

• 论著 •

ARIMA 模型的建立及对中国肺结核月报告例数的预测效果研究

张顺先 邱磊 张少言 李翠 胡骏 田黎明 鹿振辉

【摘要】 目的 建立自回归移动平均 (autoregressive integrated moving average, ARIMA) 模型, 并对全国 (不包括我国港澳台地区, 下同) 肺结核月报告患者例数进行预测效果研究, 为肺结核防控措施的制定提供科学参考。**方法** 通过中国疾病预防控制中心主办的《疾病监测》杂志公布的我国每月甲、乙、丙类传染病疫情动态简介, 搜集 2006 年 1 月至 2019 年 8 月全国肺结核月报告患者例数。采用 SPSS 26.0 统计学软件, 以 2006 年 1 月至 2018 年 12 月的全国肺结核月报告患者例数为基础建立时间序列, 初步识别和定阶 ARIMA 模型类型; 再以满足模型简洁、ARIMA 模型各参数 [包括自回归法 (AR), 平均移动法 (MA), 季节自回归法 (SAR), 季节移动平均法 (SMA)] 均有统计学意义 (P 值均 < 0.05), 以及 $P > 0.05$ 的模型总体检验指标 (Ljung-Box Q 值)、最大平稳决定系数 (R^2)、最小整体模型的标准化贝叶斯信息准则值 (NBIC)、最小均方根误差 (RMSE) 为标准筛选几种 ARIMA 模型; 继而以 2019 年 1—8 月报告患者例数作为验证数据, 参照预测值相对误差越小模型越优的原则筛选出最小相对误差的模型为最优模型; 最后再以该模型预测我国 2019 年 9 月至 2020 年 12 月肺结核月报告患者例数。**结果** 根据 2006—2018 年每年的全国肺结核月报告患者例数为基础建立时间序列, 确定需拟合 ARIMA(p, d, q) 或 ARIMA(p, d, q) \times (P, D, Q) 模型。以 Ljung-Box Q 值所对应的 P 值均 > 0.05 、模型简洁、模型各参数均有统计学意义 (P 值均 < 0.05) 筛选出 12 个基本模型, 然后再以 R^2 最大的模型 [ARIMA(1, 0, 1)(0, 1, 1)₁₂, $R^2 = 0.707$]、RMSE 最小的模型 [ARIMA(0, 1, 2)(0, 1, 1)₁₂, RMSE = 9147.85]、NBIC 最小的模型 [ARIMA(0, 1, 1)(0, 1, 1)₁₂, NBIC = 18.355]、Ljung-Box Q 值最小的模型 [ARIMA(1, 1, 1)(0, 1, 1)₁₂, Ljung-Box $Q = 8.797$] 作为备用模型, 预测 2019 年 1—8 月中国肺结核月报告患者例数, 并与实际的月报告患者例数进行比较, 确定预测平均相对误差最小 (0.55%)、MA(1) = 0.875 ($t = 19.243$, $P < 0.001$)、SMA(1) = 0.876 ($t = 7.596$, $P < 0.001$)、Ljung-Box $Q = 9.876$ ($df = 16$, $P = 0.873$) 的 ARIMA(0, 1, 1)(0, 1, 1)₁₂ 模型为最优模型。再以该模型预测我国 2019 年 9 月至 2020 年 12 月肺结核月报告患者例数, 其中 2020 年 1—12 月患者总计 1 025 863 例, 平均每月 85 489 例。**结论** ARIMA(0, 1, 1)(0, 1, 1)₁₂ 模型对预测中国肺结核月报告患者例数方面效果较好, 但应注意模型的建立和预测是个动态变化过程, 需不断根据积累的数据进行调整, 从而提高预测精度。

【关键词】 结核, 肺; 疾病报告; 流行病学研究设计; 模型, 统计学; 预测

A study of prediction effect of autoregressive integrated moving average model on the monthly reported pulmonary tuberculosis cases in China ZHANG Shun-xian, QIU Lei, ZHANG Shao-yan, LI Cui, HU Jun, TIAN Li-ming, LU Zhen-hui. Respiratory Research Institute of Longhua Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai 200032, China

Corresponding author: LU Zhen-hui, Email: Dr_luzh@shutcm.edu.cn

【Abstract】 Objective An autoregressive integrated moving average (ARIMA) model was used to predict the monthly pulmonary tuberculosis cases in China (excluding Hong Kong, Macao and Taiwan regions) to provide a reference for pulmonary tuberculosis prevention and control. **Methods** Monthly pulmonary tuberculosis cases number in China from January 2006 to December 2018 reported on *Disease Surveillance* sponsored by CDC were collected. Based on these data, time series, preliminary identification and ordering of ARIMA model types were conducted using SPSS 26.0. Several ARIMA models were selected according to that both the simplicity of the model



开放科学 (资源服务) 标识码 (OSID) 的开放科学计划以二维码为入口, 提供丰富的线上扩展功能, 包括作者对论文背景的语音介绍、该研究的附加说明、与读者的交互问答、拓展学术圈等。读者“扫一扫”此二维码即可获得上述增值服务。

基金项目: “十三五”国家科技重大专项 (2018ZX10725-509)

作者单位: 200032 上海中医药大学附属龙华医院呼吸疾病研究所 (张顺先、邱磊、张少言、李翠、田黎明、鹿振辉), 微生物室 (胡骏)

通信作者: 鹿振辉, Email: Dr_luzh@shutcm.edu.cn

and the parameters of the ARIMA model (including autoregressive method (AR), average moving method (MA), seasonal autoregressive method (SAR), seasonal moving average method (SMA)) were statistically significant ($P_s < 0.05$), as well as the overall test index (Ljung-Box Q value), maximum stationary coefficient (R^2) of the model, standardized Bayesian information criterion value (NBIC) of the smallest overall model, and minimum root mean square error (RMSE). Numbers of reported cases from January to August 2019 were used as verification, and the model with the smallest relative error was selected as the optimal model according to that the smaller the relative error of the predicted value, the better the model; finally, the model was used to predict monthly reported numbers of tuberculosis patients from September 2019 to December 2020 in China. **Results** Time series were based on cases from January 2006 to December 2018, the fitted model was ARIMA (p, d, q) or ARIMA (p, d, q) \times (P, D, Q). Twelve models were selected according to P value (which is relative to Ljung-Box Q) > 0.05 , the simplicity of the model, and parameters of the model were statistically significant (all $P < 0.05$); and models with the maximum R^2 (ARIMA (1, 0, 1) (0, 1, 1)₁₂, $R^2 = 0.707$), or with the minimum RMSE (ARIMA (0, 1, 2) (0, 1, 1)₁₂, RMSE=9147.85), or with the minimum NBIC (ARIMA (0, 1, 1) (0, 1, 1)₁₂, NBIC=18.355), or with the minimum Ljung-Box Q (ARIMA (1, 1, 1) (0, 1, 1)₁₂, Ljung-Box Q =8.797) were taken as alternative models, to predict numbers of reported cases from January to August 2019, which were then compared with the actual data, to determine the optimal ARIMA model (ARIMA (0, 1, 1) (0, 1, 1)₁₂ model), with the relative error was the smallest (0.55%), MA (1)=0.875 ($t=19.243$, $P<0.001$), SMA (1)=0.876 ($t=7.596$, $P<0.001$), Ljung-Box Q =9.876 ($df=16$, $P=0.873$). The ARIMA (0, 1, 1) (0, 1, 1)₁₂ model was used to predict numbers of monthly reported tuberculosis cases in China from September 2019 to December 2020; in 2020 year, there will be 1 025 863 cases totally with average of 85 489 cases monthly. **Conclusion** ARIMA (0, 1, 1) (0, 1, 1)₁₂ model is the better model to predict the monthly pulmonary tuberculosis cases in China. However, in order to improve accuracy of the prediction, the establishment and prediction of the model is a dynamic process needed to be adjusted continuously according to accumulated data.

【Key words】 Tuberculosis, pulmonary; Disease notification; Epidemiologic research design; Models, statistical; Forecasting

结核病是由结核分枝杆菌引起的全球最流行的传染性疾病之一,可发生于全身多脏器,以肺部受累最为常见^[1],是世界上因单一病原体感染造成死亡最多的传染病^[2-3]。2019 年全球新发肺结核患者超过 1000 万例,肺结核相关的死亡患者超过 120 万例^[2]。近年来中国作为全球 30 个结核病高负担国家之一^[4],肺结核疫情具有患病率下降缓慢、疾病负担巨大和地区不平衡等特点,防治任务艰巨^[1, 5],准确掌握肺结核新发患者例数及流行特征,是开展肺结核相关健康教育、检测技能培训和配置相关卫生资源的前提和基础。伴随计算机软件的进步,很多数学模型因其能够较好地揭示传染病随时间发展的动态变化和预期结果相对准确等特点被广泛应用于传染病疫情的预测和研究,为疾病的防控提供积极参考^[6-7]。

自回归移动平均 (autoregressive integrated moving average, ARIMA) 模型从序列自相关的角度揭示了传染病的时间序列发展规律,既能解决数据间的自相关问题,又能充分利用其中的季节效应特征拟合疾病发生的实际情况,短期预测效果较好^[8-9],被广泛应用于传染病的预测预警。本研究采用该模型对我国(不含我国香港、台湾、澳门地区,下

同)肺结核月报告患者例数进行时间序列发展规律研究,以预测中国肺结核月报告患者例数,为结核病相关药品的生产、采购和就诊入院等卫生服务的供给预算提供科学参考。

资料和方法

一、资料来源

通过中国疾病预防控制中心主办的《疾病监测》杂志公布的中国每月甲、乙、丙类传染病疫情动态简介,搜集 2006 年 1 月至 2019 年 8 月全国肺结核月报告患者例数(不含中国港澳台地区)。采用 SPSS 26.0 软件对 2006 年 1 月至 2018 年 12 月的全国肺结核月报告患者例数进行拟合 ARIMA 备用模型,再以 2019 年 1—8 月的数据评价备用模型的预测效果和筛选最优模型。疫情动态数据由中国各级各类医疗卫生机构通过《传染病报告管理信息系统》在线报告、后期复核而产生。

二、ARIMA 建模

1. 时间序列的建立、特征分析及平稳化处理:建立时间序列,然后通过自相关系数函数图 (autocorrelation function, ACF) 和偏相关系数函数图 (partial autocorrelation function, PACF) 分析序列

的平稳性,如果序列不平稳,将其通过自然对数转换、一般差分和(或)季节差分等方法处理数据^[10]。

2. 模型识别:根据 ACF 图、PACF 图和 SPSS 26.0 统计学软件初步拟合的结果,尝试对模型进行初步识别和定阶。预先考虑使用 ARIMA(p, d, q) 或 ARIMA(p, d, q) × (P, D, Q) 模型,结合以往经验, p 和 q 一般不超过 3 阶, d 不会超过 2 阶, P、D 和 Q 不会超过 2 阶, p、q、P 和 Q 可依次取 0、1 和 2 并由低阶到高阶逐个尝试,同时 d 和 D 依次取 0 和 1 进行尝试^[10-11]。

3. 备选模型的整体诊断和参数判读:通过非线性最小二乘法估计模型参数,然后对模型残差进行白噪声(Ljung-Box Q)检验,以判断模型整体的合理性。然后再根据模型整体的检验指标(选择 $P > 0.05$ 的统计量,且 Ljung-Box Q 值越小越好)、整体模型是否简洁[即 p、d、q 和(或)P、D、Q 的阶越低越简洁]、ARIMA 模型各参数[自回归法(autoregressive, AR),平均移动法(moving average, MA),季节自回归法(seasonal autoregressive, SAR),季节移动平均法(seasonal moving average, SMA)]是否有意义、整体模型的平稳决定系数(stationary R-square, R^2 ;越大越平稳)、整体模型的标准化贝叶斯信息准则值(normalized bayesian information criterion, NBIC;越小越好),及整体模型的均方根误差(root mean square error, RMSE;越小越好)来筛选备选模型^[7, 12]。

4. 最优模型的确定:在满足模型成立的基本条件、完成备选模型的选择后,从预测值相对误差越小模型越优的原则出发^[13],应用不同的备选模型去预测 2019 年 1—8 月的全国肺结核月报告患者例数,

再与 2019 年 1—8 月实际报告的患者例数进行对比,选择预测平均相对误差最小的模型为最优模型。

5. 模型的预测应用:使用最优模型预测 2019 年 9 月至 2020 年 12 月我国每月的肺结核新发患者例数。

三、统计学处理

采用 Excel 2019 软件对 2006 年 1 月至 2019 年 8 月的全国肺结核月报告患者例数进行整理,采用 SPSS 26.0 软件进行 ARIMA 模型预测分析,通过比较预测数据和实际数据的相对误差来确定最优模型,并进行 2019 年 9 月至 2020 年 12 月数据的预测应用。

结 果

一、中国肺结核月报告患者例数的发病趋势分析

2006 年 1 月至 2018 年 12 月中国肺结核报告患者总计为 16 905 732 例,平均年报告患者例数为 1 300 441 例,从 2006—2008 年,报告患者例数缓慢上升(2006 年为 1 454 232 例、2007 年为 1 389 072 例、2008 年为 1 532 472 例),从 2009 年开始年报告患者例数逐步下降。2006 年 1 月至 2018 年 12 月中国肺结核平均月报告患者例数为 108 370 例,其中,2015 年 2 月报告患者例数最少(75 541 例),2008 年 3 月报告患者例数最多(156 679 例)。中国肺结核报告患者例数总体上呈现先短暂上升再缓慢下降的流行特征(图 1),即非平稳序列,同时也呈现出一定的季节特征,每年 2—4 月为流行高峰、11 月到次年 1 月为流行低谷。

二、模型的参数估计和模型诊断

以 2006 年 1 月至 2018 年 12 月中国肺结核月报告患者例数建立时间序列,发现其 ACF 图和

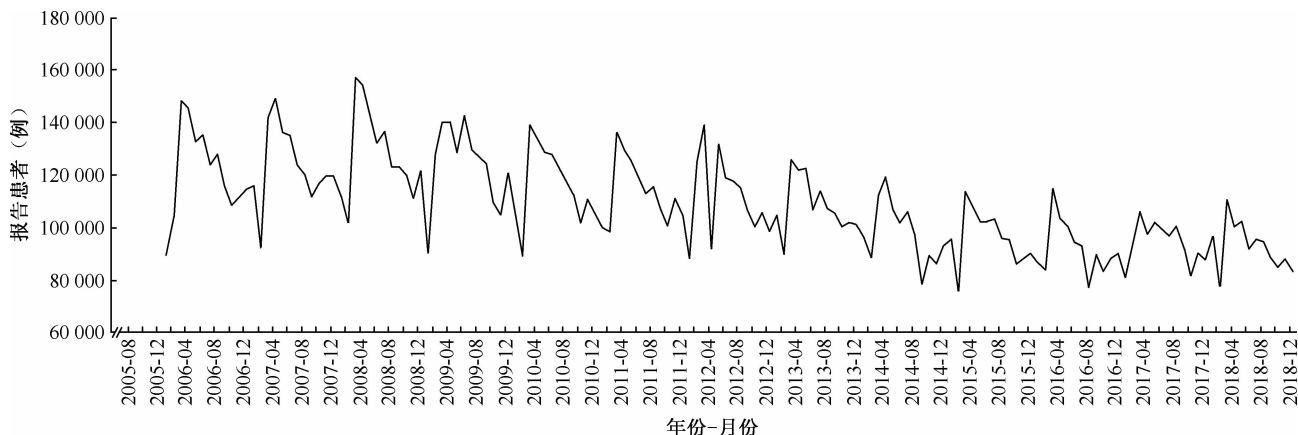


图 1 2006—2018 年全国肺结核月报告患者例数时间序列图

PACF 图的相关系数在延迟数目 12[即“滞后数(lags)”]时尚未落入 95% 的置信区间内,说明序列(连续 12 个月的数据)不稳定,原始数据可能在 12 个月的周期上有内在关联,遂将数据进行自然对数转化和 1 阶差分处理,需建立模型为 ARIMA(p,d,q)或 ARIMA(p,d,q)×(P,D,Q)。在模型总体的 Ljung-Box Q 值所对应的 P 值均 >0.05 的情况下,再根据模型是否简洁、模型各参数差异是否均有统计学意义(要求各参数的 t 值所对应的 P 值均 <0.05)筛选出 12 个基本模型(表 1)。然后再从这些基本模型中筛选出 R^2 最大的模型 ARIMA(1,0,1)(0,1,1)₁₂, $R^2=0.707$],RMSE 最小的模型

表 1 根据 2006—2018 年中国肺结核月报告患者例数建立的 ARIMA 模型

ARIMA 模型类型	参数	估计值	$s_{\bar{x}}$ 值	t 值	P 值	R^2 值	RMSE 值	NBIC 值	Ljung-Box Q 值	df 值	P 值
ARIMA(2,1,1) (1,1,0) ₁₂	AR(1)	-0.398	0.099	-4.009	<0.001	0.628	9 933.982	18.581	11.045	14	0.682
	AR(2)	-0.294	0.094	-3.119	0.002						
	MA(1)	0.754	0.072	10.458	<0.001						
	SAR(1)	-0.495	0.074	-6.652	<0.001						
ARIMA(1,1,1) (1,1,0) ₁₂	AR(1)	-0.248	0.091	-2.738	0.007	0.607	10 210.333	18.601	18.859	15	0.327
	SMA(1)	0.859	0.050	17.319	<0.001						
	SAR(1)	-0.458	0.076	-6.019	<0.001						
ARIMA(0,1,1) (1,0,1) ₁₂	MA(1)	0.88	0.042	20.777	<0.001	0.588	9 784.939	18.507	11.875	15	0.688
	SAR(1)	0.995	0.008	121.115	<0.001						
	SMA(1)	0.872	0.105	8.296	<0.001						
ARIMA(0,1,1) (0,1,1) ₁₂	MA(1)	0.875	0.045	19.243	<0.001	0.684	9 185.703	18.355	9.876	16	0.873
	SMA(1)	0.876	0.115	7.596	<0.001						
ARIMA(1,0,1) (1,0,1) ₁₂	AR(1)	0.991	0.013	76.633	<0.001	0.707	9 964.917	18.576	9.836	14	0.777
	MA(1)	0.861	0.053	16.183	<0.001						
	SAR(1)	0.990	0.011	92.778	<0.001						
	SMA(1)	0.817	0.097	8.422	<0.001						
ARIMA(0,1,1) (1,0,0) ₁₂	MR(1)	0.955	0.025	37.691	<0.001	0.482	10 815.905	18.675	21.403	16	0.164
	SAR(1)	0.646	0.063	10.312	<0.001						
ARIMA(1,0,1) (1,0,0) ₁₂	AR(1)	0.987	0.017	56.469	<0.001	0.634	10 833.238	18.710	18.407	15	0.242
	MR(1)	0.916	0.048	18.961	<0.001						
	SAR(1)	0.661	0.063	10.546	<0.001						
ARIMA(0,1,1) (1,1,0) ₁₂	MR(1)	0.907	0.039	23.467	<0.001	0.594	10 204.620	18.565	18.894	16	0.274
	SAR(1)	-0.439	0.078	-5.640	<0.001						
ARIMA(1,1,1) (0,1,1) ₁₂	AR(1)	-0.205	0.096	-2.139	0.034	0.690	9 206.244	18.394	8.797	15	0.888
	MA(1)	0.833	0.056	14.869	<0.001						
	SMA(1)	0.840	0.099	8.495	<0.001						
ARIMA(1,1,2) (0,1,1) ₁₂	AR(1)	-1.000	0.001	-1 157.500	<0.001	0.686	9 247.755	18.438	10.347	14	0.736
	MA(1)	-0.125	0.049	-2.538	0.012						
	MA(2)	0.869	0.050	17.365	<0.001						
	SMA(1)	0.915	0.163	5.625	<0.001						
ARIMA(0,1,2) (0,1,1) ₁₂	MA(1)	1.096	0.085	12.907	<0.001	0.692	9 147.85	18.381	9.155	15	0.817
	MA(2)	-0.251	0.086	-2.935	0.004						
	SMA(1)	0.819	0.093	8.840	<0.001						

注 AR:自回归法;MA:平均移动法;SAR:季节自回归法;SMA:季节移动平均法; R^2 :整体模型的平稳决定系数;NBIC:整体模型的标准化贝叶斯信息准则值;RMSE:整体模型的均方根误差

ARIMA(0,1,2)(0,1,1)₁₂, RMSE=9147.85]、NBIC 最小的模型[ARIMA(0,1,1)(0,1,1)₁₂, NBIC=18.355]、Ljung-Box Q 值最小的模型[ARIMA(1,1,1)(0,1,1)₁₂, Ljung-Box Q=8.797]作为备用模型。

三、最优模型的确定

从预测值相对误差越小模型越优的原则出发确定最优模型。分别以备用模型 ARIMA(0,1,1)(0,1,1)₁₂、ARIMA(0,1,2)(0,1,1)₁₂、ARIMA(0,1,1)(0,1,1)₁₂和 ARIMA(1,1,1)(0,1,1)₁₂预测 2019 年 1—8 月中国肺结核月报告患者例数(表 2),并与实际的月报告患者例数进行比较,发现 ARIMA(0,1,1)(0,1,1)₁₂模型的预测平均相对误差最小(0.55%),预测效果较好;其余 3 种模型的预测平均相对误差均较大(分别为 3.01%、1.16%和 0.58%),预测效果相对不理想。另外,从预测值最大相对误差的角度讲,ARIMA(1,0,1)(1,0,1)₁₂的最大相对误差为 10.88%、ARIMA(1,0,1)(1,0,1)₁₂的最大相对误差为 8.56%、ARIMA(1,1,2)(0,1,1)₁₂的最大相对误差为 9.58%,均高于 ARIMA(0,1,1)(0,1,1)₁₂的最大相对误差(8.17%),故最优模型确定为 ARIMA(0,1,1)(0,1,1)₁₂。

四、模型的预测应用

通过 ARIMA(0,1,1)(0,1,1)₁₂模型预测 2019

年 9 月至 2020 年 12 月中国肺结核月报告患者例数(图 2),结果显示 2019 年 9—12 月预测患者例数分别为 84 399(95%CI: 98 844, 71 586)例、79 928(95%CI: 93 708, 67 716)例、82 551(95%CI: 96 884, 69 858)例、82 598(95%CI: 97 042, 69 818)例;2020 年 1—12 月预测患者例数分别为 78 381(95%CI: 92 488, 65 941)例、75 614(95%CI: 89 338, 63 524)例、101 847(95%CI: 120 486, 85 444)例、95 075(95%CI: 112 616, 79 653)例、94 289(95%CI: 111 825, 78 887)例、89 862(95%CI: 106 707, 75 081)例、88 964(95%CI: 105 770, 74 231)例、84 980(95%CI: 101 157, 70 811)例、81 188(95%CI: 96 761, 67 561)例、76 873(95%CI: 91 730, 63 886)例、79 380(95%CI: 94 834, 65 882)例、79 410(95%CI: 94 984, 65 821)例(图 2)。预计 2020 年新发肺结核患者约 1 025 863 例,平均每月 85 489 例。

讨 论

近年来,ARIMA 模型、灰色模型和神经网络等数理统计学因其能相对准确的探索传染病的发生发展规律,在传染病预测研究领域的应用非常活跃,为传染病的有效防控提供了科学参考和重要依据。ARIMA 模型之所以能在疾病预测领域中广泛应用,因其既吸收了传统回归分析的优点又发挥了移

表 2 2019 年 1—8 月中国肺结核患者报告例数预测值与实际值比较情况

月份	实际报告 患者例数	预测患者例数							
		ARIMA(0,1,1) (0,1,1) ₁₂	相对误 差(%)	ARIMA(1,0,1) (1,0,1) ₁₂	相对误 差(%)	ARIMA(1,1,1) (0,1,1) ₁₂	相对误 差(%)	ARIMA(1,1,2) (0,1,1) ₁₂	相对误 差(%)
1	88 597	81 357	−8.17	84 631	−4.48	83 187	−6.11	81 645	−7.85
2	73 096	78 500	7.39	81 047	10.88	78 947	8.00	77 350	5.82
3	97 866	105 754	8.06	106 916	9.25	106 248	8.56	107 240	9.58
4	101 191	98 741	−2.42	99 946	−1.23	98 852	−2.31	97 941	−3.21
5	96 106	97 944	1.91	99 922	3.97	98 403	2.39	99 288	3.31
6	99 555	93 363	−6.22	95 174	−4.40	93 465	−6.12	92 607	−6.98
7	93 318	92 447	−0.93	95 091	1.90	92 932	−0.41	93 662	0.37
8	84 304	88 324	4.77	91 223	8.21	88 746	5.27	87 307	3.56
平均相 对误差			0.55		3.01		1.16		0.58

注 平均相对误差为 2019 年 1—8 月相对误差之和/8×100%

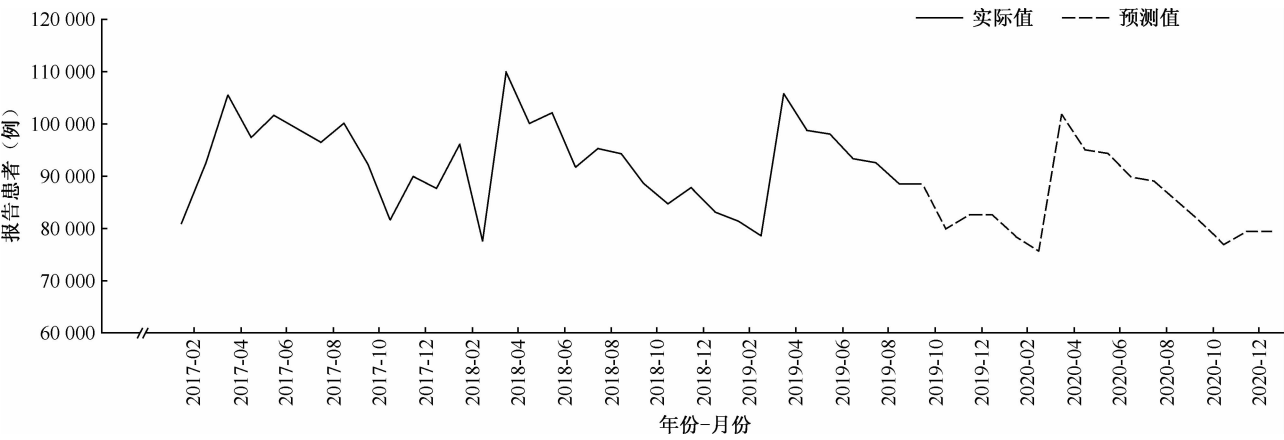


图2 最优模型 ARIMA(0,1,1)(0,1,1)₁₂预测 2019 年 9 月至 2020 年 12 月中国肺结核月报告患者例数拟合图

动平均的长处,可以将诸多影响传染病发生发展的复杂因素(如人口流动、气温和湿度等因素)综合统一并蕴含于时间变量中,借助趋势变化、周期变化、随机干扰和模型参数进行量化表达,是一种适用范围广、实用性强、预测精确度较高、预测效果优于回归模型的预测方法。但从传染病的长期趋势来看,疫苗的出现和推广会极大地改变相关传染病的流行特征,如果采用 ARIMA 进行长期预测会出现一定的偏差,故 ARIMA 只适合短期预测,不适合长期预测^[14]。另外,由于用 ARIMA 进行预测时一般需要 10 个以上的周期性变化的数据,故在急性传染病的早期阶段、未出现疾病流行的周期性变化趋势之前,ARIMA 不能做预测,如 2019 年年底开始在全球流行的新型冠状病毒肺炎疾病(COVID-19),直至 2020 年 3 月 30 日仍不具备一个完整的周期性规律变化,如果以这段时间每天的报告患者例数进行 ARIMA 预测,结果将出现很大的偏差。而对那些已出现过多个周期性趋势的慢性传染病(如肺结核),因其致病因素、影响因素、防控措施和策略(如疫苗等)不会在短时间内出现大的改变,导致 ARIMA 对这类疾病的短期预测效果较好。

肺结核是我国常见的一种对人群威胁较大的呼吸道传染病。本研究根据 2006 年 1 月至 2018 年 12 月我国传染病监测系统中肺结核报告病例数的变化趋势,发现中国肺结核新发患者例数总体上呈缓慢下降的趋势,可能与这些年中国肺结核大力推行短程督导化疗方案有关,同时也与大力推广肺结核患者的早发现、早诊断和早治疗诊治措施有关^[15-16]。由于高速铁路网的迅速发展加剧了人口

迁移和流动,以及结核分枝杆菌耐药性的出现和逐渐加重使得肺结核新发患者例数下降缓慢^[17-18]。本研究发现中国肺结核发病总体在时间分布上具有鲜明的季节特征和周期变化趋势,我国每年肺结核的发病高峰为 2—4 月,这可能与春节期间大量的结核病患者不按规定进行自我隔离治疗,不采取戴口罩等防护措施,频繁的流动和参加聚集性活动,造成较大范围的传播有关^[17-18]。同时,2—4 月是我国的冬春季,天气寒冷,空气污染指数相对较高,不利于家庭开窗通风,增加了呼吸道感染(包括肺结核)的发生和发展^[3, 19]。因此,提高肺结核患者预防传染病的自我意识,使其主动在家隔离、治疗并佩戴口罩外出;同时,建议家庭每天开窗通风,可在一定程度上降低发生肺结核的风险。

本研究应用 ARIMA 模型对我国近年来肺结核新发患者例数进行分析,在掌握肺结核发病规律和流行特征的同时,探索预测效果较好的模型。准确预测肺结核的发病趋势,不仅可评估防控措施的实施效果,同时对生产和采购抗结核相关药品、诊断试剂、耗材和卫生服务、合理分配卫生资源、制定最优的控制策略和措施具有重要意义,有助于提升肺结核防控的预见性和主动性。本研究通过预测发现 2020 年我国新发肺结核患者约 1 025 863 例,平均每月 85 489 例,这和近几年中国肺结核流行强度基本一致^[12],但稍高于 WHO 的预测结果(90 万新发患者)^[1],这可能与 ARIMA 更适合短期预测有关,ARIMA 进行长期预测可能会高估结果。因此,我们认为在没有其他因素干扰的情况下,2020 年结核病防控相关的医疗资源、防控措施和力度仍需保持

近 2 年的强度,不能降低各种相关资源的投入。

综上所述,ARIMA 模型对中国肺结核新发患者例数预测效果较好,可为肺结核的防治提供科学参考。但模型的建立和预测应用是个动态过程,需不断根据积累的数据进行调整,从而提高预测精度。

参 考 文 献

- [1] Furin J, Cox H, Pai M. Tuberculosis. *Lancet*, 2019, 393 (10181):1642-1656.
- [2] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990—2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*, 2018, 392(10159):1789-1858.
- [3] 赵珍,刘年强,依帕尔·艾海提,等. 2008—2018 年新疆维吾尔自治区肺结核空间流行病学特征分析. *中国防痨杂志*, 2019, 41(8):893-899.
- [4] Duan Q, Chen Z, Chen C, et al. The prevalence of drug-resistant tuberculosis in mainland China: an updated systematic review and meta-analysis. *PLoS One*, 2016, 11(2): e0148041.
- [5] Yang C, Luo T, Shen X, et al. Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect Dis*, 2017, 17 (3):275-284.
- [6] 言晨绮,王瑞白,刘海灿,等. ARIMA 模型预测 2018—2019 年我国肺结核发病趋势的应用. *中华流行病学杂志*, 2019, 40 (6):633-637.
- [7] 王华,田昌伟,王文明,等. SARIMA 模型和 SARIMA-GRNN 组合模型在肺结核发病率预测中的比较. *寄生虫病与感染性疾病*, 2019, 17(1):28-31.
- [8] 孙娜,许小珊,冯佳宁,等. ARIMA 与 GM(1,1)模型对我国肺结核年发病人数预测情况的比较. *中国卫生统计*, 2019, 36 (1):71-74.
- [9] Huang L, Li XX, Abe EM, et al. Spatial-temporal analysis of pulmonary tuberculosis in the northeast of the Yunnan province, People's Republic of China. *Infect Dis Poverty*, 2017, 6(1):53.
- [10] 谭恩丽,侯慧玉,包海荣,等. 采用自回归移动平均模型预测中国流感病例数. *病毒学报*, 2017, 33(5):699-705.
- [11] 陈之源,常玉雪,叶尔扎提·吾瓦特,等. 灰色模型在新疆维吾尔自治区伊宁市肺结核发病率预测中的应用. *中国防痨杂志*, 2019, 41(12):1314-1317.
- [12] 薛亚妮,张梅,李存龙. 基于灰色模型的全国肺结核疫情预测及分析. *中国防痨杂志*, 2019, 41(7):782-789.
- [13] 谭恩丽,王正峰,周文策,等. 我国包虫病报告病例数自回归移动平均模型预测研究. *中国血吸虫病防治杂志*, 2018, 30 (1):47-53.
- [14] 王娜,王黎霞,李仁忠. 四地市结核病定点医院住院初治涂阳肺结核患者医疗费用及经济负担分析. *中国防痨杂志*, 2012, 34(2):79-84.
- [15] 赵飞,杜昕,李涛,等. 基于世界卫生组织公共数据库的中国结核病流行趋势与预测. *临床药物治疗杂志*, 2018, 16(4):1-3.
- [16] Dong D, Jiang WX, Long Q, et al. Impact of an innovative tuberculosis financing and payment model on health service utilization by tuberculosis patients in China: do the poor fare better than the rich? *Infect Dis Poverty*, 2019, 8(1):44.
- [17] Lange C, Dheda K, Chesov D, et al. Management of drug-resistant tuberculosis. *Lancet*, 2019, 394(10202):953-966.
- [18] 李雨晴,李李,陈杏,等. 中国西北四省(区)结核分枝杆菌分离株一线药物耐药状况及其影响因素分析. *中国人兽共患病学报*, 2017, 33(5):398-402.
- [19] 丛明瑶,云妙英,阿德娜依·阿力肯. 气候因子对肺结核发病率影响的分析. *中华疾病控制杂志*, 2014, 18(11):1051-1054.

(收稿日期:2020-01-17)

(本文编辑:孟莉 薛爱华)