· 调查报告与分析 ·

中国手足口病发病率 ARIMA、RBF 及 ARIMA-RBF 组合模型拟合及预测效果比较*

王永斌,许春杰,尹素凤,武建辉,柴峰,李向文,袁聚祥

摘　要:目的　比较求和自回归滑动平均混合模型(ARIMA)、径向基函数神经网络模型(RBF)和 ARIMA-RBF 组合模型对中国手足口病月发病率的拟合及预测效果,探讨预测手足口病发病率的优化模型。**方法**　收集 2008 年 1 月—2014 年 12 月中国手足口病月发病率资料,基于 2008 年 1 月—2014 年 6 月的 78 个数据分别建立 ARIMA 模型、RBF 模型和 ARIMA-RBF 组合模型,采用 2014 年 7—12 月的 6 个数据进行外回代验证模型的外推预测效果,评价指标包括相对误差(*RE*)、平均相对误差(*MRE*)、均方误差(*MSE*),均方根误差(*RMSE*)和平均绝对误差(*MAE*)。**结果**　ARIMA 模型拟合和预测的 *MRE*、*MSE*、*RMSE*、*MAE* 分别为 14.006、4.689、2.165、0.916 和 13.565、4.416、2.101、0.577,RBF 模型分别为 9.031、1.559、1.249、0.508 和 8.964、1.504、1.226、0.503,ARIMA-RBF 组合模型分别为 6.397、1.357、1.165、0.416 和 6.655、1.485、1.218、0.433,ARIMA-RBF 组合模型的拟合及预测曲线与原始值最接近。**结论**　ARIMA-RBF 组合模型拟合及预测效果均优于 ARIMA 模型和 RBF 模型。

关键词:手足口病;发病率;求和自回归滑动平均混合模型(ARIMA);径向基函数神经网络模型(RBF);ARIMA-RBF 组合模型

中图分类号:R 195.1　**文献标志码:**A　**文章编号:**1001-0580(2017)05-0760-04　DOI:10.11847/zgggws2017-33-05-19

Comparison on predictive capacity and fitting efficiency of ARIMA, RBF and ARIMA-RBF combination model for incidence of hand-foot-mouth disease

WANG Yong-bin, XU Chun-jie, YIN Su-feng, et al (*School of Public Health, North China University of Science and Technology, Tangshan, Hebei Province* 063000, China)

Abstract: Objective　To compare predictive capacity and fitting efficiency of autoregressive integrated moving average model (ARIMA), radical basis function model (RBF), and ARIMA-RBF combined model for monthly incidence of hand-foot-mouth disease (HFMD) in China and to explore an optimized model. **Methods**　Data on monthly incidence of HFMD from January 2008 to December 2014 in China were collected and ARIMA, RBF, and ARIMA-RBF combined models were established based on the data; then the data for the period from July to December 2014 were introduced back to the established models to evaluate their predictive capacity. The indexes of fitting efficiency of the models included relative error (RE), mean relative error (MRE), mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE). **Results**　The MRE, MSE, RMSE, and MAE were 14.006, 4.689, 2.165, and 0.916 for ARIMA fitting model; 13.565, 4.416, 2.101, and 0.577 for ARIMA predictive model; 9.031, 1.559, 1.249, and 0.508 for RBF fitting model; 8.964, 1.504, 1.226, and 0.503 for RBF predictive model; 6.397, 1.357, 1.165, and 0.416 for ARIMA-RBF combined fitting model; and 6.655, 1.485, 1.218, and 0.433 for ARIMA-RBF combined predictive model, respectively. For the three established models, the fitting and predictive values of ARIMA-RBF combined model were closest to the original data. **Conclusion**　The ARIMA-RBF combination model is superior to ARIMA model and RBF model in fitting and prediction of HFMD monthly incidence in China.

Key words: hand-foot-mouth disease; incidence; autoregressive integrated moving average model; radical basis function model; ARIMA-RBF combined model

手足口病是由多种人肠道病毒引起的一种儿童常见传染病,2008 年卫生部列为丙类传染病,其常出现暴发或流行[1]。为此了解其流行规律及流行趋势,对于探索手足口病的病因线索和疾病预防策略及措施的制定具有至关重要的作用。由于手足口病月发病率序列具有季节性特点,采用季节性自回归滑动平均混合模型(autoregressive integrated moving average model, ARIMA)预测手足口病发病是目前最常用的方法之一。ARIMA 模型通过对每一个季节周期中相同时间点的序列值进行分析,可以提取其中的季节趋势,并针对每个季节周期内部序列的变化可以提取其中的非季节性成分[2]。但 ARIMA 模型作为线性模型对数据序列的非线性趋势提取效果较差,而中国手足口病的发病率序列既

* **基金项目:**河北省卫生厅医学科学研究重点课题(20130055)
作者单位:华北理工大学公共卫生学院,河北 唐山 063000
作者简介:王永斌(1989 -),男,甘肃正宁人,博士在读,研究方向:人群传染病防治。
通讯作者:袁聚祥,E-mail:yuanjx@heuu.edu.cn
数字出版日期:2016 - 5 - 19 11:02
数字出版网址:http://www.cnki.net/kcms/detail/21.1234.R.20160519.1102.026.html

有线性趋势又有非线性趋势。因此,本研究基于中国 2008 年 1 月 1 日—2014 年 12 月 31 日手足口病发病序列,将线性 ARIMA 模型与非线性径向基函数神经网络模型(radical basis function, RBF)进行组合来弥补 ARIMA 模型非线性映射性能不足的弊端[3],并比较其与 ARIMA 模型和 RBF 模型对手足口病月发病率预测的准确性,以探讨手足口病预测预警的优化模型。

1　资料与方法

1.1　资料来源　资料来源于 2008 年 1 月 1 日—2014 年 12 月 31 日中国疾病预防控制中心传染病直报系统中各地区上报的手足口病发病序列资料,人口资料来源于中国卫生统计局。

1.2　统计分析　应用 SPSS 13.0 软件构建 ARIMA 模型,通过单位根检验来确定序列是否为平稳序列,对于非平稳序列通过对数转换和差分的方式来消除季节和趋势的影响,从而获得平稳的时间序列[2];依据赤池信息准则(Akaike information criterion, AIC)和 Schwarz 贝叶斯准则(Schwarz Bayesian criterion, SBC)最小,对数似然函数值(log likelihood)最大的模型为最优 ARIMA 模型,模型诊断时残差序列应是白噪声,其自相关系数(autocorrelation coefficient, ACF)和偏自相关系数(partial autocorrelation coefficient, PACF)不应与 0 有显著的差异,Box-Ljung Q 统计量应无统计学意义,且模型各系数估计值均应有统计学意义[3]。ARIMA 模型公式为:$\square(B)\Phi(B^s)\nabla^d\ \nabla_S{}^D X_t = \theta(B)\Theta(B^s)\varepsilon_t$;$E(\varepsilon_t)=0, var(\varepsilon_t)=\sigma^2, E(\varepsilon_t\varepsilon_s)=0, s\neq t\ E(x_t\varepsilon_t)=0, \vee s<t$;$\nabla^d=(1-B)^d$;$\nabla_S{}^D=(1-B)^{SD}$;$\square(B)=1-\square_1 B-\cdots\square_p B^p$;$\theta(B)=1-\theta_1 B-\cdots\theta_q B^q$;$\Phi(B^s)=1-\Phi_1 B^s-\cdots\Phi_P B^{Ps}$;$\Theta(B^s)=1-\Theta_1 B^s-\cdots\Theta_Q B^{Qs}$。式中,B 为后移算子,$\varepsilon_t$ 为残差,d 和 D 分别为非季节和季节差分次数,p 和 q 分别为自回归和滑动平均阶数,P 和 Q 分别为季节自回归和滑动平均阶数[4-5]。应用 Matlab 8.0 软件建立 RBF 模型和 ARIMA-RBF 组合模型时,原始数据序列使用峰值法进行归一化处理。在 RBF 模型训练过程中,隐含层神经元个数和光滑因子 spread 的确定是 1 个关键问题。本研究中利用函数 newrbe 创建一个精确的神经网络,该函数在创建 RBF 模型时,自动选择隐含层的数目,使得误差为 0[6]。光滑因子 spread 的确定方法是在用于学习的样本数据中选取 6 个月的数据作为待估点,将剩余的学习样本数据用来训练神经网络,用训练好的神经网络对这 6 个待估点进行预测,预测值与实际值的网络逼近误差和均方误差的值最小时对应的光滑因子为最佳 spread。通过手足口病发病序列建立 ARIMA 模型,由于 ARIMA 模型对手足口病发病序列的拟合值与实际值存在高度的关联性可以将 ARIMA 模型的拟合值作为 RBF 模型的输入,实际值作为 RBF 模型的输出,通过调整光滑因子 spread 构建 1 个一维输入、一维输出的优化 RBF 模型。基于 2008 年 1 月—2014 年 6 月的 78 个数据建立相关模型,用 2014 年 7—12 月的 6 个数据验证模型的外推预测效果,评价指标包括平均相对误差(average relative error, *MRE*)、相对误差(relative error, *RE*)、均方误差(mean square predict error, *MSE*)、均方根误差(root mean squared predict error, *RMSE*)和平均绝对误差(mean absolute error, *MAE*)[2]。

2　结　果

2.1　ARIMA 模型建立及预测(图 1)　单位根检验结果显示,中国手足口病发病率序列为非平稳时间序列数据。因此首先对原始序列进行对数转换和 D = 1 的季节差分、d = 1 的非季节差分来消除季节和趋势的影响,从而使序列转化为平稳序列。再结合平稳序列的自我相关函数和偏自我相关函数图、残差序列的自我相关函数和偏自我相关函数图以及相关系数之间的关联性选取 log likelihood 值最大,AIC 和 SBC 值最小,且模型各系数估计值均有统计学意义的模型为最优模型,通过反复比较,得到最优的模型是 ARIMA(2,1,1) × $(0,1,1)_{12}$,具体方程为 $(1-B)(1-B^{12})X_t=(1-0.637B)(1-0.448B^{12})\varepsilon_t/(1-0.891B+0.501B^2)$,其残差的自我相关函数和偏自我相关函数见图 1。季节性滑动平均系数为 0.448($t=3.492$, $P=0.001$),非季节性滑动平均系数为 0.637($t=3.652$, $P<0.001$),非季节性自回归系数为 0.891($t=6.060$, $P<0.001$),AR2 = −0.501($t=-4.626$, $P<0.001$),AIC = 64.078,

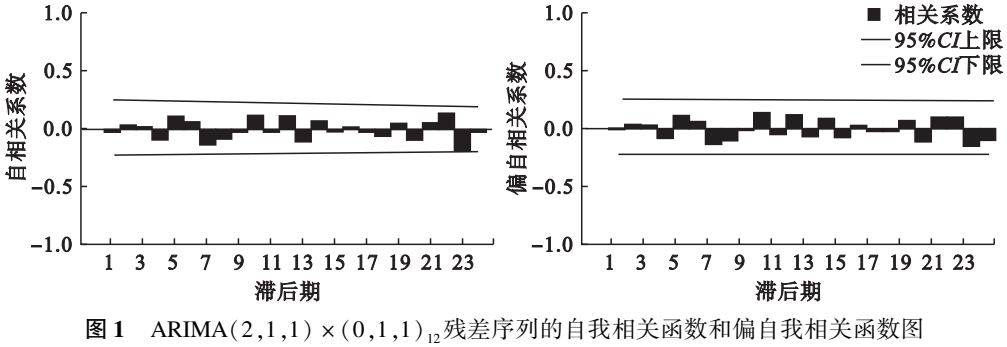

图 1　ARIMA(2,1,1) × $(0,1,1)_{12}$ 残差序列的自我相关函数和偏自我相关函数图

SBC = 75.391, log likelihood = −27.039。且残差序列的 Box-Ljung Q 统计量差异均无统计学意义(均 $P > 0.05$),可认为残差序列是白噪声。

2.2　RBF 模型建立及预测(图 2)　由于手足口病发病率变化的周期为年,因此本研究中 RBF 模型输入层和输出层神经元个数分别为 12 和 1。光滑因子 spread 确定时,以 2008 年 1 月—2013 年 12 月作为网络的训练样本,2014 年 1—6 月的数据作为网络的测试样本,经多次尝试后,最终确定从 1 开始,每次增加 1 直至 5,分别得到测试样本的预测值。计算预测值和实际值的网络预测误差、逼近误差和 *MSE*,当它们的值最小时,对应的光滑因子为最优值。本研究中不同光滑因子对应的网络预测误差见图 2,网络逼近误差分别为 0.861 6、3.595 1、0.694 6、0.729 6、0.756 9,*MSE* 分别为 0.123 7、2.154 1、0.080 4、0.088 7、0.095 5。提示当光滑因子为 3 时,对应的网络预测误差、逼近误差和 *MSE* 的值最小,所以最终确定的光滑因子为 3。用确定好的光滑因子构建 RBF 模型,然后对 2014 年 7—12 月数据序列进行预测,将预测值进行反归一化后得到原始序列的预测值。

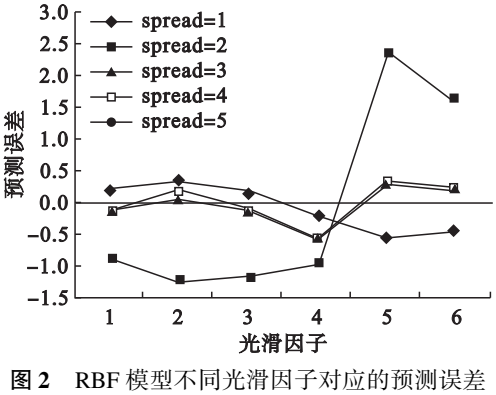

图 2　RBF 模型不同光滑因子对应的预测误差

2.3　ARIMA-RBF 组合模型建立及预测(图 3)　构建 ARIMA-RBF 组合模型时,将 ARIMA 的拟合及预测值与原始数据序列采用峰值法进行归一化处理。由于对手足口病原始序列进行差分导致 13 个数据损失,因此以 2009 年 2 月—2014 年 6 月 ARIMA 模型的拟合值作为 RBF 模型的输入值,以对应月份的手足口病实际发病率作为RBF模型的输出值,构建一维输入、一维输出的 ARIMA-RBF 组合模型。光滑因子 spread 确定时,以 2009 年 2 月—2013 年 12 月作为网络的训练样本,2014 年 1—6 月的数据作为网络的测试样本。经多次尝试后,最终确定从 1 开始,每次增加 1 直至 5,最终得到网络逼近误差分别为 16.635 8、11.455 7、11.746 7、0.777 2、0.780 4,*MSE* 分别为 29.639 2、21.872 3、22.997 6、0.100 7、0.101 5,不同光滑因子对应的网络预测误差见图 3。提示当光滑因子为 4 时,对应的网络预测误差、逼近误差和 *MSE* 的值最小,所以最终确定组合模型的光滑因子为 4。用确定好的光滑因子构建 ARIMA-RBF 组合模型,然后以 2014 年 7—12 月 ARIMA 模型的预测值为输入值,用组合模型对 2014 年 7—12 月的数据进行预测,将预测值进行反归一化后得到原始序列的预测值。

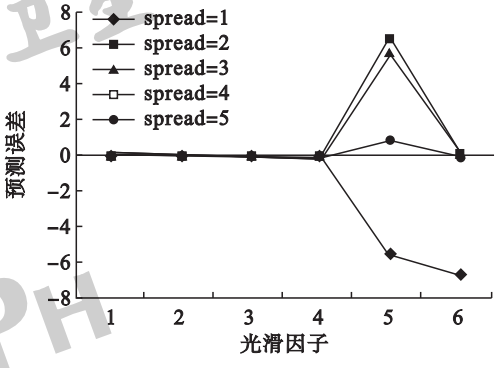

图 3　ARIMA-RBF 组合模型不同光滑因子对应的预测误差

2.4　3 种模型拟合及预测效果比较(表 1、图 4)　ARIMA 模型、RBF 模型和ARIMA-RBF组合模型 3 种模型对中国 2014 年 7—12 月手足口病月发病率的预测结果见表 1。ARIMA 模型拟合和预测的 *MRE*、*MSE*、*RMSE*、*MAE* 分别为 14.006、4.689、2.165、0.916 和 13.565、4.416、2.101、0.577,RBF 模型分别为 9.031、1.559、1.249、0.508 和 8.964、1.504、1.226、0.503,ARIMA-RBF 组合模型分别为 6.397、1.357、1.165、0.416 和 6.655、1.485、1.218、0.433,ARIMA-RBF 组合模型的拟合及预测性能均优于 ARIMA 模型和 RBF 模型。图 4 为 3 种模型对中国手足口病发病率的拟合及预测曲线,可见 ARIMA-RBF 组合模型的拟合及预测曲线与原始值最接近。

表 1　中国 2014 年 7—12 月手足口病发病率(1/10 万)3 种模型预测外回代验证值

月份	实际值	ARIMA 模型		RBF 模型		ARIMA-RBF 组合模型	
		预测值	*RE*(%)	预测值	*RE*(%)	预测值	*RE*(%)
7 月	23.503	25.039	6.536	24.772	5.399	25.379	7.982
8 月	10.502	11.792	12.284	11.751	11.893	11.492	9.427
9 月	14.292	12.279	14.088	14.984	4.842	13.933	2.512
10 月	15.577	13.607	13.644	17.135	10.002	13.652	12.358
11 月	11.010	12.094	9.850	12.642	14.823	11.765	6.857
12 月	7.798	7.733	0.827	7.266	6.822	7.860	0.795

中国公共卫生 CPH

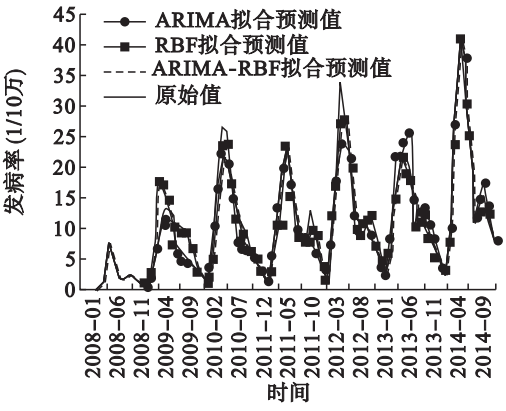

图 4　中国手足口病发病率 3 种模型对拟合及预测值对比图

3　讨　论

近年来在医学和公共卫生领域,多种统计预测方法已被广泛应用于疾病发病率、患者预后估计、人口预测、卫生人力资源规划以及各种卫生经费的预算[7]。传染病的发病往往受到多种因素的影响和制约,应用传统的线性模型对其预测存在较大的局限性[8]。而神经网络所具有的非线性映射能力和高度的并行分布处理方式为预测这类非线性动态关系的时间序列提供了一条思路。但单一的线性模型和非线性模型皆存在信息源不广泛问题,在应用中常常受到限制或达不到高精度的要求[9],而组合预测就是将不同的单一预测模型按一定方式进行组合,综合利用各种方法所能够提供的信息取长补短,具有很强的优势互补性,从而达到提高预测精度和增加预测稳定性的效果。RBF 模型具有结构简单、局部响应、快速的训练过程及与初始权重无关的优良特性[6]。因此本研究建立具有 ARIMA 模型和 RBF 模型优点的 ARIMA-RBF 组合模型对我国手足口病发病率进行预测。研究结果表明,ARIMA-RBF 组合模型拟合及预测效果优于 ARIMA 模型和 RBF 模型。有文献报道,模型拟合及预测的 *MRE*≤10%时,用于预测时准确度和精度较高,其中 *MRE*≤5%时为理想状态[4]。本研究中,ARIMA-RBF 组合模型拟合和预测的 *MRE* 均≤10%,且 ARIMA-RBF 组合模型拟合及预测值与实际值的走向有着极为相似的升降规律,提示 ARIMA-RBF 组合模型更能较好地反映我国手足口病发病率序列的内部规律和未来趋势,是可以推广应用的。但本研究中 ARIMA-RBF 组合模型拟合及预测效果还未达到理想状态,且目前已有相关研究将 ARIMA-BP 组合模型[3]、ARIMA-ARCHA 组合模型[8]、ARIMA-GRNN 组合模型[9]和 ARIMA-NARNN 组合模型[10]应用于预测传染病发病率,研究结果表明它们的拟合及预测效果优于 ARIMA 模型,且 ARIMA-NARNN 组合模型拟合及预测性能优于 ARIMA-GRNN 组合模型。因此,下一步可考虑将 ARIMA-RBF 组合模型的拟合及预测效果同上述模型进行比较,寻找预测手足口病发病的最佳组合模型。

尽管本研究中 ARIMA 模型和 RBF 模型拟合及预测精度均低于 ARIMA-RBF 组合模型,但就 ARIMA 模型和 RBF 模型拟合及预测的 *MRE* 来看,其仍然可以用于预测我国手足口病发病率。需注意的是 RBF 模型和 ARIMA-RBF 组合模型建模过中对光滑因子 spread 的确定是关键,spread 的值越大,网络对样本的拟合就越平滑,spread 的值越小,网络能更精确地对数据进行拟合。本研究中在学习样本中选取 6 个月的数据,通过反复训练,综合网络预测误差、逼近误差和 *MSE* 来确定最优的 spread,从而使模型的拟合及预测性能达到平衡。

综上所述,ARIMA-RBF 组合模型在一定程度上具有较高的预测精度与准确度。因此可借助 ARIMA-RBF 组合模型对我国手足口病发病率进行早期预测、预警,为未来制定相应的政策提供参考依据,从而减少决策的盲目性。但在实际工作中应用组合模型时需注意:(1)在满足模型使用条件的情况下,根据数据序列的特征,选择合适的单项模型进行组合预测,否则模型的效能不但得不到提高,反而会提供一些不正确的信息;(2)传染病发病是多因素综合作用的结果,在条件允许的情况下,尽可能多得收集影响传染病发生的因素来构建组合模型;(3)单次分析构建的组合模型,只适用于短期进行预测。在实际应用过程中,应不断收集新的数据序列对已建立的模型进行外回代验证[11],以建立更能反映时间序列的内部规律和未来趋势的组合模型。

参考文献

[1] Zhang W, Huang B, She C, et al. An epidemic analysis of hand, foot, and mouth disease in Zunyi, China between 2012 and 2014[J]. Saudi Med J, 2015, 36(5): 593 – 598.
[2] Zhang X, Liu Y, Yang M, et al. Comparative study of four time series methods in forecasting typhoid fever incidence in China[J]. PLoS One, 2013, 8(5): e63116.
[3] Zhang G, Huang S, Duan Q, et al. Application of a hybrid model for predicting the incidence of tuberculosis in Hubei, China[J]. PLoS One, 2013, 8(11): e80969.
[4] 王永斌,郑瑶,柴峰,等. 基于周期分解的 ARIMA 模型在甲肝发病率预测中的应用[J]. 现代预防医学, 2015, 42(23): 4225 – 4229.
[5] 梅树江,周志峰,马汉武,等. 深圳市 ARIMA 在肾综合征出血热发病预测中应用[J]. 中国公共卫生, 2015, 31(7): 936 – 938.
[6] 葛哲学. 神经网络理论与 MATLABR2007 实现[M]. 北京:电子工业出版社, 2007: 111 – 147.
[7] 刘桂芬. 医学统计学[M]. 2 版. 北京:中国协和医科大学出版社, 2009: 346 – 365.
[8] Yan W, Xu Y, Yang X, et al. A hybrid model for short-term bacillary dysentery prediction in Yichang city, China[J]. Jpn J Infect Dis, 2010, 63(4): 264 – 270.
[9] Zheng YL, Zhang LP, Zhang XL, et al. Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China[J]. PLoS One, 2015, 10(3): e0116832.
[10] Wu W, Guo J, An S, et al. Comparison of two hybrid models for forecasting the incidence of hemorrhagic fever with renal syndrome in Jiangsu province, China[J]. PLoS One, 2015, 10(8): e0135492.
[11] Liu L, Luan RS, Yin F, et al. Predicting the incidence of hand, foot and mouth disease in Sichuan province, China using the ARIMA model[J]. Epidemiol Infect, 2015, 12(3): 1 – 8.

收稿日期: 2015-11-23　　(郭薇编校)