, which is able to acquire high prediction accuracy despite requiring small sample size (but the sample size must be at least 4). The GM (1, 1) model is suitable for sequences that show an obvious exponential pattern and can be used to describe monotonic changes. As for nonmonotonic wavelike development sequences, or saturated sigmoid sequences, one can consider establishing GM (2, 1) model.

The establishment for a GM (1,1) model is derived as follows.

(1) Let nonnegative time sequence expressing $X^0 = (x^0(1), x^0(2), \dots, x^0(n))$ be an original time sequence. Where *n* is the sample size of the data.

(2) First-order accumulative generation operation (1-AGO) is used to convert X^0 into $X^1 = (x^1(1), x^1(2), ..., x^1(n)) = (\sum_{i=1}^1 x^0(i), \sum_{i=1}^2 x^0(i), ..., \sum_{i=1}^n x^0(i))$ (3) Let $Z^1 = (z^1(2), z^1(3), ..., z^1(n))$ be the sequence generated from X^1 by adjacent neighbor means. That is,

 $z^{1}(t) = 0.5(x^{1}(t) + x^{1}(t-1)), t = 2, 3, ..., n$. The least square estimate sequence of the grey difference equation of GM(1, 1) is defined as $x^{0}(t) + az^{1}(t) = b$, where -a and b are referred to as the development coefficient and grey action quantity, respectively.

Then

$$[ab]^{T} = \left(B_{1}^{T}B_{1}\right)^{-1}B_{1}^{T}Y_{1}, \qquad (1)$$

where

$$Y_{1} = \begin{bmatrix} x^{0} (2) \\ x^{0} (3) \\ \vdots \\ x^{0} (n) \end{bmatrix},$$

$$B_{1} = \begin{bmatrix} -z^{1} (2) & 1 \\ -z^{1} (3) & 1 \\ \vdots & \vdots \\ -z^{1} (n) & 1 \end{bmatrix}.$$
(2)

(4) The whitenization equation is given by $dx^{1}/dt + ax^{1} =$ b.

(5) The forecasting model can be obtained by solving the above equation, which is shown as follows:

$$\hat{x}^{1}(t+1) = \left(x^{0}(1) - \frac{b}{a}\right)e^{-at} + \frac{b}{a}.$$
 (3)

(6) The predicted value of the primitive data at time point (t+1) is extracted:

$$\hat{x}^{0}(t+1) = \hat{x}^{1}(t+1) - \hat{x}^{1}(t) = (1 - e^{a}) \left(x^{0}(1) - \frac{b}{a} \right) e^{-at}.$$
(4)

The procedure for a GM (2, 1) model is derived as follows. (1) For a given sequence of original data $X^0 = (x^0(1),$ $x^{0}(2), \ldots, x^{0}(n)$, let its sequences of accumulation generation and inverse accumulation generation be $X^1 = (x^1(1),$ $x^{1}(2), \ldots, x^{1}(n)$ and $\alpha^{1} X^{0} = (\alpha^{1} x^{0}(1), \ldots, \alpha^{1} x^{0}(n))$, respectively, where $\alpha^1 x^0 = x^0(t) - x^0(t-1)$, t = 2, 3, ..., n and the sequence of adjacent neighbor mean generation of X^1 is $Z^1 = (z^1(2), z^1(3), \dots, z^1(n)).$

(2) The GM (2, 1) model is $\alpha^1 x^0(t) + a_1 x^0(t) + a_2 z^1(t) =$ u and the whitenization equation is given by d^2x^1/dt^2 + $\alpha_1(dx^1/dt) + \alpha_2 x^1 = u$. The least squares estimate of the parametric sequence is

$$\widehat{a} = [a_1, a_2, u]^T = (B_2^T B_2)^{-1} B_2^T Y_2,$$
 (5)

where

$$Y_{2} = \begin{bmatrix} \alpha^{1} x^{0} (2) \\ \alpha^{1} x^{0} (3) \\ \dots \\ \alpha^{1} x^{0} (n) \end{bmatrix} = \begin{bmatrix} x^{0} (2) - x^{0} (1) \\ x^{0} (3) - x^{0} (2) \\ \dots \\ x^{0} (n) - x^{0} (n-1) \end{bmatrix},$$

$$B_{2} = \begin{bmatrix} -x^{0} (2) -z^{1} (2) & 1 \\ -x^{0} (3) -z^{1} (3) & 1 \\ \dots & \dots \\ -x^{0} (n) -z^{1} (n) & 1 \end{bmatrix}.$$
(6)

(3) Solve the whitenization equation. If X^{1*} is a special solution of the whitenization equation and \overline{X}^1 the general solution of the corresponding homogeneous equation $d^2x^1/dt^2 + \alpha_1(dx^1/dt) + \alpha_2x^1 = 0$, then $X^{1*} + \overline{X}^1$ is the general solution of the GM (2, 1) whitenization equation. There are three cases for the general solution of the homogeneous equation: (i) when the characteristic equation $r^2 + \alpha_1 r$ + $\alpha_2 = 0$ has two distinct real roots $\overline{X}^1 = c_1 e^{r_1 t} + c_2 e^{r_2 t}$; (ii) when the characteristic equation has a repeated root \overline{X}^1 = $e^{rt}(c_1 + c_2 t)$; (iii) when the characteristic equation has two complex conjugate roots $r_1 = \alpha + i\beta$ and $r_2 = \alpha - i\beta$, $\overline{X}^1 = e^{\alpha t} (c_1 \cos \beta t + c_2 \sin \beta t)$. A special solution of the whitenization equation may take of the three possibilities: (i) when 0 is not a root of the characteristic equation, $X^{1*} = C$; (ii) when 0 is one of the two distinct roots of the characteristic equation, $X^{1*} = Cx$; and (iii) when 0 is the only root of the characteristic equation, $X^{1*} = Cx^2$.

The steps of the forecasting method can be described as follows.

Step 1 (train the BP-ANN). In order to obtain the input of the BP-ANN, the GM (1, 1) and GM (2, 1) model are used to predict for the original time series of hepatitis B, respectively. The two groups of prediction are taken as the input of the BP-ANN. At the same time, the original time series are taken as the output. Thus the structure of a three-layer BP-ANN is constructed and the trained BP-ANN model will be obtained by training.

Step 2 (forecast by the trained BP-ANN). The GM (1, 1) and GM (2, 1) model are used to forecast for the original time series of hepatitis B, respectively, at first. Then the two groups of forecasted data are taken as the input of the trained BP-ANN. Finally, the forecasted of hepatitis B will be obtained by running the trained BP-ANN.

The method flow chart is shown in Figure 1.

3. Evaluation Criteria

The metrics used are relative error (RE), RE_{*i*} = $|\hat{y}_i - y_i|/y_i$, *i* = 1, 2, ..., *n*, where \hat{y}_i is the forecasted data and y_i is the actual data; correlation coefficient (*R*), $R = R_{\hat{y}y}/\text{std}(y)\text{std}(\hat{y})$, where



FIGURE 1: Flow chart of the hybrid method.



FIGURE 2: The incidence number of hepatitis B in China from 2002 to 2012.

 $R_{\hat{y}y}$ is the covariance between y and \hat{y} ; mean square error (MSE), MSE = $(1/n)\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$; root mean square error (RMSE); RMSE = $\sqrt{(1/n)\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$; mean average error (MAE), MAE = $(1/n)\sum_{i=1}^{n}|\hat{y}_i - y_i|$; mean average percentage error (MAPE), MAPE = $(1/n)\sum_{i=1}^{n}|\hat{y}_i - y_i| + 100$, and sum of squared error (SSE), SSE = $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$.

4. Result

The incidence data of hepatitis B are collected year by year from 2002 to 2012 in China and taken as the original time series, which is shown in Figure 2.

4.1. The GM (1, 1) and GM (2, 1) Models Calculation. The GM (1, 1) and GM (2, 1) models are calculated and shown as follows.



FIGURE 3: The topology structure of the proposal method.

4.1.1. The Parameters

GM (1, 1) *Model.* Consider the following: -a = 0.027523, b = 893075.402372. Therefore, the GM (1, 1) model of this time series can be forecasted for long term forecasting for the reasons that GM (1, 1) can be used for long term forecasting when $-a \le 0.3$ and for short term forecasting when $0.3 < -a \le -0.5$. -a reflects the development states of the accumulation generated sequence $\hat{x}^{(1)}$ and the sequence of raw data $\hat{x}^{(0)}$.

GM (2, 1) *Model.* Consider the following: $a_1 = 0.4083$, $a_2 = 0.0158$, u = 583425.10336, $r_1 = -0.04329$, $r_2 = -0.36502$, $c_1 = -38764735.45099$, and $c_2 = 2527585.93078$.

4.1.2. The Forecasting Models

GM (1, 1) *Model*. Consider the following:

$$x(t+1) = -33116202.802344e^{0.027523t} + 32447876.802344.$$
(7)

GM (2, 1) *Model*. Consider the following:

$$x^{1}(t+1) = -38764735.45099e^{-0.04329t} + 2527585.93078e^{-0.36502t}$$
(8)
+ 36918146.77020.

4.2. The Forecasted. In the three-layer BP-ANN, the hidden node n_2 and the input node n_1 are related by $n_2 = 2n_1 + 1$. The two groups of prediction created by the two GM models are taken as the input of the BP-ANN and the observed data is taken as the output. Therefore, a three-layer proposed model with 2 input nodes, 5 hidden nodes, and 1 output node is obtained. The topology structure is shown in Figure 3.

The weights and thresholds of the proposal model will be obtained by training. Let the training time be 1000, the learning rate be 0.9, the momentum factor be 0.95, and the error be 0.001; Levenberg Marquardt is used as training algorithm.



FIGURE 4: The forecasted incidence of hepatitis B in China from 2013 to 2021 by the proposal method.



FIGURE 5: The scatter diagram of the relationship between the observed data and the prediction.

The prediction of the original time series by the GM (1, 1) and GM (2, 1) model, respectively, are taken as the input of the trained BP-ANN. Then the forecasted of hepatitis B will be obtained by running the trained BP-ANN. The forecasted is shown in Figure 4.

5. Discussion

In order to compare the prediction created by the two GM models and the proposed method, a prediction is performed under the same conditions. The results are listed in Table 1 and the scatter diagram is shown in Figure 5. It can be seen from Figure 5 that the prediction generated by the two GM models has greater dispersion than that by the proposed method.

The RE of prediction is shown in Figure 6. From the figure, we know that the prediction obtained by the proposed method has higher accuracy and smaller RE than that by the GM approaches. Figure 6 indicates that the smaller the relative error is, the closer prediction is to the observed data.

\mathbf{T}	TABLE 1: The	prediction created	ov the GM	(1, 1), GM	(2, 1), and the	proposal model.
--------------	--------------	--------------------	-----------	------------	-----------------	-----------------

Year	The observed data	GM (1, 1)	GM (2, 1)	The proposal method
2003	719011	924130	882183	736600
2004	916396	949918	1036306	884300
2005	982297	976426	1133758	1027700
2006	1109130	1003674	1183850	1096300
2007	1169946	1031682	1201811	1118200
2008	1169569	1060471	1198178	1124300
2009	1179607	1090065	1180237	1125900
2010	1060582	1120484	1153018	1126400
2011	1093335	1151751	1119982	1126500
2012	1087086	1183892	1083509	1126500

TABLE 2: The evaluation indexes comparison.

Index	R	MSE	MAE	RMSE	MAPE	SSE
The proposal method	0.9495	2.3649×10^{7}	$3.9704 imes 10^4$	4.863×10^{3}	3.9704×10^{6}	1.8162×10^{10}
GM (1, 1) model	0.6365	2.2867×10^8	1.0492×10^{6}	1.5122×10^4	1.0492×10^8	1.1078×10^{13}
GM (2, 1) model	0.9392	1.6798×10^{8}	1.1173×10^{6}	1.5122×10^4	1.1173×10^{8}	1.2570×10^{13}

TABLE 3: The forecasted generated by the three methods.

Year	GM (1, 1)	GM (2, 1)	The proposal method
2013	1216929.0	1045223.6	1077864.1
2014	1250888.2	1006228.6	1074038.2
2015	1285795.1	967267.6	1012371.7
2016	1321676.0	928834.0	976301.8
2017	1358558.3	891248.9	946959.7
2018	1396469.7	854715.0	939194.1
2019	1435439.1	819353.3	937881.5
2020	1475496.0	785229.0	937607.7
2021	1516670.7	752369.4	937531.5



The comparison of *R*, MSE, MAE, RMSE, MAPE, and SSE are listed in Table 2. It can be seen that the proposed method has advantages over GMs in all the evaluation indexes.

The forecasted generated by the GM (1, 1), GM (2, 1), and the proposal model are listed in Table 3 and shown in Figure 7.

The weights and thresholds of BP-ANN will generate randomly at first when the model is training. This will make the predicted and forecasted uncertainty. To describe this clearer, the proposal model is ran 100 times and the mean value will be taken as predicted or forecasted value. The 95% confidence interval and predicted or forecasted value are shown in Figures 8 and 9, respectively.

Although the prediction result created by the proposal method in the paper has more accurate than that by the two gray models, the proposal model has its limitations. Firstly, since the proposal model is built on the basis of gray model, the sample size, namely, the number of historical data must be not less than 4. Secondly, the prediction result

FIGURE 6: Comparison of the RE of the prediction by the proposal method and the GMs.

will be inaccuracy if the weights and thresholds in BP-ANN ran into local optimum in the process of training. Intelligent algorithms can be used to optimize the weights and thresholds of BP-ANN [21].

6. Conclusion

The hepatitis B epidemiological information is often difficult to obtain. Forecasting of hepatitis B will be inaccurate by the limited data. The grey systems theory focuses on uncertainty problems with small samples and incomplete information. At the same time, the BP-ANN is a method of forecasting with understanding of the relationship among variables and nonlinear relationships. The research proposes a new forecasting method, which combines the GM and BP-ANN, to forecast hepatitis B in China. The useful information can generate and



FIGURE 7: The forecasted incidence of hepatitis B in China from 2013 to 2021 by the three methods.



FIGURE 8: The predicted incidence of hepatitis B in China from 2003 to 2012 by the proposal methods.



FIGURE 9: The forecasted incidence of hepatitis B in China from 2013 to 2021 by the proposal methods.

extract from the small samples and the BP neural networks can train data more sufficiently. The prediction results show that this method can obtain better forecasting.

Conflict of Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors thank Zola Banh for carefully correcting grammar errors. This work was supported by Youth Science Foundation of Guangxi Medical University (GXMUYSF201208), the open project of Guangxi Medical Science Experimental Center (KFJJ2011-31), the training programs of innovation and entrepreneurship for undergraduates in Guangxi province (2012xjcxcy006), and the training programs of innovation and entrepreneurship for undergraduates in China (201210598002). The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the paper.

References

- S. A. Gourley, Y. Kuang, and J. D. Nagy, "Dynamics of a delay differential equation model of hepatitis B virus infection," *Journal of Biological Dynamics*, vol. 2, no. 2, pp. 140–153, 2008.
- [2] F. Liu, L. Chen, D.-M. Yu et al., "Evolutionary patterns of hepatitis B virus quasispecies under different selective pressures: correlation with antiviral efficacy," *Gut*, vol. 60, no. 9, pp. 1269– 1277, 2011.
- [3] J. M. Torpy, A. E. Burke, and R. M. Golub, "JAMA patient page. Hepatitis B," *Journal of the American Medical Association*, vol. 305, no. 14, article 1500, 2011.
- [4] S. C. Nwokediuko, "Chronic hepatitis B: management challenges in resource-poor countries," *Hepatitis Monthly*, vol. 11, no. 10, pp. 786–793, 2011.
- [5] "The guideline of prevention and treatment for chronic hepatitis B (2010 version)," *Chinese Journal of Hepatology*, vol. 19, no. 1, pp. 13–24, 2011.
- [6] T. Zhang and Z. Teng, "On a nonautonomous SEIRS model in epidemiology," *Bulletin of Mathematical Biology*, vol. 69, no. 8, pp. 2537–2559, 2007.
- [7] M. Tizzoni, P. Bajardi, C. Poletto et al., "Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm," *BMC Medicine*, vol. 10, article 165, 2012.
- [8] S. Eubank, H. Guclu, V. S. A. Kumar et al., "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, pp. 180–184, 2004.
- [9] H. Ren, J. Li, Z.-A. Yuan, J.-Y. Hu, Y. Yu, and Y.-H. Lu, "The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China," *BMC Infectious Diseases*, vol. 13, no. 1, article 421, 2013.
- [10] Y. Bai and Z. Jin, "Prediction of SARS epidemic by BP neural networks with online prediction strategy," *Chaos, Solitons and Fractals*, vol. 26, no. 2, pp. 559–569, 2005.
- [11] M. Ture and I. Kurt, "Comparison of four different time series methods to forecast hepatitis A virus infection," *Expert Systems with Applications*, vol. 31, no. 1, pp. 41–46, 2006.

- [12] S. Thornley, C. Bullen, and M. Roberts, "Hepatitis B in a high prevalence New Zealand population: a mathematical model applied to infection control policy," *Journal of Theoretical Biology*, vol. 254, no. 3, pp. 599–603, 2008.
- [13] X. Shen, L. Ou, X. Chen, X. Zhang, and X. Tan, "The application of the grey disaster model to forecast epidemic peaks of typhoid and paratyphoid fever in China," *PLoS ONE*, vol. 8, no. 4, Article ID e60601, 2013.
- [14] J. L. Deng, "Introduction to grey system theory," *The Journal of Grey System*, vol. 1, no. 1, pp. 1–24, 1989.
- [15] S. F. Liu and Y. Lin, *Grey Systems: Theory and Applications*, Springer, Berlin, Germany, 2010.
- [16] Y. Amaritsakul, C.-K. Chao, and J. Lin, "Multiobjective optimization design of spinal pedicle screws using neural networks and genetic algorithm: mathematical models and mechanical validation," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 462875, 9 pages, 2013.
- [17] J. Zou, Y. Han, and S. S. So, "Overview of artificial neural networks," *Methods in Molecular Biology*, vol. 458, pp. 15–23, 2008.
- [18] N. Terrin, C. H. Schmid, J. L. Griffith, R. B. D'Agostino, and H. P. Selker, "External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks," *Journal of Clinical Epidemiology*, vol. 56, no. 8, pp. 721–729, 2003.
- [19] E. Grossi and M. Buscema, "Introduction to artificial neural networks," *European Journal of Gastroenterology and Hepatol*ogy, vol. 19, no. 12, pp. 1046–1054, 2007.
- [20] The Science Data Center of Public Health of China, http://www .phsciencedata.cn/Share/ky_sjml.jsp?id=8defcfc2-b9a4-4225b92c-ebb002321cea&show=0.
- [21] D. Huang, R. Gong, and S. Gong, "Prediction of wind power by chaos and BP artificial neural networks approach based on genetic algorithm," *Journal of Electrical Engineering & Technology*, vol. 10, no. 1, pp. 41–46, 2015.