Statistics and Application 统计学与应用, 2022, 11(4), 732-738
Published Online August 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.114077

Hans汉斯

ARIMA乘积季节模型在新疆肺结核发病预测中的应用

姚艳茹

石河子大学，新疆 石河子

收稿日期：2022年7月9日；录用日期：2022年7月20日；发布日期：2022年8月2日

摘　要

目的：根据新疆地区肺结核月发病数的季节性以及趋势性，建立求和自回归移动平均(ARIMA)乘积季节模型，并对新疆肺结核的发病趋势进行预测，调整防控措施。方法：利用R语言以2012年1月至2021年5月新疆地区肺结核每月发病人数为基础，建立并选出最适合的模型，对该地区2021年6月至2022年5月的肺结核发病人数进行一个预测，再将预测值与实际值作对比，以此为标准来讨论这个模型的预测效果。结果：通过赤池信息量(AIC = 46.23)与贝叶斯信息量(BIC = 57.1)最小原则可以得出，ARIMA(1, 1, 1)(1, 0, 0)$_{12}$是最优模型，2012年1月至2021年5月拟合结果，2021年6月至2022年5月模型预测值都落在置信区间95%内。结论：本文建立的ARIMA(1, 1, 1)(1, 0, 0)$_{12}$能较为准确地预测新疆地区肺结核的月发病数。

关键词

ARIMA乘积季节模型，肺结核，月发病数，预测

Application of ARIMA Multiplicative Seasonal Model in the Prediction of Pulmonary Tuberculosis Incidence in Xinjiang

Yanru Yao

Shihezi University, Shihezi Xinjiang

Received: Jul. 9th, 2022; accepted: Jul. 20th, 2022; published: Aug. 2nd, 2022

文章引用: 姚艳茹. ARIMA 乘积季节模型在新疆肺结核发病预测中的应用[J]. 统计学与应用, 2022, 11(4): 732-738.
DOI: 10.12677/sa.2022.114077

Abstract

Objective: According to the seasonality and trend of the monthly incidence of pulmonary tuberculosis in Xinjiang, a Autoregressive Integrated Moving Average (ARIMA) multiplicative seasonal model was established to predict the incidence trend of tuberculosis in Xinjiang and adjust the prevention and control measures. Methods: Based on the monthly incidence of tuberculosis in Xinjiang from January 2012 to May 2021, the most suitable model is established and selected by using R language to predict the incidence of tuberculosis in this region from June 2021 to May 2022. Then, the predicted value is compared with the actual value, and the prediction effect of this model is discussed based on this standard. Results: ARIMA(1, 1, 1)(1, 0, 0)$_{12}$ can be obtained by the principle of minimum Akachi information (AIC = 46.23) and Bayesian information (BIC = 57.1) is the optimal model. The fitting results from January 2012 to May 2021, and the predicted values of the model from June 2021 to May 2022 fall within the 95% confidence interval. Conclusions: ARIMA(1, 1, 1)(1, 0, 0)$_{12}$ established in this paper can accurately predict the monthly incidence of tuberculosis in Xinjiang.

Keywords

ARIMA Multiplicative Seasonal Model, Pulmonary Tuberculosis, Monthly Incidence, Forecast

1. 引言

结核病是由结核分枝杆菌感染身体引起的慢性传染病。在我国每年都有大量的病例报告，也一直被列为全球公共卫生问题。2022 年在新疆维吾尔自治区法定报告传染病疫情中显示，肺结核的报告病例数居新疆地区前五位。因此，为了减轻我们国家在该病上的投入，预测肺结核发病数就是一个很好的方法。求和自回归移动乘积季节模型作为经典的季节性时间序列预测模型，已应用于传染病及其他学科[1] [2] [3] [4]，实践证明预测效果理想。因此本研究拟用新疆维吾尔族自治区 2012 年 1 月份到 2022 年 5 月份结核发病数进行拟合和预测，为新疆地区的疫情防控政策提供科学依据。

2. 资料与方法

2.1. 数据来源

数据来自公共卫生科学数据中心和新疆维吾尔自治区卫生健康委员会－疾病监测与评价，统计 2012 年 1 月~2022 年 5 月报告的新疆地区肺结核月发病数。

2012 年 1 月~2021 年 5 月的数据用于建立 ARIMA 乘积季节预测模型，2021 年 6 月~2022 年 5 月数据用于模型的预测效果检验。

2.2. 研究方法

ARIMA 模型是用来拟合稳定的，或者是对数转换和差分处理后的稳定的时间序列，序列的属性不会随着时间的流逝而改变[5]。具有季节性因子的时间序列资料可以分为趋势因子、季节性因子和随机性因

子(错误因子) ARIMA 乘积季节模型表达式为 ARIMA$(p, d, q)\times(P, D, Q)$，参数 p, P 和 q, Q 表示自回归，移动平均阶数，D, d 为差分次数，S 为循环长度。

ARIMA 乘积季节模型建模过程包括平稳性检验、模型识别、模型诊断、模型预测[6]，肺结核月发病数经过对数转换后，利用 R 语言中的 forecast 和 tseries 等包，对新疆地区 2012 年~2021 年的月发病数据建立模型，在使用 R4.2.0 中的 auto.arima()代码拟合出最优模型，并对 2021 年 6 月~2022 年 5 月的月发病数进行预测，与真实值比较，评价模型预测效果。检验水平 $\alpha = 0.05$。

3. 结果

3.1. 新疆地区肺结核月发病情况

2012 年 1 月至 2022 年 5 月新疆维吾尔自治区肺结核的累积发病人数是 728,886 例。根据 2012 年 1 月至 2022 年 5 月新疆地区肺结核月发病数作出时间序列图(图 1)和经过自然对数转化后的时间序列图(图 2)。通过时序图 1 和图 2 可以看出，新疆地区 2012 年 1 月~2021 年 5 月肺结核月发病数据中，每一年 1~2 月份是发病高峰期，此后逐月下降，故有明显的季节性趋势，且呈现年周期性波动。因此，对取对数后的数据作趋势分解图(图 3)。由图 3 可知，2019 年之前年发病人数逐年小幅度增长，但在 2019 年 6~7 月骤增到最高峰值，随后呈现明显降低趋势。这里的数据显然是非平稳的，并且拥有一些季节性的特征，因此我们会考虑到对这些数据进行差分或季节差分，从而使这些数据变得平稳[7]。对取对数后的序列进行 1 阶差分和消除季节性影响后(图 4)，时间序列已趋于平稳。用增广迪基－福勒检验(augmented Dickey-Fuller, ADF)方法检验处理后序列的平稳性，结果显示 $p = 0.01 < \alpha$，表明序列是平稳的。

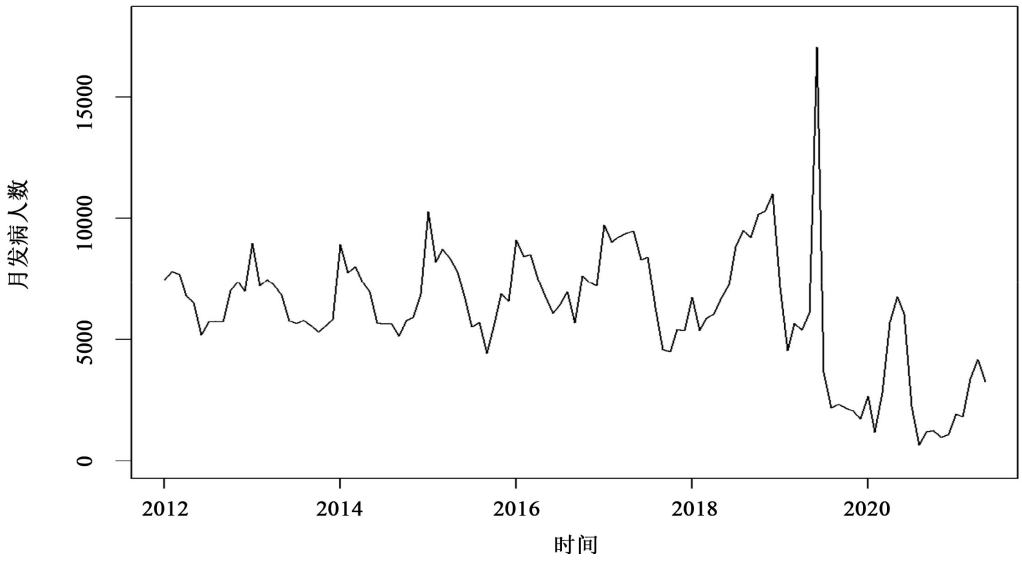

Figure 1. Time series of monthly incidence of tuberculosis in Xinjiang from January 2012 to May 2022
图 1. 2012 年 1 月至 2022 年 5 月新疆地区肺结核月发病数时序图

3.2. 建立 ARIMA 乘积季节模型

利用 R 语言中的 auto.arima()代码自动选取的最优模型为 ARIMA$(1, 1, 1)(1, 0, 0)_{12}$。作出该模型的残差诊断图(图 5)，根据图 5 可知该模型的残差在 0 附近随机波动，且呈现正态分布。经 Ljung-Box 检验后，该模型的残差为白噪声($Q = 26.648, p = 0.1455$)，并且从其自相关图(ACF)来看，该模型的残差大部分都落入置信度为 0.95 的置信区间内，说明已经将时间序列的信息进行了充分的提取。同时，由

$\mathrm{ARIMA}(1,1,1)(1,0,0)_{12}$ 模型的平均绝对百分误差 $\mathrm{MAPE}=2.078<10$，说明模型的预测精度较高。因此该模型拟合新疆地区 2012 年 1 月~2021 年 5 月的肺结核发病数据是合适的。

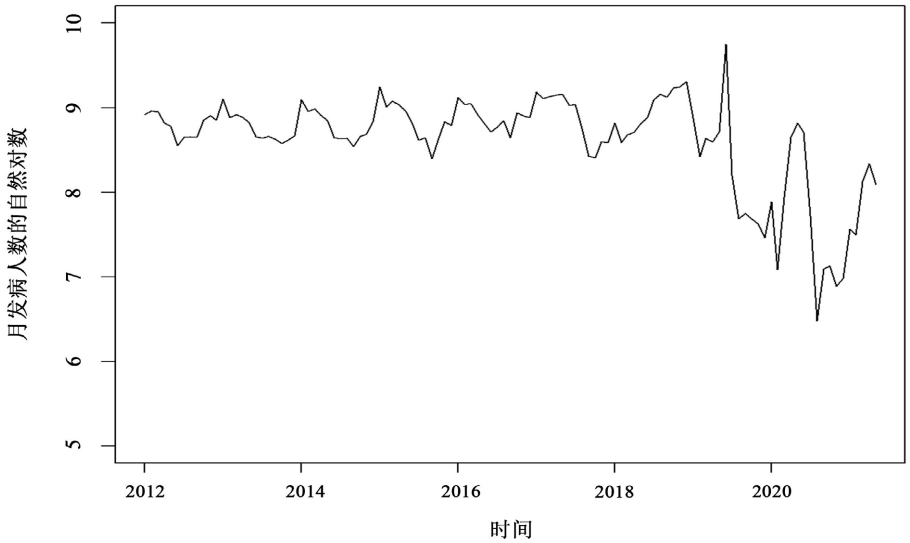

Figure 2. Natural logarithm time series of monthly incidence of tuberculosis in Xinjiang from January 2012 to May 2022
图 2. 2012 年 1 月至 2022 年 5 月新疆地区肺结核月发病数的自然对数时序图

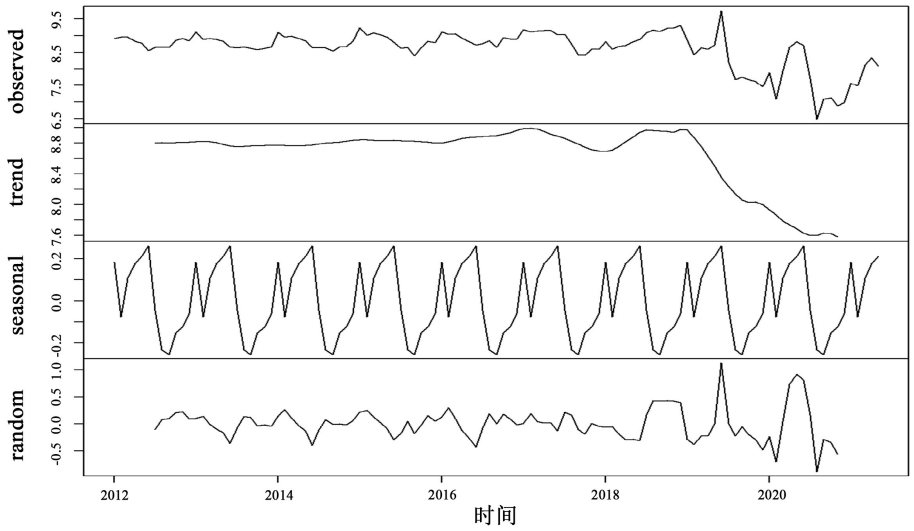

Figure 3. Breakdown of incidence trend of tuberculosis in Xinjiang from January 2012 to May 2022
图 3. 2012 年 1 月至 2022 年 5 月新疆地区肺结核发病趋势分解图

3.3. 模型预测

图 6 中的黑线为真实值，蓝线部分为置信度为 95%的样本后面 12 月的预测值，可以看出预测序列的变化趋势与原序列基本一致。因此模型 $\mathrm{ARIMA}(1,1,1)(1,0,0)_{12}$ 可用于预测 2021 年 6 月至 2022 年 5 月的肺结核流行趋势。利用建立的模型得到预测值，并将其与真实发病人数进行比较，除 2021 年 8 月和 2022 年 3 月的相对误差分别为 38%和 23%，其余月份的相对误差均在 20%以下，而预测值与真实值的相

对误差最小是 2021 年 7 月，仅为 1%。结果表明其平均相对误差百分比等于 12%，说明预测效果较好，见表 1。

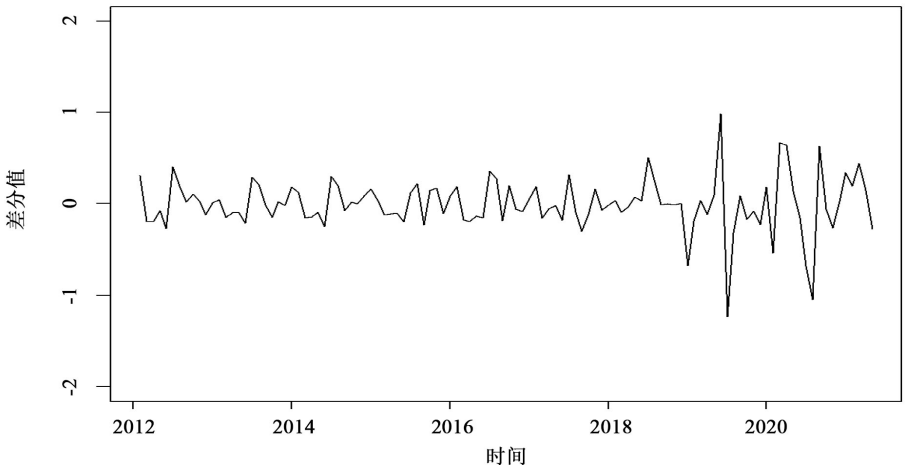

Figure 4. The season of incidence and elimination of tuberculosis in Xinjiang from January 2012 to May 2022 and the first-order difference diagram
图 4. 2012 年 1 月至 2022 年 5 月新疆地区肺结核发病消除季节和 1 阶差分图

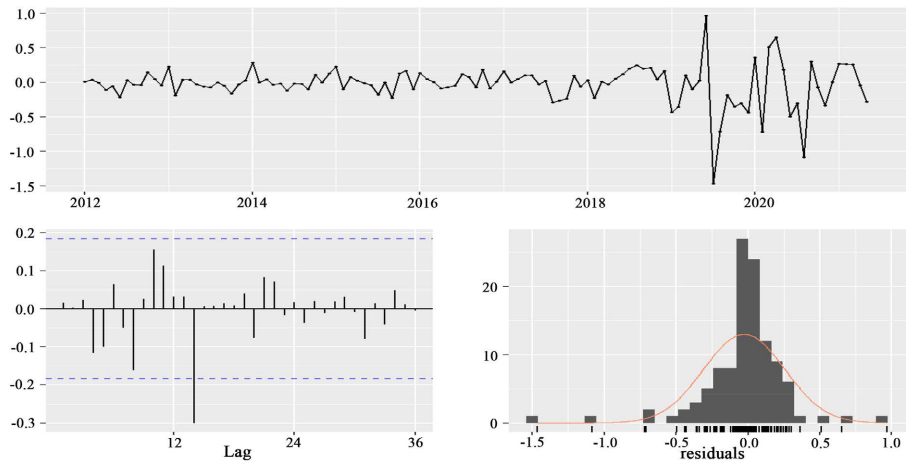

Figure 5. Residual distribution of ARIMA(1, 1, 1)(1, 0, 0)$_{12}$ model
图 5. ARIMA(1, 1, 1)(1, 0, 0)$_{12}$ 模型的残差分布图

4. 讨论

肺结核是一种慢性的疾病，那就意味着肺结核会长期地影响人们的健康，所以它又被国际上列为重大公共卫生问题。进入二十一世纪以来，我们国家在肺结核问题上也是煞费苦心，各方面的投入都在不断加大，也取得了很明显的效果。但是却难以根除，每年的统计中都有很多人感染肺结核，我们所研究的新疆地区其肺结核发病人数在我国各个省份中一直居高不下。

本文利用 R 语言中的 auto.arima()选择最优模型 ARIMA$(1, 1, 1)(1, 0, 0)_{12}$，残差检验为白噪声过程，预测的平均相对误差仅为 12%。结果表明预测值均落在 95%的置信区间内，说明当月的结核病疫情正常，若落在置信区间外，则说明结核病疫情会有大的波动，提示我们应该重视，做好防控预案。因此本文可通过 ARIMA 乘积季节模型对未来结核病发病人数进行早期的预测和预警，能够及早做好防疫措施。

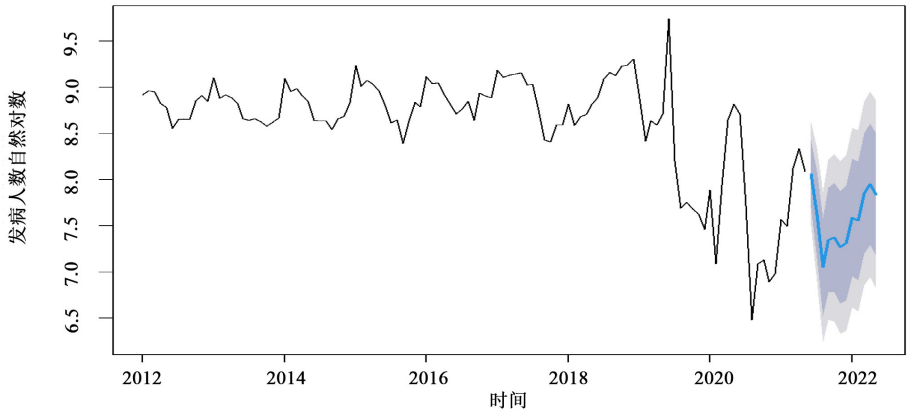

Figure 6. Predicted and true number of tuberculosis cases in Xinjiang
图 6. 新疆地区肺结核发病人数预测值和真实值

Table 1. Comparison between actual value and predicted value in Xinjiang from February 2021 to May 2022
表 1. 2021 年 2 月~2022 年 5 月新疆地区真实值与预测值比较

时间	预测值(例)	真实值(例)	差值	相对误差(%)
2021 年 6 月	3168	2684	484	18
2021 年 7 月	2036	2025	11	1
2021 年 8 月	1160	1870	−710	38
2021 年 9 月	1553	1669	−116	7
2021 年 10 月	1594	1455	139	10
2021 年 11 月	1434	1534	−100	7
2021 年 12 月	1501	1574	−73	5
2022 年 1 月	1976	1789	187	10
2022 年 2 月	1915	1700	215	13
2022 年 3 月	2567	3335	−768	23
2022 年 4 月	2839	3297	−458	14
2022 年 5 月	2532	2619	−87	3

该模型是在历史数据的基础上建立的，但是结核病的发病规律会受到多种因素的影响，如人口、环境和经济等。因此要通过不断地加入新监测数据对模型进行修正，提高模型预测的精确度。ARIMA 乘积季节模型能够较好地预测肺结核的短期流行趋势，为提升防控的预见性和主动性提供依据。

参考文献

[1] 安晓丹, 李晓霞. 基于 ARIMA 模型的运城市 GDP 预测分析[J]. 运城学院学报, 2022, 40(3): 69-73+90.

[2] 杨聪, 彭巨擘, 伍美珍, 张合生. 基 ARIMA 模型的铟电解槽异常预测研究[J]. 仪表技术, 2022(3): 30-34.

[3] 孙梦彩, 周权, 权戴戴, 周海军, 陈清泉, 何青, 张晓阳, 徐幽琼. ARIMA 模型在登革热高发城市发病预测中的应用[J/OL]. 中国预防医学杂志: 1-8. http://kns.cnki.net/kcms/detail/11.4529.R.20220519.1625.008.html, 2022-06-28.

[4] 陈奎, 董晨雪, 卢佳月, 葛国曙. 基于 ARIMA 时间序列模型的我国艾滋病发病人数预测[J]. 中国初级卫生保健, 2022, 36(3): 91-93.

[5] 许明燕. 基于 ARIMA 模型和 BP 神经网络模型的江苏省 GDP 预测分析[D]: [硕士学位论文]. 济南: 山东大学, 2020.

[6] 黄艳华. 乘积季节模型 *ARIMA*(*p*,*d*,*q*)(*P*,*D*,*Q*)*s* 在 CPI 分析中的应用[J]. 重庆工商大学学报(自然科学版), 2016, 33(3): 70-75.

[7] 张蓓蓓, 彭献镇, 王建明, 王欣怡, 于新航. 中国肺结核发病趋势的 ARIMA 乘积季节模型构建[J]. 江苏预防医学, 2021, 32(4): 400-402+408.