

基于季节性 ARIMA 模型的新疆肺结核发病预测分析

聂艳武¹, 郑彦玲², 孙亚红¹, 杨磊³, 张利萍²

1. 新疆医科大学公共卫生学院省部共建中亚高发病因与防治国家重点实验室, 新疆 乌鲁木齐 830011;
2. 新疆医科大学医学工程技术学院, 新疆 乌鲁木齐 830011; 3. 新疆医科大学护理学院, 新疆 乌鲁木齐 830011

摘要: **目的** 探讨季节性时间序列模型 (autoregressive integrated moving average, ARIMA) 在新疆肺结核发病预测中的应用, 并验证模型的可行性和适用性。 **方法** 采用季节性 ARIMA(p, d, q)(P, D, Q) 拟合 2005 年 1 月—2019 年 8 月新疆地区肺结核月发病人数, 建立多个季节时间序列模型并进行比较, 选出最优模型对 2019 年 9—12 月肺结核发病人数进行预测。 **结果** 2005 年 1 月—2019 年 8 月新疆地区肺结核累积发病人数为 627 869 例, 年平均发病人数为 3 567 例。新疆地区肺结核月发病数具有季节性, 1—5 月平均发病数高于平均水平, 6—12 月平均发病数低于平均水平, 发病高峰为 1 月和 3 月, 发病低谷为 9 月。通过赤池信息量 (Akaike Information Criterion, AIC) 和贝叶斯信息量 (Bayesian Information Criterion, BIC) 最小原则得出, ARIMA(1, 1, 1)(0, 1, 2)₁₂ 是最优模型, 其残差序列为白噪声, 参数的回归系数均具有统计学意义, 拟合的平均绝对百分比误差 MAPE 为 8.723%。预测的 MAPE 为 18.674%, 真实值均处于预测值的 95% 置信区间内。 **结论** ARIMA(1, 1, 1)(0, 1, 2)₁₂ 模型能够较好地拟合新疆肺结核发病数据, 并进行短期预测, 对新疆卫生防控措施的制定具有一定指导意义。

关键词: 肺结核; ARIMA; 时间序列; 预测

中图分类号: R521 **文献标识码:** A **文章编号:** 1006-3110(2021)11-1324-05 **DOI:** 10.3969/j.issn.1006-3110.2021.11.011

Prediction of pulmonary tuberculosis incidence in Xinjiang based on seasonal ARIMA model

NIE Yan-wu¹, ZHENG Yan-ling², SUN Ya-hong¹, YANG Lei³, ZHANG Li-ping²

1. State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, School of Public Health, Xinjiang Medical University, Urumqi, Xinjiang 830011, China;
2. School of Medical Engineering and Technology, Xinjiang Medical University, Xinjiang, Urumqi 830011, China;
3. School of Nursing, Xinjiang Medical University, Xinjiang, Urumqi 830011, China
Corresponding author: ZHANG Li-ping, E-mail: zhanglp1219@163.com

Abstract: **Objective** To explore the application of seasonal autoregressive integrated moving average (ARIMA) model to prediction of pulmonary tuberculosis incidence in Xinjiang, and to verify the feasibility and applicability of the model. **Methods** Seasonal ARIMA (p, d, q)(P, D, Q) was used to fit the monthly incidence of pulmonary tuberculosis in Xinjiang from January 2005 to August 2019. Multiple seasonal time series models were established and compared to select the optimal model to predict the incidence of pulmonary tuberculosis from September to December 2019. **Results** From January 2005 to August 2019, the cumulative incidence of pulmonary tuberculosis in Xinjiang was 627,869 cases, with an average annual incidence of 3,567 cases. The monthly incidence of pulmonary tuberculosis in Xinjiang showed a seasonal pattern. The average incidence from January to May was higher than the average level, while the average incidence from June to December was lower than the average level. The incidence peak was in January and March, whereas the incidence was found to be lower in September. According to the minimum principle of Akaike Information Criterion and Bayesian Information Criterion, ARIMA (1, 1, 1)(0, 1, 2)₁₂ was the optimal model, the residual sequence was white noise. The regression coefficients of parameters were statistically significant, and the average absolute percentage error of fitting was 8.723%. The predicted mean absolute percentage error was 18.674%, and the real values were within the 95% confidence interval of the predicted values. **Conclusion** ARIMA (1, 1, 1)(0, 1, 2)₁₂ model can better fit the incidence data of pulmonary tuberculosis in Xinjiang and make short-term prediction, which has a certain guiding significance for the formulation of health prevention and control measures in Xinjiang.

Keywords: pulmonary tuberculosis; autoregressive integrated moving average; time series; prediction

基金项目: 省部共建中亚高发病因与防治国家重点实验室开放课题资助项目 (SKL-HIDCA-2020-9)

作者简介: 聂艳武 (1998-), 男, 河南周口人, 硕士研究生, 研究方向: 流行病与卫生统计学。

通信作者: 张利萍, E-mail: zhanglp1219@163.com。

结核病是由结核分枝杆菌所引起的慢性传染病,全身的器官都可发生,但以肺结核最为常见,约占结核病人的 80%。肺结核是一种感染性较强的疾病,一旦出现传染源,若不加以控制干预,极易暴发流行^[1-2]。因此,肺结核具有较高的防治需求,各地区需加强对其的流行病学分析^[3]。近年来,我国加大了对肺结核防控的力度,伴随着医疗卫生事业的发展和医疗条件的改善,结核疫情得到了有效的控制。但统计数据表明,个别地区肺结核报告发病率仍呈现缓慢上升趋势。2019 年 WHO 结核病报告显示,结核病是导致死亡的十大病因之一,我国肺结核报告发病数较多,是全球肺结核排行第二的国家,耐药肺结核新发病例占全球的 14%^[4]。新疆是一个多民族居住的地区,地处我国的西部,受到气候条件、生活习惯、经济发展水平、医疗卫生条件等方面的限制,结核病的高发对新疆人民的生命健康安全造成严重的危害。数据显示,2008—2018 年新疆地区肺结核报告发病率总体呈上升趋势(202.93/10 万~304.94/10 万),远远高出全国平均发病水平(61/10 万),新疆地区肺结核的预防控制形势严峻。本研究通过对新疆地区 2005—2019 年肺结核的流行病学特征及疫情变化情况进行分析,构建肺结核季节性 ARIMA 模型,并对 2019 年 9—12 月发病人数进行预测,以期为该地区肺结核的预防与治疗提供理论依据。

1 数据与方法

1.1 数据来源 本研究基于 2005—2019 年新疆肺结核报告月发病例数。其中 2005—2017 年数据来源于公共卫生科学数据中心(<http://www.phsciencedata.cn/Share/>),2018—2019 年数据来源于新疆维吾尔自治区卫生健康委员会法定传染病开放数据。

1.2 模型简介 ARIMA 模型是时间序列分析中重要的组成部分,分析过程简便,是其他预测方法不可替代的^[5]。基于 R 语言的时间序列常用的分析步骤为:(1)将一组同时间相关的数据转换为时间序列,一般要求数据量 50 个以上较好^[6]。(2)对时间序列作平稳性检验和白噪声检验^[7]。平稳性检验通常绘制时序图或用单位根检验,非平稳时间序列通过差分使序列平稳化。用 Ljung-Box 方法进行白噪声检验,显著性水平 $\alpha=0.05$ 。(3)根据数据检验结果选择合适的时间序列建模方法,对模型定阶,需要选择合适的阶数。(4)选取常用的统计指标对模型精度进行检验并选择最优模型,通常比较模型的赤池信息量(Akaike Information Criterion, AIC)和贝叶斯信息量(Bayesian

Information Criterion, BIC)水平,根据最小信息量准则来判断。(5)时间序列趋势的预测,模型的拟合和预测效果通过平均绝对百分比误差来判断。

1.3 季节指数 季节指数用来验证序列是否具有季节性,能判断出疾病的好发月份^[8]。计算公式为: $S_j = \frac{\bar{x}_j}{\bar{x}}$,公式中 \bar{x} 为肺结核的月平均发病数, \bar{x}_j 为第 j 个月

肺结核的平均发病数, S_j 表示第 j 个月的季节指数。若 $S_j=1$,表明该月无季节效应,若 $S_j>1$,表明当月的平均发病数高于平均水平,若 $S_j<1$,当月的平均发病数低于平均水平。

1.4 统计学分析 数据由 2 位人员使用 Excel 2007 对数据进行录入、核查以及处理,确保结果的准确性。使用 R-4.0.2 进行时间序列建模。

2 结果

2.1 新疆地区肺结核发病情况 2005 年 1 月—2019 年 8 月新疆地区肺结核累积发病人数为 627 869 例,年均发病人数为 3 567 例。由时序图 1 可知,肺结核发病规律呈现明显的季节特征,并且于 2010—2019 年呈现缓慢增长趋势,值得一提的是 2017、2018 年平均发病人数分别高达 5 007、4 535 例。

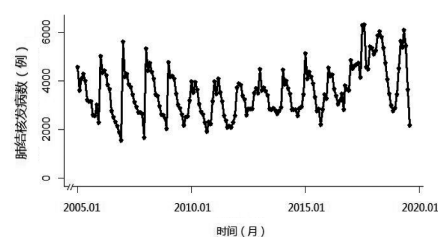


图 1 2005 年 1 月—2019 年 8 月新疆地区肺结核月报告发病数时序图

2.2 基于新疆肺结核月发病数据建立 ARIMA 模型

2.2.1 平稳性和季节性分析 基于 KPSS 检验的 $ndiffs()$ 函数可判断出,需要对时间序列进行一阶差分。对数据做 1~12 阶白噪声检验,结果均显示 $P<0.05$,表明原始序列为非白噪声,具有研究意义。图 2 中 observed、trend、seasonal、random 部分分别代表时序图、季节效应图、趋势图和随机波动项。趋势图表明近年肺结核发病数有所下降,但仍处于较高水平。通过季节效应图得出新疆肺结核具有季节性趋势,提示需做季节差分来消除数据的季节性影响。因此对原始数据进行一阶差分和一阶季节差分,处理后数据经单位根检验有统计学意义($t=-6.667, P<0.01$),认为数据平稳。通过季节指数对季节性进行定量分析,具体结果见表 1,季节指数图见图 3。结果显示新疆肺结核发

病数季节效应明显,以一年为周期出现 1 月和 3 月两个高峰,9 月为一个低谷,并且 1—5 月季节指数大于 1,其余月份季节指数小于 1。表明 1—5 月的平均发病数高于平均水平,6—12 月的平均发病数低于平均水平,呈现出季节效应。

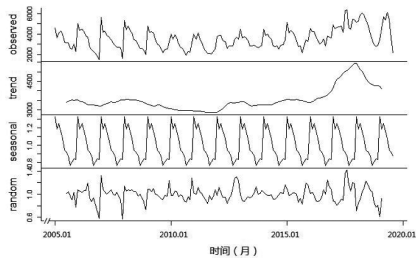


图 2 新疆地区 2005 年 1 月—2019 年 8 月肺结核季节效应分解

表 1 新疆肺结核每月报告数季节指数

时间(月)	平均月报告数(\bar{x}_j)	季节指数(S_j)
1	4 606	1.303 31
2	4 118	1.165 29
3	4 413	1.248 57
4	4 181	1.183 02
5	3 965	1.121 81
6	3 452	0.976 77
7	3 301	0.933 89
8	3 074	0.869 83
9	2 702	0.764 39
10	2 865	0.810 68
11	2 989	0.845 70
12	2 961	0.837 74

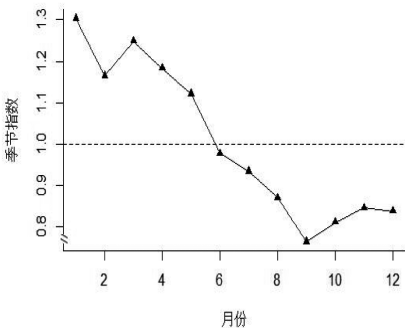


图 3 新疆地区肺结核月报告数季节指数图

2.2.2 ARIMA 乘积季节模型识别 因原始数据进行一阶差分和一阶季节差分处理之后达到平稳状态,所以模型 $ARIMA(p, d, q)(P, D, Q)_s$ 中 $d=1, D=1$ 。根据平稳时间序列绘制出自相关图和偏自相关图。

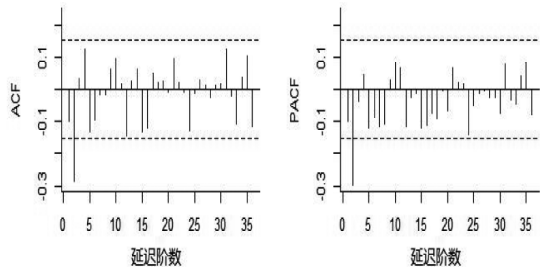


图 4 平稳时间序列的自相关图和偏自相关图

如图 4 所示,自相关图和偏自相关图均是二阶截尾。考虑到模型阶数过高将造成过拟合,因此参数取值范围为 0~2,对参数从低阶到高阶进行尝试,筛选出通过参数检验的模型,并结合实际情况,选取 $AIC = 2\,503$ 为临界值,排除 $AIC > 2\,503$ 的模型,最终列举 11 个模型的 AIC、BIC 值(表 2)。采用最小信息量准则对上述 11 个模型进行判断,具体为 AIC、BIC 的值越小模型的拟合优度越高来建立最优模型。确定了最优模型为 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$,其 $AIC = 2\,499.033$, $BIC = 2\,514.502$ 。

表 2 模型的 AIC、BIC 比较

模型	AIC	BIC
$ARIMA(0, 1, 2)(0, 1, 2)_{12}$	2 502.377	2 517.846
$ARIMA(1, 1, 1)(0, 1, 2)_{12}$	2 499.033	2 514.502
$ARIMA(1, 1, 1)(1, 1, 1)_{12}$	2 500.699	2 516.167
$ARIMA(1, 1, 1)(1, 1, 2)_{12}$	2 500.527	2 519.09
$ARIMA(1, 1, 1)(2, 1, 1)_{12}$	2 499.798	2 518.361
$ARIMA(1, 1, 1)(2, 1, 2)_{12}$	2 501.092	2 522.748
$ARIMA(1, 1, 2)(0, 1, 2)_{12}$	2 500.897	2 519.459
$ARIMA(1, 1, 2)(1, 1, 1)_{12}$	2 502.405	2 520.967
$ARIMA(1, 1, 2)(1, 1, 2)_{12}$	2 502.352	2 524.008
$ARIMA(2, 1, 2)(0, 1, 2)_{12}$	2 501.532	2 523.188
$ARIMA(2, 1, 2)(1, 1, 2)_{12}$	2 502.968	2 527.718

2.2.3 模型参数估计与检验 采用最大似然估计法来估计参数,结果表明参数具有统计学意义。证明 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$ 可以用来预测新疆肺结核的发病率(详见表 3)。利用最优模型 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$,得到实际值和预测值之差即残差,对模型残差进行白噪声检验,模型的 LB 检验统计量显示 $P=0.264(P>0.05)$,可认为残差序列为白噪声,模型提取较完整,此模型来预测新疆肺结核的发病人数是合理的。对残差进行自相关和偏自相关分析显示,自相关系数都落在区间内,因此不能认为残差序列各数值间具有相关性(图 4),模型拟合较好。

表 3 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$ 模型的参数估计与检验

参数	估计值	标准误差	t 值	P 值
AR1	0.745	0.056	13.304	<0.0001
MA1	-1	0.148	-6.757	<0.0001
SMA1	-0.322	0.085	-3.788	<0.0005
SMA2	-0.308	0.091	-3.385	<0.0005

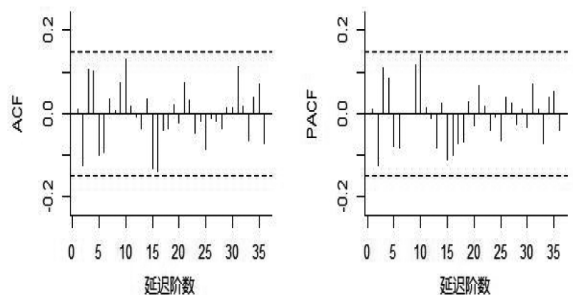


图 5 残差的自相关和偏自相关函数图

2.2.4 模型预测 采用 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$ 模型拟合新疆 2005 年 1 月—2019 年 8 月期间肺结核发病数(图 5)。模型拟合的精度通过平均百分比误差来评估。根据计算公式得出 $MAPE = 8.723\%$, 表明模型拟合效果较好。根据最优模型预测 2019 年 9—12 月的肺结核报告发病数并给出 $95\%CI$ (表 4)。由表 4 可以看出,2019 年 9—12 月新疆地区肺结核月发病人数真实值与预测值稍有不同,但都落在 95% 置信区间中,模型预测的 $MAPE = 18.674\%$, 表明模型预测尚可。查阅中华人民共和国交通运输部得知,2019 年 12 月份公路旅客运输量为去年同期的 90.4% , 引起真实值略微降低,不符合季节指数变化规律,且真实值远低于 12 月平均发病人数(2 961),这可能是引起 12 月份相对误差较大的原因。

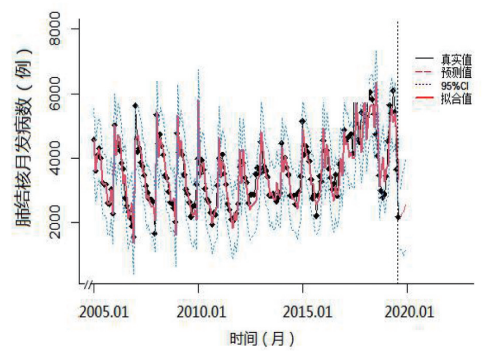


图 6 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$ 模型拟合及预测结果

表 4 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$ 模型预测
新疆地区 2019 年 9—12 月期间人群发病数

月份	真实值	预测值	相对误差	UCL	LCL
9 月	2 328	2 111	0.093	3 067	1 155
10 月	2 166	2 313	0.068	3 510	1 117
11 月	2 051	2 295	0.119	3 609	980
12 月	1 737	2 540	0.462	3 926	1 170

3 讨论

肺结核作为一种慢性消耗性疾病,长期危害患者健康,且救治延误预后差,是重大的公共卫生问题。虽然我国加大了对公共卫生的关注与资金的投入,使得

肺结核疫情得到较明显的改善^[9],但目前仍有较多的人感染结核分枝杆菌,同时新疆地区肺结核的发病率长期位于全国前列。相关研究显示,与全国对比,2004—2019 年新疆肺结核的发病率和死亡率总体呈上升趋势,提示新疆需加强结核病的防治措施^[10]。肺结核发病人数变化呈现季节性趋势,同时,季节性 ARIMA 模型能较好地拟合具有趋势性和季节性的时间序列,因此通过构建并运用季节性 ARIMA 模型将很好地拟合出新疆地区肺结核发病人数的变化趋势。

本研究发现新疆地区肺结核月发病数随时间的推移而变化,根据季节性分解图可确定新疆肺结核具有季节性,同时季节指数表明 1—5 月的平均发病数高于平均水平,6—12 月的平均发病数低于平均水平,这与王薇^[11]研究结果一致。其中,1 月份报告发病人数通常最多,在 3 月和 10—12 月也容易出现小幅上升趋势。这是因为,人口流动易增加肺结核患病风险^[12],而新疆地区的流动人口较多,尤其是 1 月份至春节期间,流动人口返乡,为新疆冬季最寒冷时节,更容易造成肺结核发病人数增加。通过自相关图和偏自相关图确定相关参数,并综合比较模型 AIC 和 BIC 指数,选择最优模型,进行模型的参数检验,最终确定模型为 $ARIMA(1, 1, 1)(0, 1, 2)_{12}$,该模型较好地拟合了 2005 年 1 月—2019 年 8 月新疆地区肺结核的流行趋势($MAPE = 8.723\%$),并预测 2019 年 9—12 月新疆地区肺结核月发病人数,得出预测的 $MAPE = 18.674\%$,表明预测结果尚可。

时间序列在短期预测精度较高,但长期预测精度会大大减少^[13]。本研究的预测结果与近年来数据波动有较大关联,导致预测结果出现偏差。2020 年,受到新冠疫情影响,肺结核检测工作未能及时有序开展,导致报告发病数减少,以往的模型用于预测 2020 年数据将会产生较大误差(实际值偏小)。将 2020 年数据纳入模型,会预测得到大幅减少的发病数,这同实际情况是不相符的。因为检测工作制约得到减少的发病数并不能作为新疆结核疫情得到了有效控制的佐证,所以应警惕潜在未检测出的结核病人,警惕延迟的肺结核峰值。因此,本研究并未将 2020 年 1 月以来的发病数纳入单纯的时间序列模型,这部分数据挖掘应综合考虑多方面影响因素。同时已有证据表明,COVID-19 与肺结核病之间在临床、流行病学和防控等方面必然会相互关联、相互影响^[14],疫情期间,肺结核相关防控措施被中断,患者得不到及时诊断与治疗,故在疫情过后,防疫工作人员应加大排查力度,及时采取预防及治疗措施,防止肺结核扩散。