

SARIMA 模型在苏州市细菌性痢疾发病预测中的应用

王建树, 刘强, 覃江纯, 杭惠, 杨海兵
苏州市疾病预防控制中心, 江苏 苏州 215004

摘要: **目的** 探讨季节性差分自回归求和滑动平均(seasonal auto-regressive integrated moving average, SARIMA)模型在苏州市细菌性痢疾月发病数预测中的应用。**方法** 利用 R i386 3.2.3 软件对 2005 年 1 月-2018 年 4 月苏州市细菌性痢疾月发病数据构建 SARIMA 模型,对 2018 年 5-7 月份细菌性痢疾的月发病人数进行预测,验证预测效果。**结果** 建立了 SARIMA(0,1,2)×(0,1,1)₁₂模型,Ljung-Box 检验结果为 $Q=19.494, P=0.244$,模型拟合效果良好,与 2018 年 5-7 月实际发病人数比较,实际值均在预测值 95%可信区间内,相对误差的平均值为-0.147。**结论** SARIMA(0,1,2)×(0,1,1)₁₂模型可以对苏州市细菌性痢疾月发病人数进行较好的预测。

关键词: 细菌性痢疾;季节性差分自回归滑动平均模型;预测

中图分类号:R516.4 文献标识码:A 文章编号:1006-3110(2019)06-0656-03 DOI:10.3969/j.issn.1006-3110.2019.06.005

Application of seasonal auto-regressive integrated moving average model to predicting the incidence of bacillary dysentery in Suzhou City

WANG Jian-shu, LIU Qiang, QIN Jiang-chun, HANG Hui, YANG Hai-bing
Suzhou Center for Disease Control and Prevention, Suzhou, Jiangsu 215004, China
Corresponding author: YANG Hai-bing, E-mail: yhb111@163.com

Abstract: **Objective** To explore the feasibility of application of seasonal auto-regressive integrated moving average (SARIMA) model in predicting the monthly number of bacillary dysentery in Suzhou City. **Methods** R i386 3.2.3 software was used to establish SARIMA model based on the data regarding the monthly number of bacillary dysentery in Suzhou City from January 2005 to April 2018. The monthly number of cases of bacillary dysentery in Suzhou City from May to July in 2018 was forecasted, and the prediction effect was evaluated. **Results** The model of SARIMA (0, 1, 2)×(0, 1, 1)₁₂ was established. Ljung-Box test showed that the prediction results with the model accorded well with the actual data ($Q=19.194, P=0.244$), all the actual values of monthly number of cases of bacillary dysentery in Suzhou City from May to July in 2018 fell in the 95% confidence intervals of expected values, and the mean relative error was -0.147. **Conclusion** The SARIMA (0, 1, 2)×(0, 1, 1)₁₂ model is well fit for predicting the monthly number of cases of bacillary dysentery in Suzhou City.

Key words: bacillary dysentery; seasonal auto-regressive integrated moving average model; prediction

细菌性痢疾作为一种常见的介水传染病,是法定的乙类传染病,在苏州市介水传染病的发病中一直处于较高水平,其对公众健康的影响不容忽视^[1]。细菌

基金项目:江苏省卫生计生委科研课题(Y2015021);苏州市科学技术局课题(SYS201582)

作者简介:王建树(1985-),男,硕士,主管医师,研究方向:环境与健康。

通信作者:杨海兵, E-mail: yhb111@163.com。

性痢疾的发病具有一定的季节特征,一般以夏秋季节多见。季节性差分自回归求和滑动平均(seasonal auto-regressive integrated moving average, SARIMA)模型,基于线性模型估计,是疾病发病预测的重要时间序列模型之一,在介水传染病的发病中已有应用^[2-3]。

为探索苏州市细菌性痢疾的流行特征,本文采用 R i386 3.2.3 软件,基于 SARIMA 模型对苏州市 2005 年 1 月-2018 年 4 月细菌性痢疾发病情况进行拟合分

- [13] 陈满连,蔡木蔚,李笑梅,等. 长期低剂量电离辐射对放射工作人员甲状腺功能的影响[J]. 现代诊断与治疗, 2017, 28(11):2056-2057.
- [14] 钱小莲. 南京市 2 019 名放射工作人员甲状腺激素水平分析[J]. 中国辐射卫生, 2017, 26(1):52-54.
- [15] 王娜,梁婧,罗环. 放射医护人员甲状腺 FT3、FT4 和 TSH 水平分析[J]. 职业与健康, 2015, 31(24):3498-3500.
- [16] 伍岳,梁婧,夏春娟,等. 甲状腺功能检查在医学放射工作人员职业健康检查中的应用价值[J]. 职业与健康, 2014, 30(19):2709-2712.

- [17] 黄敏,黄志红,钟晓红,等. 血清 TH 水平与甲状腺功能异常患者关系的探讨[J]. 实用预防医学, 2016, 23(7):876-878.
- [18] 龚婧婧,黄海涛,米其林,等. 微核技术研究进展[J]. 生物技术通报, 2012,28(3):49-56.
- [19] 牛丽梅,刘刚,雷红玉,等. 2011-2012 年某铀矿放射工作人员健康状况调查[J]. 疾病预防控制中心通报, 2014,29(1):42-43.
- [20] 马子建,杨晓兰. 300 例放射工作人员淋巴细胞微核及染色体畸变情况调查[J]. 中国辐射卫生, 2011, 20(2):192-193.

收稿日期:2018-10-30

析,并使用 2018 年 5-7 月份细菌性痢疾月发病数对模型预测效果进行评价,探索该模型在苏州市细菌性痢疾月发病数预测中的运用前景,为细菌性痢疾等介水传染病的防控预警提供一定借鉴。

1 资料与方法

1.1 资料 2005 年 1 月-2018 年 7 月苏州市细菌性痢疾月发病数来自传染病报告系统。其中 2005 年 1 月-2018 年 4 月细菌性痢疾月发病人数用于模型拟合,2018 年 5 月-2018 年 7 月细菌性痢疾月发病人数用于模型有效性的验证。

1.2 建立 SARIMA 模型^[4-6]

1.2.1 数据处理 SARIMA 模型一般表示为 $SARIMA(p,d,q) \times (P,D,Q)_s$,是在 ARIMA 模型的基础上增加了季节性的分析,其建立的前提条件之一是时间序列的平稳性,通过对 2005 年 1 月-2018 年 4 月苏州市细菌性痢疾月发病数数据序列进行平稳性识别,判断其趋势性变化,可进一步通过差分和数据变换进行数据的平稳化预处理。

1.2.2 模型建立与识别 通过转换后序列的样本的自相关系数(autocorrelation function, ACF)和偏自相关系数(partial autocorrelation function, PACF)对拟合模型阶数进行前期判断,初步确定模型的阶数。

1.2.3 参数估计及模型诊断 采用最大似然估计对初步估计的模型进行检验。根据最小信息量准则(Akaike information criterion, AIC)等对比各拟合模型的优劣,同时根据 Q 统计量的结果对残差序列是否为白噪声序列来对模型进行诊断。

1.2.4 序列预测 模型选定后,观察模型的拟合效果,利用拟合的模型对 2018 年 5-7 月苏州市细菌性痢疾月发病数进行预测。

1.3 统计分析 运用 R i386 3.2.3 软件程序包中的“tseries”和“forecast”软件包进行数据处理和预测,检验水准 $\alpha = 0.05$ 。

2 结果

2.1 细菌性痢疾月发病情况 用“plot”函数绘制 2005 年 1 月-2018 年 4 月苏州市细菌性痢疾月发病数的时间序列图(图 1),可见细菌性痢疾月发病数呈现一定的季节性趋势,发病的高峰主要集中在夏秋季的 8-10 月份,提示苏州市细菌性痢疾的月发病数具有一定的季节性周期,为非平稳序列。

2.2 模型的识别与建立 由图 1 看出,数据较为离散,为获得平稳的序列,先对原始数据进行了对数转

换,然后进行一阶普通差分和一阶季节性差分,来消除趋势和季节的影响,使序列平稳,两次差分后的 ACF 图见图 2、PACF 图见图 3,可见拆分后序列基本平稳,没有明显的周期性,符合数据平稳性的要求,由于进行了一次普通差分和周期为 12 的季节性差分,因此 d 取值 1, D 取值 1,季节性周期 s 取 12。 p, q, P, Q 的阶数通常不超过 2 阶,同时,基于 AIC 或者 BIC 准则,运用 R 软件的 auto.arima 函数可以得出推荐模型,在参数检验和相关模型诊断的基础上完成模型的筛选。本文利用 R 软件初步筛选出最优预测模型 $SARIMA(0,1,2) \times (0,1,1)_{12}$ 。

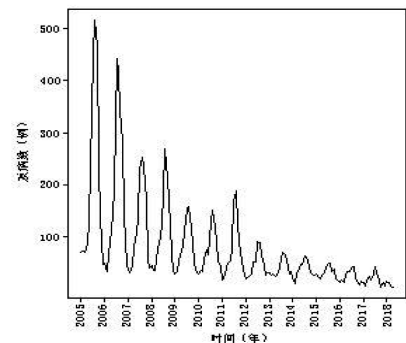


图 1 2005 年 1 月-2018 年 4 月苏州市细菌性痢疾月发病情况

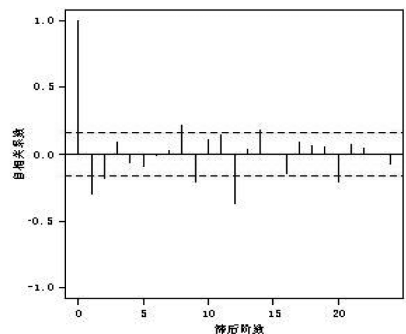


图 2 拆分后数据序列自相关图

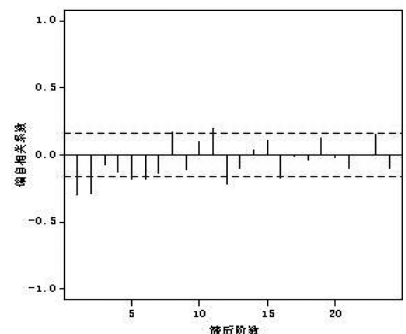


图 3 拆分后数据序列偏自相关图

2.3 模型参数估计与诊断 采用最大似然估计(maximum likelihood estimation, MLE)对初步定阶后模型进行参数估计,其结果差异均有统计学意义,见表 1。

表 1 模型参数估计及检验

参数	估计值	标准误	t 值	P 值
MA1	-0.637	0.081	-7.852	0.000
MA2	-0.290	0.078	-3.698	0.000
SMA1	-0.573	0.080	-7.157	0.000

模型 SARIMA(0,1,2)×(0,1,1)₁₂ 的对数似然值 0.098 6, 对数似然值为 -41.91, aic 值 = 91.82。用“Box. test”语句对模型的残差进行 Ljung-Box 检验, 结果显示 $P>0.05$, 差异无统计学意义 ($Q=19.494, P=0.244$), 表明残差序列是白噪声。综合残差检验和参数检验的结果表明模型对序列相关的信息提取充分。

2.4 模型的预测及效果分析 根据建立的 SARIMA(0,1,2)×(0,1,1)₁₂ 模型用“forecast”函数包中“forecast. Arima”函数对苏州市 2018 年 5-7 月细菌性痢疾的月发病人数进行预测, 预测结果见表 2。结果显示, 预测月发病人数的趋势与实际情况基本一致, 预测值与实际值较为接近, 其平均绝对误差为 -3, 平均相对误差为 -0.147, 实际值都在预测拟合值的 95% 的可信区间之内, 模型预测效果较好。

表 2 2018 年 5-7 月苏州市细菌性痢疾月发病数预测结果

月份	预测	95%可信区间		实际	绝对误差	相对误差
	发病数	下限	上限	发病数		
5	11	6	20	17	-6	-0.353
6	19	10	37	20	-1	-0.050
7	25	13	49	26	-1	-0.038

3 讨论

细菌性痢疾等介水传染病对居民健康具有重要的影响, 通过发病情况预测, 可以在早期预测发病的趋势, 有利于做到提前预警, 及时发现疾病暴发的可能性。已经有学者运用相关统计模型对不同地区细菌性痢疾发病进行了预测, 不同模型其适用性也不尽相同, 一些时间序列分析模型如回归分析等, 在模型构建过程中纳入了多种因变量因素的作用, 其预测精度受到一定影响^[7-9]。而 SARIMA 模型利用时间序列数据的线性组合, 通过综合时间序列变化特点而不用考虑疾病的其他影响因素, 具有模型相对简单、预测精度较高等优点^[10-11], SARIMA 模型在结合了 ARIMA 模型的基础上整合了季节趋势, 对于细菌性痢疾等具有季节周期性特点的发病预测具有良好的适用性。

本文基于 R 软件及相关软件包, 利用苏州市 2005-2018 年细菌性痢疾月发病人数进行模拟分析, 构建了适用于苏州市的 SARIMA(0,1,2)×(0,1,1)₁₂ 模型, 通过模型对 2018 年 5-7 月份苏州市细菌性痢疾月发病人数进行了预测。从模型预测结果来看, 预测

值动态趋势与实际发病数据基本一致, 2018 年 5-7 月份各月的发病预测数值虽然与实际发病有一定差异, 但实际值均在发病预测数值 95% 的可信区间内, 其平均相对误差为 -0.147, 表明用该模型进行苏州市细菌性痢疾的月发病人数预测具有良好的可行性。但通过模型预测的细菌性痢疾月发病人数均低于实际发病人数, 因此在运用 SARIMA(0,1,2)×(0,1,1)₁₂ 模型对苏州市细菌性痢疾的月发病人数预测时, 应考虑对发病人数低估的效应。

细菌性菌痢发病的影响因素较多, 其发病具有一定的地区差异性, 而预测模型的构建也是一个动态调整的过程。本文以苏州市细菌性痢疾的月发病数为基础构建了 SARIMA(0,1,2)×(0,1,1)₁₂ 模型, 可以为苏州市细菌性痢疾的防控提供一定的参考。由于 ARIMA 模型通常适用于短期预测^[12], 因此, 需要在加入新数据的基础上, 及时对模型进行调整。同时, 今后可以通过尝试借助与其他非线性模型如自回归神经网络模型组合, 增强模型的适用性和预测精度。新的卫生政策的颁布以及突发公共卫生事件等也会对传染病的发病产生影响^[13], 因此在对细菌性痢疾等介水传染病进行发病预测时, 在借助于构建的预测模型基础上, 进一步结合卫生政策及突发公共卫生事件等因素的影响, 以提高对细菌性痢疾等介水传染病发病预测的准确性。

参考文献

[1] 袁志平, 李建华, 姚玉斌, 等. 赣州市农村饮用水与介水传染病相关性分析[J]. 环境与健康杂志, 2017, 34(10):885-888.

[2] 王建华, 刘强, 覃江纯, 等. 基于 ARIMA 乘积季节模型的苏州市介水传染病发病预测研究[J]. 环境卫生学杂志, 2017, 7(6):417-420.

[3] 郑磊, 刘德坚, 许贤. ARIMA 模型与 GM(1,1)模型在细菌性痢疾发病率预测中的比较研究[J]. 实用预防医学, 2015, 22(3):365-367.

[4] 刘雷, 张连生, 汤恒, 等. ARIMA 乘积季节模型在丙肝发病预测中的应用[J]. 中华疾病控制杂志, 2014, 18(4):366-367.

[5] Sato RC. Disease management with ARIMA model in time series [J]. Einstein (Sao Pau 10), 2013, 11(1):128-131.

[6] 孟凡东, 吴迪, 隋承光, 等. 2004-2015 年中国狂犬病发病数据 ARIMA 乘积季节模型的建立及预测[J]. 中国卫生统计, 2016, 33(3):389-391, 395.

[7] 石雷. 细菌性痢疾月发病率 ARIMA 季节模型预测分析[J]. 中国公共卫生, 2014, 30(9):1234-1235.

[8] 关静, 张燕, 宋静, 等. ARIMA 模型在北京市西城区细菌性痢疾发病预测中的应用[J]. 职业与健康, 2015, 31(23):3243-3245, 3248.

[9] 张斯冬, 张建陶, 钱建东, 等. 常州市细菌性痢疾发病趋势灰色模型 GM(1,1) 预测[J]. 实用预防医学, 2013, 20(3):310-311, 299.

[10] 陈玲, 程丽君, 赵向军. 恶性肿瘤住院量与住院费用的 ARIMA 乘积季节模型预测研究[J]. 中国卫生统计, 2017, 34(4):554-557.

[11] 原凌云, 周以军, 朱妮, 等. 多种数据模型在手足口病发病预测的应用探讨[J]. 实用预防医学, 2018, 25(11):1400-1402.

[12] 王永斌, 李向文, 柴峰, 等. ARIMA 模型在我国梅毒发病率预测中的应用[J]. 现代预防医学, 2015, 42(3):385-388, 417.

[13] 王橙, 许沛尧, 马爱军, 等. 基于 ARIMA 模型对传染病监测数据异常点的探测研究[J]. 现代预防医学, 2018, 45(4):577-581, 590.