

# X-12-SARIMA 在甘肃省猩红热疫情分析及短期预测中的应用

蒋小娟, 刘新风, 成瑶, 杨筱婷

甘肃省疾病预防控制中心, 甘肃 兰州 730000

**摘要:** **目的** 探讨 X-12 和季节性差分自回归滑动平均模型(seasonal autoregressive integrated moving average, SARIMA)在甘肃省猩红热疫情分析及短期预测中的应用。 **方法** 利用 2010—2017 年甘肃省猩红热月发病率数据建立 SARIMA 模型并进行短期预测,运用 X-12 季节调整法分析疾病流行的季节波动等特征。 **结果** 建立的 SARIMA(3,1,1)(1,1,1)<sub>12</sub>模型参数估计值均有统计学意义,残差为白噪声序列,预测值的精度评价指标和误差衡量指标均符合标准。2010—2018 年甘肃省猩红热月发病率存在明显季节波动,季节因子影响 6 月最大且呈缓慢下降趋势,循环-趋势成分影响呈缓慢上升趋势,不规则因子影响规律平稳。 **结论** X-12 季节调整方法能较好地分析具有一定季节波动和长期趋势传染病的时间变化规律,SARIMA 模型对于甘肃省猩红热的短期预测效果较好。

**关键词:** X-12-SARIMA 模型;猩红热;季节调整;预测

**中图分类号:**R181.2;R515.1 **文献标识码:**A **文章编号:**1006-3110(2022)12-1541-04 **DOI:**10.3969/j.issn.1006-3110.2022.12.033

## Application of X-12-SARIMA to epidemic analysis and short-term forecasting of scarlet fever in Gansu Province

JIANG Xiao-juan, LIU Xin-feng, CHENG Yao, YANG Xiao-ting

Gansu Provincial Center for Disease Control and Prevention, Lanzhou, Gansu 730000, China

**Abstract:** **Objective** To explore the application of X-12 and seasonal autoregressive integrated moving average (X-12-SARIMA) model to epidemic analysis and short-term prediction of scarlet fever in Gansu Province. **Methods** Using the monthly incidence data of scarlet fever in Gansu Province from 2010 to 2017, a SARIMA model was established for short-term prediction. X-12 seasonal adjustment method was used to analyze the seasonal fluctuation characteristics of the disease epidemic. **Results** The estimated values of all parameters in the SARIMA (3,1,1)(1,1,1)<sub>12</sub> model established were significant in statistics. The residuals were proved to be white-noise series. The accuracy evaluation and error measurement indicators of the predicted value were up to the standards. The monthly incidence rates of scarlet fever in Gansu Province from 2010 to 2018 had obvious seasonal fluctuations. The influence of seasonal factors was most obvious in June and showed a slow decreasing trend, while the influence of cycle-trend factors showed a slow increasing trend. The influence of irregular factors was regular and smooth. **Conclusion** X-12 seasonal adjustment method is fit for the analysis of time variation of infectious diseases with seasonal fluctuations and long-term trends. The SARIMA model is effective to forecast short-term epidemic level of scarlet fever in Gansu Province.

**Keywords:** X-12-SARIMA model; scarlet fever; seasonal adjustment; forecast

猩红热是由 A 组  $\beta$  型溶血性链球菌(group A- $\beta$  hemolytic Streptococcus, GAS)引起的急性呼吸道传染病,人群普遍易感,2~10 岁儿童高发,冬春季为流行季节,尚无特异性免疫手段。近年来,甘肃省猩红热疫情呈上升趋势。本研究利用 2010—2017 年甘肃省猩红

**基金项目:**“十三五”国家科技重大专项“甘肃及周边省区传染病病原谱流行规律研究”课题(2017ZX10103006);甘肃省卫生行业科研计划项目(GSWKY-2019-67)

**作者简介:**蒋小娟(1982-),女,硕士研究生,副主任医师,主要从事急性传染病防控工作。

热月发病率建立季节性差分自回归滑动平均模型(seasonal autoregressive integrated moving average, SARIMA),对甘肃省猩红热发病率进行短期预测,利用 X-12 法对延长的猩红热发病率序列进行季节波动、长期趋势和不规则因子的分解,分析疾病流行的季节性及影响因素,为猩红热的防控提供科学依据。

### 1 资料与方法

**1.1 数据来源** 2010—2018 年甘肃省猩红热月发病数和发病率数据来源于中国疾病预防控制中心信息系统。

**1.2 方法** SARIMA 模型 SARIMA 模型是由 Box 和



Jenkins 于 20 世纪 70 年代初提出的著名时间序列预测方法,是将非平稳时间序列转化为平稳时间序列,然后将因变量仅对其的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型<sup>[1]</sup>。

1.2.1 数据预处理 观察甘肃省 2010 年 1 月—2017 年 12 月猩红热月发病率时间序列图,并通过单位根检验(Augmented Dickey-Fuller, ADF)判断序列的平稳性,如果为非平稳时间序列,则对原序列进行差分处理;对差分后序列再次进行 ADF 检验,同时采用 Box-Ljung  $Q$  进行白噪声检验,确定差分后序列为平稳非白噪声序列。

1.2.2 模型识别、诊断 根据预处理后平稳时间序列的自相关和偏自相关图,估计  $ARIMA(p,d,q)(P,D,Q)_s$  的自回归阶数  $p$ 、差分阶数  $d$ 、移动平均阶数  $q$  及季节周期的自回归阶数  $P$ 、差分次数  $D$ 、移动平均阶数  $Q$ ,  $s$  是季节性的周期,本研究为 12;在所有参数均通过  $t$  检验的模型中,根据 AIC 值和 SBC 值确定模型阶数;对识别的模型残差序列进行白噪声检验,若残差序列为白噪声,则证明模型适应性良好。

X12 季节调整 X-12 季节调整是美国普查局季节调整首席研究员 David Findley 关于季节调整研究的最新成果,是以 X-11 和 X-11-ARIMA 为基础扩展的 X-11 季节调整程序<sup>[2]</sup>。X-12 核心算法是扩展的 X-11 季节调整程序,除了包括 X-11 的全部过程外,还扩展了贸易日和节假日影响的调节功能,增加了季节因子(seasonal factor, SF)、循环趋势因子(trend factor, TC)、不规则因子(irregularity factor, IR)分解模型的选择功能,季节调整结果稳定诊断功能以及 X-12 的建模和选择功能。

1.3 统计学分析 数据处理、模型建立和季节分解利用 EViews 9.0 软件完成。

## 2 结果

2.1 建立 SARIMA 模型及短期预测 将甘肃省 2010 年 1 月—2017 年 12 月的月发病率绘制成时间序列图(见图 1),显示甘肃省猩红热发病呈现明显的季节性和周期性,每年的 5—6 月和 11—12 月发病率相对较高,且 2010—2017 年发病率呈逐年上升的趋势,可初步识别该序列为非平稳序列。经 ADF 检验,结果显示单位根统计量  $ADF = -1.891$ ,大于显著性水平 1%~10% 的 ADF 临界值,证明该序列的确为非平稳序列。

采用 1 阶差分法消除序列趋势性,再进行一次步长为 12 的季节差分后,时序图显示序列基本平稳。经 ADF 检验,结果显示  $ADF = -7.471$ ,小于显著性水平

1%~10% 的 ADF 临界值,证明经差分处理后的序列已转换为平稳序列。差分后序列的自相关和偏自相关分析显示  $Q$  值对应的  $P$  值均  $<0.05$ ,证明该序列为非白噪声序列。

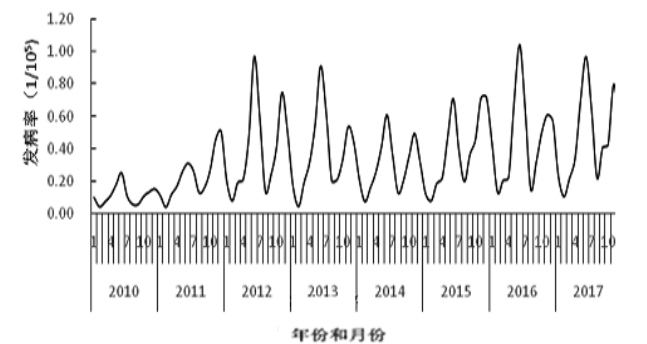


图 1 甘肃省 2010—2017 年猩红热月发病率时间序列图  
表 1 经 1 阶-12 步差分后序列的自相关和偏自相关分析结果

Autocorrelation		Partial Correlation		AC	PAC	Q-Stat	Prob
1		1		-0.005	-0.005	0.002	0.966
2		2		-0.371	-0.371	11.986	0.002
3		3		-0.241	-0.284	17.116	0.001
4		4		-0.008	-0.220	17.122	0.002
5		5		0.121	-0.140	18.448	0.002
6		6		0.110	-0.069	19.550	0.003
7		7		0.070	0.038	20.000	0.006
8		8		-0.162	-0.138	22.478	0.004
9		9		0.033	0.119	22.581	0.007
10		10		0.070	0.060	23.049	0.011
11		11		0.069	0.151	23.510	0.015
12		12		-0.240	-0.204	29.230	0.004

对差分后序列进行自相关和偏自相关分析(见表 1),自相关系数(autocorrelation coefficient function, ACF)和偏自相关系数(partical autocorrelation coefficient function, PACF)显示均拖尾,且序列进行了差分运算,因此可设定为 ARIMA 过程;ACF 呈 2、3、12 阶截尾, PACF 呈 2、3、4 阶截尾,因此可初步设  $p=4,3,2,1$  或 0,  $P=0$  或 1,  $q=3,2,1$  或 0,  $Q=0$  或 1, 且  $p$  和  $q$ ,  $P$  和  $Q$  不同时为 0;之前采用 1 阶-12 步差分法消除序列的趋势性和季节性,因此  $d=1, D=1$ , 选用乘积季节模型  $SARIMA(p,1,q) \times (P,1,Q)_{12}$ 。模型检验结果所有参数均通过  $t$  检验的模型中,  $SARIMA(3,1,1)(1,1,1)_{12}$  的 AIC 值(-1.528)和 SBC 值(-1.295)均最小,因此  $SARIMA(3,1,1)(1,1,1)_{12}$  为最优模型。生成残差序列并对其进行相关和偏自相关分析,  $Q$  值较大且其对应的  $P$  值均大于 0.05,显示残差序列不存在自相关性,且为白噪声,模型适应性良好。应用  $SARIMA(3,1,1)(1,1,1)_{12}$  对 2018 年 1—12 月甘肃省猩红热发病率进行预测,与实际发病率进



行比较,并对预测结果进行精度评价(见图 2)。预测结果精度评价指标:均方根误差(root mean squared error, RMSE) = 0.079, 平均绝对误差(mean absolute error, MAE) = 0.069, 平均绝对百分比误差(mean absolute percent error, MAPE) = 14.477, 希尔不等系数(theil inequality coefficient, TIC) = 0.079。其中, RMSE 和 TIC 取值范围为 0~1, MAPE 取值范围为 0~100, 且均越小表明模型预测效果越优。常见误差衡量指标:偏倚比例(bias proportion) = 0.029, 方差比例(variance proportion) = 0.270, 协方差比例(covariance proportion) = 0.701; 三者之和为 1, 偏倚比例和方差比例之和越小, 协方差比例越大, 表明模型预测结果越理想。综合以上各项指标, SARIMA(3,1,1)(1,1,1)<sub>12</sub> 短期预测效果良好。

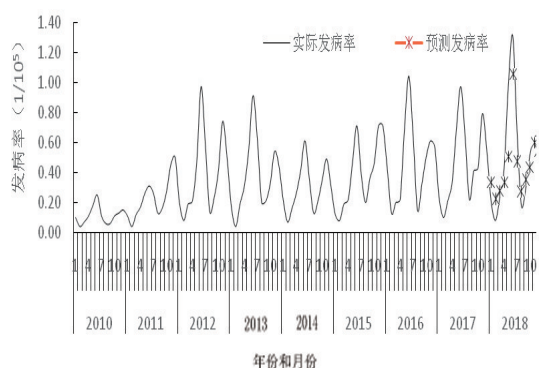


图 2 甘肃省 2018 年猩红热月发病率预测

2.2 X-12-SARIMA 模型预测及季节调整 应用 X-12 过程对延长后的时间序列 2007 年 1 月—2018 年 12 月猩红热发病率数据进行季节调整。诊断报告中, 季节调整质量统计量 M1-M11 均小于 1, 季节调整质量复合指标  $Q=0.470$ , 对季节调整结果为接受。

季节调整后序列: 经季节调整后的序列整体趋势与原序列基本一致; 每年均是 2 月发病率最低, 并低于原序列; 6 月发病率最高, 并高于原序列, 见图 3。

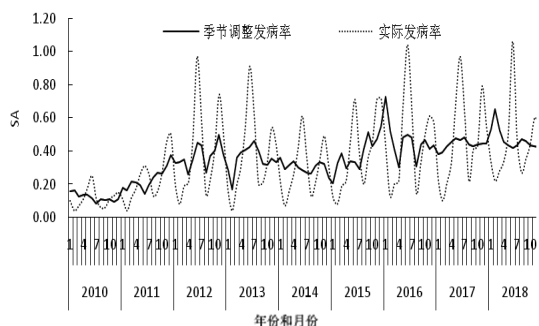


图 3 季节调整后发病率时序图

SF: 季节调整后, 分解出的季节因子时序图显示, 影响甘肃省猩红热发病率的季节性因素以年度为周期, 每年 2 月影响最小, 6 月影响最大, 且呈逐年缓慢下降趋势, 见图 4。

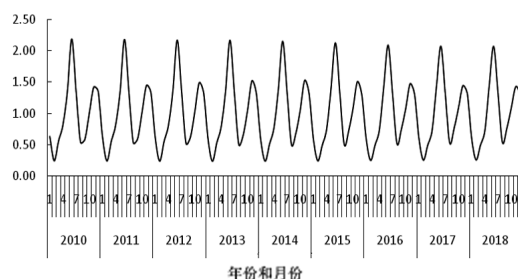


图 4 SF 时序图

TC: 趋势-循环成分是从时间序列中过滤掉季节因子和不规则波动, 暴露出时间序列的长期变化趋势, 包含长期趋势和周期循环。甘肃省猩红热发病率的循环-趋势成分以年为周期, 总体为逐年上升趋势, 见图 5。

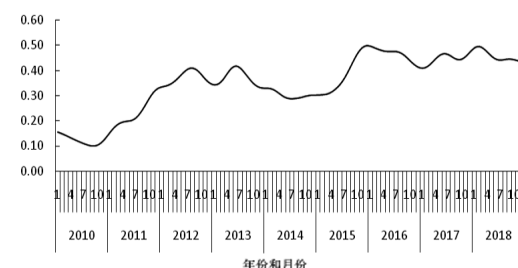


图 5 趋势-循环因子(TC)时序图

IR: 时间序列去除季节因子、循环-趋势成分后, 剩余的不规则成分则是不规则波动, 包含其他各种可引起疾病发病率波动的偶然因素。每年影响甘肃省猩红热发病率的不规则因子以年为周期, 且较规律和平稳, 见图 6。

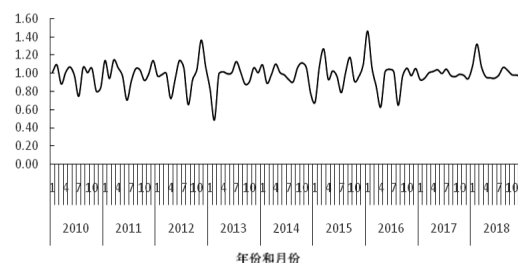


图 6 不规则因子(IR)时序图

### 3 讨论

猩红热曾经是全球广泛流行的严重传染病。由于青霉素的发现以及二战结束后社会经济和医疗条件的改善, 猩红热的发病率和病死率大幅下降<sup>[3]</sup>。但从 20 世纪 80 年代中期到 90 年代, 严重侵袭性 GAS 感染在欧洲和北美卷土重来。近 10 年, 世界各国有关猩红热暴发的报道也再次增多, 除中国之外还有越南、英国、西班牙、加拿大和美国等<sup>[4-9]</sup>。2011 年以来, 全国报告的猩红热病例数急剧上升, 2004—2010 年的年均发病率为 1.91/10 万, 2011—2016 年为 4.01/10 万, 2011—2016 年的年均发病率约是 2004—2010 年的



2.1 倍<sup>[10]</sup>。2015 年 1—7 月全国监测数据显示,甘肃省报告病例数为在全国处于中等水平<sup>[11]</sup>,2015—2018 年甘肃省猩红热发病率依次为 4.61/10 万、5.50/10 万、5.57/10 万和 6.15/10 万,在全国处于中等偏上水平,与 2018 年全国猩红热发病水平较前一年有所上升,病例主要集中在东北、华北和西北地区的结论一致<sup>[12]</sup>。目前,猩红热尚无有效可用的疫苗,因此发病率的预测对于猩红热的防控具有重要意义<sup>[13-14]</sup>。

许多传染病都存在季节性<sup>[15]</sup>。本研究对 2010 年以来甘肃省猩红热发病率进行时间序列分析,建立 SARIMA 模型开展短期预测,并采用 X-12 季节调整探讨影响疾病发病率的因子,寻找其规律。相关研究显示,只有基于至少 50 个以上的时间序列数据构建的 ARIMA 模型才能得到较满意的预测效果<sup>[13,16-18]</sup>。本研究利用甘肃省 2010—2017 年猩红热月发病率共 96 个数据构建 SARIMA(3,1,1)(1,1,1)<sub>12</sub> 并进行短期预测,结合各项预测评价指标,显示预测效果良好。2010—2018 年甘肃省猩红热以年为周期,每年 1 月发病率最低,逐渐上升到 6 月达到年度发病率最高峰,随后逐步下降到 8 月,再次上升到 12 月达到一个小高峰,与全国传染病的发病基本一致,说明每年的夏季和秋冬交替时段为传染病的高发阶段<sup>[19]</sup>。

X-12-SARIMA 模型是由 X12 方法和 SARIMA 时间序列模型组合而成的季节调整方法。本研究应用该方法进行季节调整,研究影响甘肃省 2010—2017 年猩红热月发病率的季节因子等因素。结果显示:经过季节调整的发病率时间序列总体趋势与原序列基本一致,发病率最低和最高的时间有较强的规律性,每年 2 月的发病率最低,且低于对应时间的原始序列值,6 月发病率最高,且对应时间的高于原始序列,说明每年 2 月、季节因子、趋势-循环成分及不规则波动对发病率的综合影响为负向,反之 6 月则为正向。季节因子时序图显示,影响甘肃省猩红热发病率的季节性因素以年度为周期的规律明显且平稳。每年 2 月季节因子影响最小,6 月影响最大,与调整后疾病发病率的时序一致,可见,季节因子在疾病的流行中起着重要的作用,也就是说,猩红热的发病存在显著的季节性。此外,研究结果显示,循环-趋势成分为逐年上升趋势,不规则波动的影响则以年为周期,规律且平稳。

X-12 另一重要的扩展功能,即涉及到贸易日和节假日影响的调节功能,但设定节假日均为西方节日,因此在使用时存在限制。X-12-ARIMA 只适合季度和月度统计数据,样本观察值和估测数据个数均有限

制。但在对存在季节性的传染病流行的季节性分析及短期预测方面确实具有较强的适应性,在国内相关研究中也广泛应用<sup>[20]</sup>。

## 参考文献

- [1] 李小松,冯子健,殷菲,等. 传染病时空聚集性探测与预测预警方法[M]. 第 3 版. 北京:高等教育出版社,2014.
- [2] 耿娟娟. 基于 X-12-ARIMA 和 SARMA 模型及其组合模型的 CPI 预测研究[D]. 成都:西南石油大学,2015.
- [3] Quinn RW. Comprehensive review of morbidity and mortality trends for rheumatic fever, streptococcal disease, and scarlet fever: the decline of rheumatic fever [J]. Clin Infect Dis, 1989, 11(6): 928-953.
- [4] Wong SSY, Yuen KY. Streptococcus pyogenes and re-emergence of scarlet fever as a public health problem [J]. Emerg Microbes Infect, 2012, 1(7): 1-10.
- [5] Yang P, Peng XM, Zhang DT, et al. Characteristics of group A Streptococcus strains circulating during scarlet fever epidemic, Beijing, China, 2011 [J]. Emerg Infect Dis, 2013, 19(6): 909-915.
- [6] Chen ML, Yao WL, Wang XH, et al. Outbreak of scarlet fever associated with emm12 type group A Streptococcus in 2011 in Shanghai, China [J]. Pediatr Infect Dis J, 2012, 31(9): 158-162.
- [7] Guy R, Williams C, Irvine N, et al. Increase in scarlet fever notifications in the United Kingdom, 2013/2014 [J]. Euro Surveill, 2014, 19(12): 179-184.
- [8] Guy R, Chand M. Resurgence of scarlet fever in England, 2014-2016: a population based surveillance study [J]. Lancet Infect Dis, 2018, 18(2): 180-182.
- [9] Walker MJ, Brouwer S. Scarlet fever makes a comeback [J]. Lancet Infect Dis, 2018, 18(2): 128-129.
- [10] Liu YH, Chan TC, Yap LW, et al. Resurgence of scarlet fever in China: a 13 year population based surveillance study [J]. Lancet Infect Dis, 2018, 28(6): 566-571.
- [11] 秦颖,冯录召,余宏杰. 2015 年春夏季全国猩红热疫情流行病学特征分析[J]. 疾病监测, 2015, 30(12): 1002-1007.
- [12] 高福,冯子健,王健,等. 2018 年中国传染病监测报告[R]. 北京:中国疾病预防控制中心,2019 年.
- [13] 孔德川,潘浩,郑雅旭,等. ARIMA 模型在上海市猩红热发病率预测中的应用[J]. 实用预防医学, 2020, 27(8): 1011-1013.
- [14] Zhang X, Liu YC. The resurgence of scarlet fever in China [J]. Lancet Infect Dis, 2018, 18(8): 823-824.
- [15] 罗静雯,刘元元,李晓松. 基于 X11-ARIMA 方法的猩红热季节波动分析及短期预测[J]. 现代预防医学, 2010, 37(12): 3816-3818.
- [16] 陆波,闵思韬,闵红星,等. 应用 ARIMA 模型预测麻疹发病率的可行性研究[J]. 中国卫生统计, 2015, 32(1): 106-107.
- [17] 杨其松,朱蒙蒙,张天琛,等. ARIMA 模型在宜春市肾综合征出血热发病率预测中的应用[J]. 中国卫生统计, 2018, 35(5): 713-715.
- [18] 宋媛媛,王雷,熊甜,等. ARIMA 模型与 GM(1,1) 模型在痢疾发病数预测中的比较研究[J]. 实用预防医学, 2019, 26(7): 888-892.
- [19] 王怡,张震,范俊杰,等. ARIMA 模型在传染病预测中的应用[J]. 中国预防医学杂志, 2015, 16(6): 424-428.
- [20] 范维,张磊,石刚. 季节调整方法综述及比较[J]. 统计研究, 2006, 2(2): 70-73.