DOI:10.12138/j.issn.1671－9638.20205807

·论著·

# 基于 Python 语言的 ARIMA 模型在天津市结核病发病率预测中的应用

张晓卉,姚婷婷,陈　阳,张甜甜,马　骏
(天津医科大学公共卫生学院流行病与卫生统计学系,天津　300041)

[摘　要] **目的**　探讨差分自回归移动平均(ARIMA)模型在结核病发病率预测中的可行性。**方法**　基于 Python 语言的 statsmodels 模块,以天津市 2004 年 1 月—2015 年 12 月结核病月发病率数据作为训练集建立最优季节性差分自回归移动平均(SARIMA)模型,以 2016 年 1—12 月数据对 SARIMA 模型进行效果评价,并对 2017 年 1 月—2019 年 12 月天津市结核病月发病率进行预测。**结果**　流行病学结果显示,2004 年 1 月—2015 年 12 月天津市结核病月发病率总体呈下降趋势。2005—2008 年出现一个发病高峰,2009 年后大幅度下降,随后趋于平稳。2017 年 1 月—2019 年 12 月天津市结核病月发病率与往年相比平稳下降。建立的最佳模型为 SARIMA$(1,1,1)\times(3,1,1)_{12}$,该模型残差 BOX-Ljung 统计量 $P$ 值为 0.493 ,提示残差为白噪声序列,模型拟合良好。预测结果实际值均在预测值的 95%置信区间。**结论**　SARIMA$(1,1,1)\times(3,1,1)_{12}$模型可对天津市结核病月发病率进行较准确的预测。

[关　键　词]　结核病; ARIMA 时间序列; Python 语言; 发病率; 预测
[中图分类号]　R521

# Application of ARIMA model in predicting the incidence of tuberculosis in Tianjin City based on Python language

*ZHANG Xiao-hui*, *YAO Ting-ting*, *CHEN Yang*, *ZHANG Tian-tian*, *MA Jun* (*Department of Health Statistics*, *School of Public Health*, *Tianjin Medical University*, *Tianjin* 300041, *China*)

[**Abstract**]　**Objective**　To evaluate feasibility of autoregressive integrated moving average (ARIMA) model in predicting the incidence of tuberculosis (TB). **Methods**　Using statsmodels module-based Python language, incidence of TB in Tianjin City from January 2004 to December 2015 was as training set, the optimal seasonal ARIMA (SARIMA) model was established, data from January to December 2016 were used to evaluate the efficacy of SARIMA model, and monthly incidence of TB in Tianjin City from January 2017 to December 2019 was predicted. **Results**　Epidemiological results showed that monthly incidence of TB in Tianjin showed a overall downward trend from January 2004 to December 2015. There was a of peak disease incidence in 2005－2008, which dropped sharply after 2009 and then stabilized. From January 2017 to December 2019, monthly incidence of TB in Tianjin City declined steadily compared with previous years. The established optimal model was SARIMA$(1,1,1)\times(3,1,1)_{12}$, residual BOX-Ljung statistic of the model was $P=0.493$, which indicated that the residual was a white noise sequence and the model fitted well. The actual value of predicted results was within 95% confidence interval of predicted value. **Conclusion**　SARIMA $(1,1,1)\times(3,1,1)_{12}$ model can accurately predict the monthly incidence of tuberculosis in Tianjin City.

[**Key words**]　tuberculosis; ARIMA time series; Python language; incidence; prediction

[收稿日期]　2019－09－16
[作者简介]　张晓卉(1993－),女(汉族),内蒙古乌兰察布市人,硕士研究生,主要从事生物统计学应用研究。
[通信作者]　马骏　　E-mail: junma@tmu.edu.cn

结核病是影响我国人民健康的重大传染病之一，发病率和病死率均位于我国传染病第二位[1-2]。世界卫生组织（WHO）2018 年报告显示，结核病合并免疫系统受损（如感染人类免疫缺陷病毒）、营养不良、糖尿病以及吸烟等将大大提高其病死率[3]。因此，对结核病发病率进行预测，根据其未来预测趋势为相关部门制定防控措施，具有实际意义。目前，国内已有学者利用时间序列模型对结核病的发病趋势进行预测并取得了较好的效果[4-6]，但对天津市结核病月发病率预测时间序列研究鲜有报道。鉴于关于天津市结核病发病率水平的研究较少，本研究基于 2004 年 1 月—2016 年 12 月天津市结核病月发病率数据，通过时间序列分析建立乘积季节差分自回归移动平均（autoregressive integrated moving average，ARIMA）模型，拟对天津市结核病月发病率进行预测，以期增强结核病防控预警。当前常见的统计分析软件为 SAS、SPSS、R 语言等，三者主要针对统计分析，对于大数据的挖掘开发可视化，用户普及性等不及 Python 语言，故本研究应用 Python 语言进行预测模型的拟合。

## 1 资料与方法

1.1 资料来源 天津市 2004—2016 结核病月发病率数据来源于传染病网络直报，数据获取平台为国家人口与健康科学数据共享平台公共卫生科学数据中心（http://www.ncmi.cn/info/69/1544），资料可靠。

1.2 Python 语言 Python 语言作为当前最热门的编程语言之一，仅次于 java 语言和 C 语言[7]。Python 语言不仅可用于统计分析，还被广泛的应用于数据爬取、人工智能等领域[8]，其语言简单、优美，且免费开源。强大的第三方库支持多种科学计算和统计分析[9]。在时间序列分析中，Python 语言建模过程简单，图形直观。当处理的时间序列数据量较大时，Python 语言可利用其第三方库 pandas，从而规避循环，极大地节省程序运行时间，具有 R 语言等不具有的自身优势。

1.3 季节性差分自回归移动平均（SARIMA）模型 ARIMA 模型将数据视为随时间变化的随机变量，根据序列的历史值预测将来值[10]。考虑到季节性，使用乘积 SARIMA 模型。模型常表述为 SARIMA(p，d，q)（P，D，Q），其中 AR 是自回归，p 为自回归项，MA 为移动平均，q 为移动平均项数，d 为使时间序列平稳的差分次数，“P，D，Q”为相应带有季节性的参数。

1.3.1 建模步骤[11-14] （1）平稳性检验：依据时间序列图、自相关图（autocorrelation function，ACF）、偏自相关图（partial autocorrelation function，PACF）及迪基－福勒检验（Dickey-Fuller Test）判断数据是否平稳，若不平稳需要进行平稳性处理。（2）平稳性处理：可采取对数并差分、季节差分、移动平均等方法使序列平稳。（3）白噪声检验：当 BOX-Ljung 统计量 $P \leqslant 0.05$ 时，判断序列为非白噪声序列，可进行后续分析。（4）定阶：根据差分次数确定 d、D，根据 ACF 和 PACF 确定 p、q、P、Q，参考赤池信息准则（Akaike information criterion，AIC）及贝叶斯信息准则（Bayesian information criterion，BIC）确定最优模型。（5）参数估计及残差分析：确定模型参数并检验其显著性，对残差进行白噪声检验。若检验通过，则模型建模良好。

1.3.2 模型预测及评价 以天津市 2004 年 1 月—2015 年 12 月结核病月发病率数据作为训练集拟合模型，以 2016 年 1—12 月数据作为验证集验证模型拟合精度，预测 2017 年 1 月—2019 年 12 月天津市结核病月发病率[15]。以误差均方（mean square error，MSE）、平均绝对误差（mean absolute error，MAE）衡量模型的预测精度和建模效果，其中 MSE、MAE 越小，表示预测精度越好，建模效果越好[16]。

1.4 统计学方法 建模平台采用基于 64 位 Anaconda（Python3.7），调用的模块主要有 numpy、pandas、matplotlib 以及 statsmodels。

Python 建模过程，（1）导入一系列包：import pandas as pd，import numpy as np，import statsmodels.api as sm，import matplotlib.pyplot as plt；（2）导入 2004 年 1 月—2015 年 12 月数据并将其转换为时间类型数据：data = pd.read_csv('jiehebin.csv')，dateparse = lambda dates：pd.datetime.strptime(dates，'%Y－%m')，data = pd.read_csv('jiehebin.csv'，parse_dates = ['month']，index_col = 'month'，date_parser = dateparse)；（3）差分使序列平稳：data_first_difference = data.diff(1)，data_seasonal_first_difference = data_first_difference.diff(12)；（4）模型拟合及预测：Mod = sm.tsa.statespace.SARIMAX(data，trend = "n"，order = (1，1，1)，seasonal_order = (3，1，1，12))，results = mod.fit()，predict_ts = results.predict()。

## 2　结果

2.1　流行病学趋势　以 2004 年 1 月—2015 年 12 月数据建模，将时间序列图分解为趋势性成分、季节性成分和随机成分三部分，结果显示 2004 年 1 月—2015 年 12 月天津市结核病月发病率总体呈下降趋势，2005—2008 年出现一个发病高峰，2009 年后大幅度下降，随后趋于平稳。结核病发病率于每年 3—8 月份高发，冬季骤然降低，提示其具有季节性。见图 1、2。

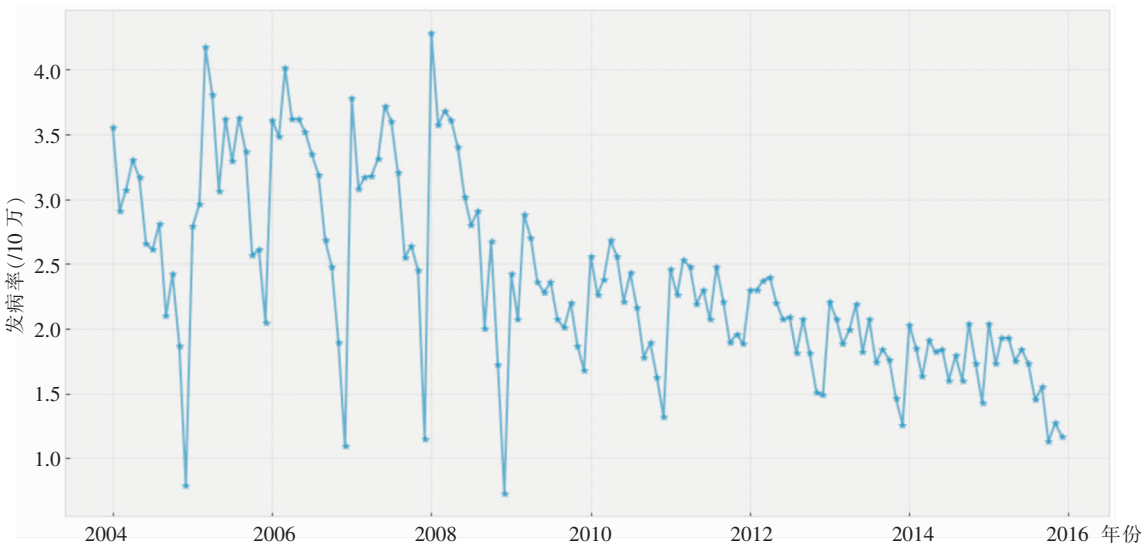


**图 1**　天津市 2004 年 1 月—2015 年 12 月结核病月发病率时间序列图

**Figure 1**　Time-series of monthly incidence of TB in Tianjin City from January 2004 to December 2015

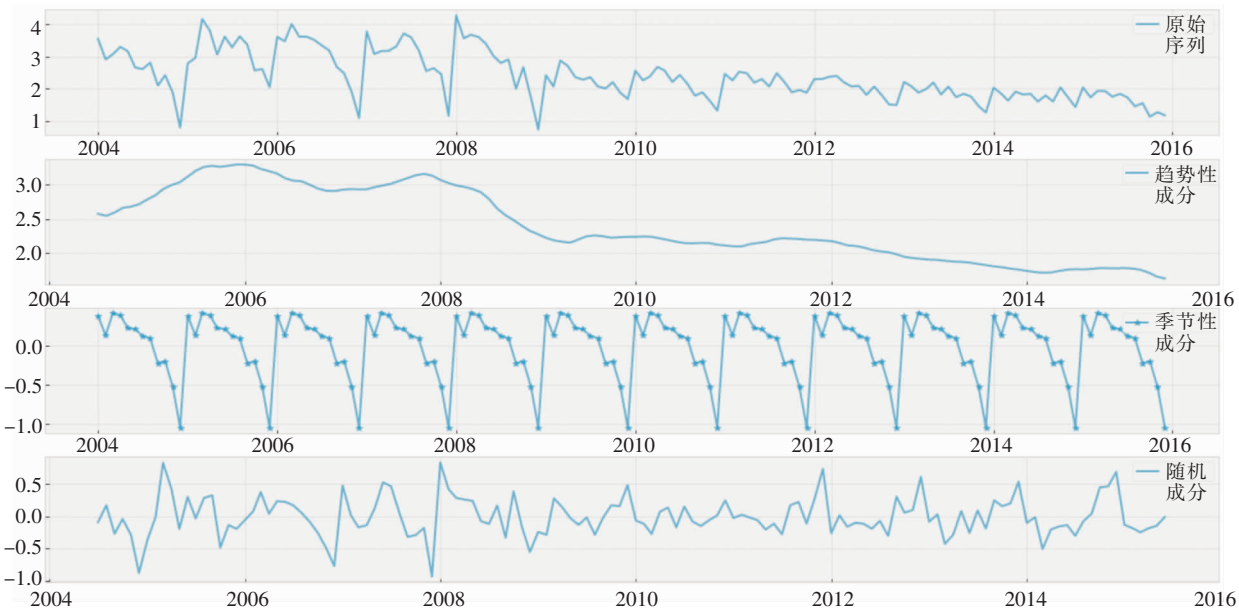


**图 2**　天津市 2004 年 1 月—2015 年 12 月结核病月发病率时间序列分解图

**Figure 2**　Time-series breakdown of monthly incidence of TB in Tianjin City from January 2004 to December 2015

2.2　SARIMA 模型建模结果

2.2.1　数据预处理　原始时间序列图经 Dickey-Fuller Test，结果显示 $P$ 值为 0.81，原始序列不平稳，需要进行平稳性处理。对原始序列进行一次差分和一次季节差分后序列图平稳，见图 3。ACF 和 PACF 亦显示数据已平稳，见图 4。Dickey-Fuller

Test 结果显示 $P$ 值为 0.000029，亦说明序列已平稳。经差分后的平稳序列 BOX-Ljung 统计量 $P<0.05$，判断该序列为非白噪声序列。

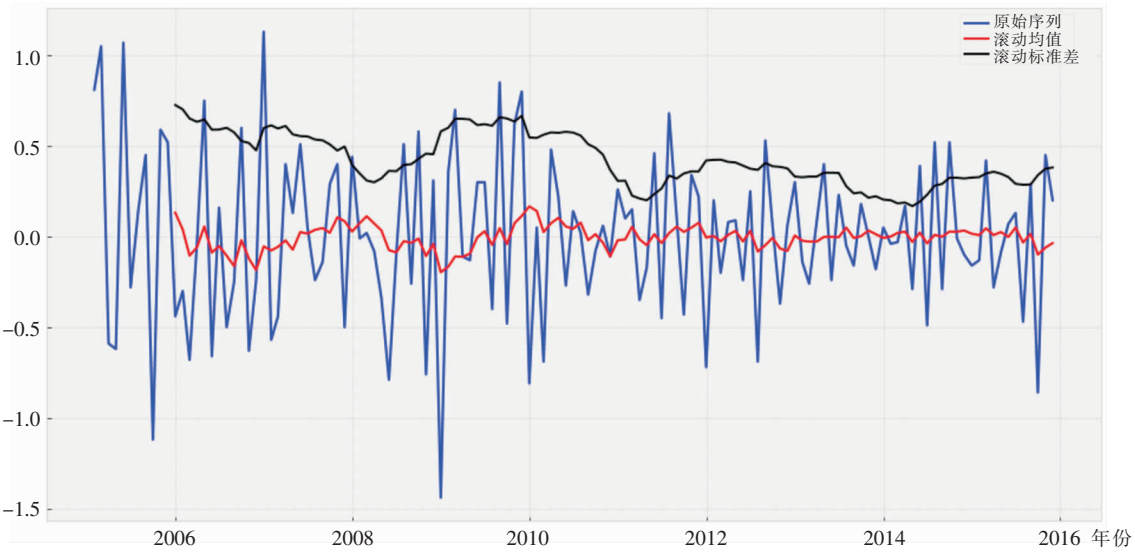


**图 3**　差分后序列时间序列图及滚动均值、滚动标准差图

**Figure 3**　Timing sequence diagram of the difference and rolling mean and rolling standard

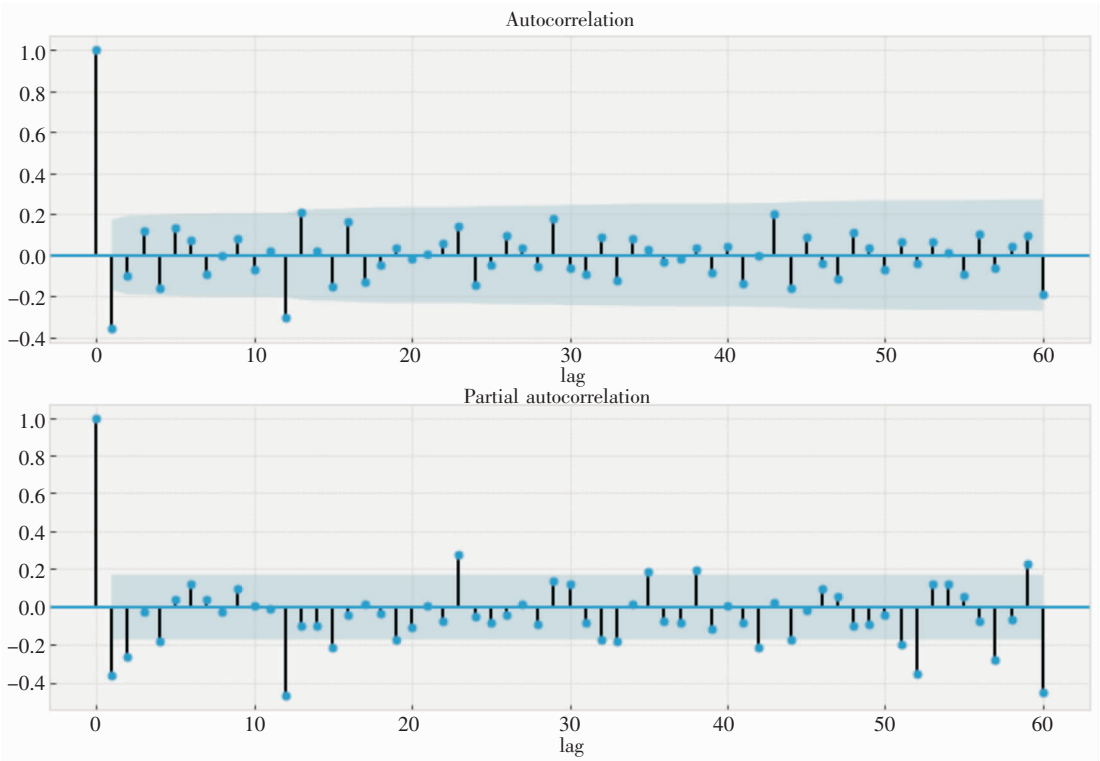


**图 4**　差分后序列的 ACF、PACF

**Figure 4**　ACF and PACF of the difference sequence

2.2.2　模型识别　由序列一阶差分和一阶季节差分初步确定 d = 1，D = 1，初步确定模型形式为 SARIMA(p,1,q) × (P,1,Q)$_{12}$。观察差分后的 ACF 和 PACF（见图 4），可见 ACF 呈截尾，PACF 呈拖尾，提示季节性模型为 SARIMA(3,1,1)$_{12}$模型。观察 SARIMA(1,1,1) × (3,1,1)$_{12}$残差序列的 ACF

和 PACF(见图 5),提示非季节模型为 SARIMA(1,1,1),故原始序列初步拟合为乘积混合效应模型 SARIMA(1,1,1)×(3,1,1)$_{12}$。为确保筛选出最优模型,采用从低阶到高阶逐个进行尝试的方法挑选模型参数[17-18]。初步纳入 SARIMA(3,1,1)×(3,1,1)$_{12}$、SARIMA(2,1,1)×(3,1,1)$_{12}$、SARIMA(1,1,1)×(3,1,1)$_{12}$、SARIMA(1,1,1)×(3,1,2)$_{12}$、SARIMA(2,1,1)×(3,1,2)$_{12}$、SARIMA(3,1,1)×(3,1,2)$_{12}$六个模型进行试验,根据 AIC、BIC 准则选取其中最小值作为最优模型(见表 1),因此原始序列拟合为 SARIMA(1,1,1)×(3,1,1)$_{12}$,拟合后残差的 ACF、PACF 见图 5。残差 BOX-Ljung 统计量 $P$ 值为0.493,判断模型拟合后残差为白噪声序列。

**表 1**　拟纳入模型 AIC、BIC 值

**Table 1**　AIC and BIC values to be included in the model

| 模型 | AIC | BIC |
|---|---|---|
| SARIMA(3,1,1)×(3,1,1)$_{12}$ | 110.686 | 136.563 |
| SARIMA(2,1,1)×(3,1,1)$_{12}$ | 110.493 | 133.495 |
| SARIMA(1,1,1)×(3,1,1)$_{12}$ | 108.971 | 129.097 |
| SARIMA(1,1,1)×(3,1,2)$_{12}$ | 108.351 | 131.353 |
| SARIMA(2,1,1)×(3,1,2)$_{12}$ | 109.387 | 135.713 |
| SARIMA(3,1,1)×(3,1,2)$_{12}$ | 110.453 | 139.205 |

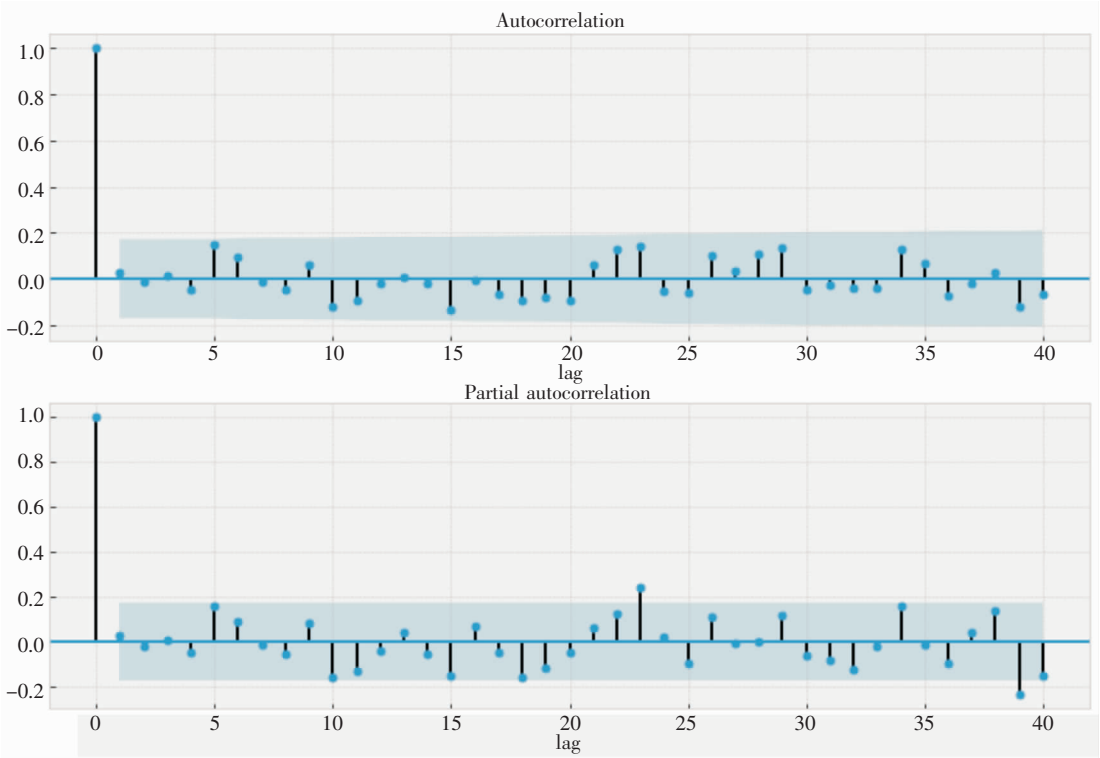


**图 5**　SARIMA(1,1,1)×(3,1,1)$_{12}$模型拟合后残差的 ACF、PACF

**Figure 5**　Residual ACF and PACF after fitting the SARIMA(1,1,1)×(3,1,1)$_{12}$ model

2.2.3　参数估计及检验　模型的参数估计结果除 ar.L1 无统计学意义外,其他参数均有统计学意义。因此,除去 ar.L1,将其他参数全部列入 SARIMA(1,1,1)×(3,1,1)$_{12}$模型。见表 2。

**表 2**　SARIMA(1,1,1)×(3,1,1)$_{12}$模型的参数估计

**Table 2**　Parameter estimation of SARIMA(1,1,1)×(3,1,1)$_{12}$ model

| 参数 | 系数 | SE | $t$ | $P$ | 95%$CI$ |
|---|---|---|---|---|---|
| ar.L1 | 0.1020 | 0.168 | 0.606 | 0.544 | −0.288～0.432 |
| ma.L1 | −0.7314 | 0.125 | −5.037 | <0.001 | −0.877～−0.386 |
| ar.S.L12 | −1.1726 | 0.224 | −5.243 | <0.001 | −1.611～−0.734 |
| ar.S.L24 | −0.7837 | 0.143 | −4.771 | <0.001 | −0.965～−0.403 |
| ar.S.L36 | −0.3998 | 0.085 | −4.69 | <0.001 | −0.567～−0.233 |
| ma.S.L12 | 0.8310 | 0.285 | 2.352 | 0.019 | 0.112～1.230 |
| sigma2 | 0.1128 | 0.016 | 6.891 | <0.001 | 0.081～0.144 |

2.2.4　模型拟合　将天津市 2004 年 1 月—2015 年 12 月结核病月发病率数据作为训练集拟合 SARIMA$(1,1,1)\times(3,1,1)_{12}$ 模型，其中 MAE = 0.306，MSE = 0.224。见图 6。

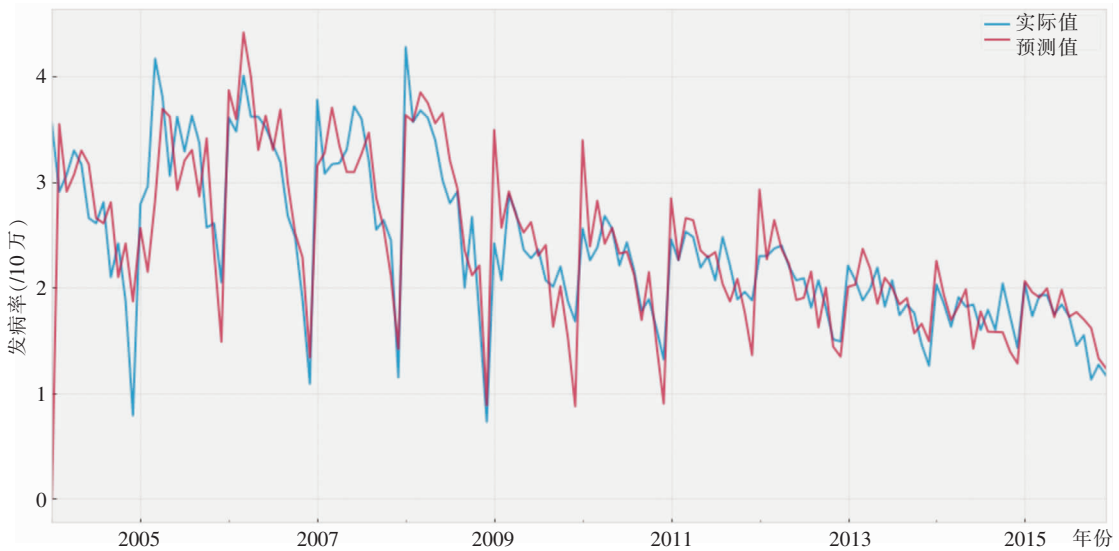


**图 6**　2004 年 1 月—2015 年 12 月天津市结核病月发病率拟合结果

**Figure 6**　Fitting result of monthly incidence of TB in Tianjin City from January 2004 to December 2015

2.2.5　模型效果评价　将 2016 年 1—12 月结核病月发病率进行回代预测，实际发病例数均在预测发病例数的 95%*CI* 内，模型拟合良好，具有较好的预测性能，其中 MAE = 0.169，MSE = 0.081，可对天津市结核病的发病数进行较准确的预测。见图 7、表 3。

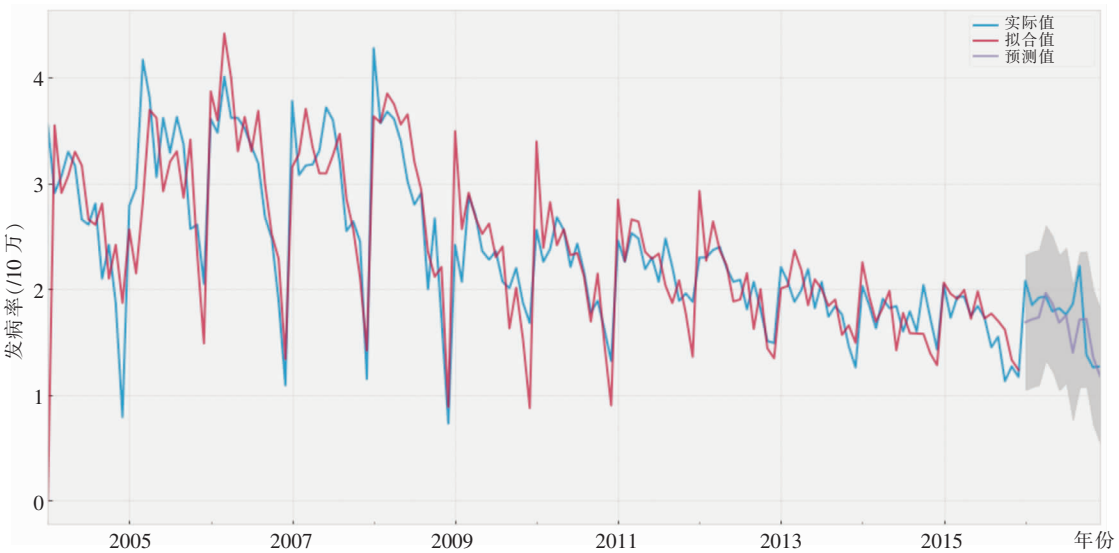


**图 7**　2016 年 1—12 月天津市结核病月发病率预测结果

**Figure 7**　Prediction result of monthly incidence of TB in Tianjin City from January to December 2016

**表 3**　2016 年 1—12 月实际发病率与预测值比较(/10 万)

**Table 3**　Comparison of actual incidence and predicted values from January to December 2016 (/100 000)

| 月份 | 预测值 | 实际值 | 绝对误差 | 相对误差(%) |
|---|---|---|---|---|
| 1 | 1.7606 | 2.0815 | −0.3209 | −15.417 |
| 2 | 1.7611 | 1.8488 | −0.0877 | −4.743 |
| 3 | 1.7701 | 1.9199 | −0.1498 | −7.802 |
| 4 | 1.9354 | 1.9264 | 0.0090 | 0.467 |
| 5 | 1.8123 | 1.7906 | 0.0217 | 1.212 |
| 6 | 1.7623 | 1.8229 | −0.0606 | −3.324 |
| 7 | 1.7632 | 1.7648 | −0.0016 | −0.091 |
| 8 | 1.2786 | 1.8553 | −0.5767 | −31.080 |
| 9 | 1.7534 | 2.2237 | −0.4703 | −21.150 |
| 10 | 1.6589 | 1.3769 | 0.2820 | 20.480 |
| 11 | 1.2796 | 1.2605 | 0.0191 | 1.515 |
| 12 | 1.2505 | 1.2735 | −0.0230 | −1.806 |

2.2.6　模型预测　利用 SARIMA(1,1,1)×(3,1,1)$_{12}$对天津市 2017 年 1 月—2019 年 12 月肺结核发病率进行预测,结果显示天津市结核病月发病率将总体呈现下降趋势,春季高发,冬季发病率降低,符合结核病发病规律,预测结果可供参考。见图 8、表 4。

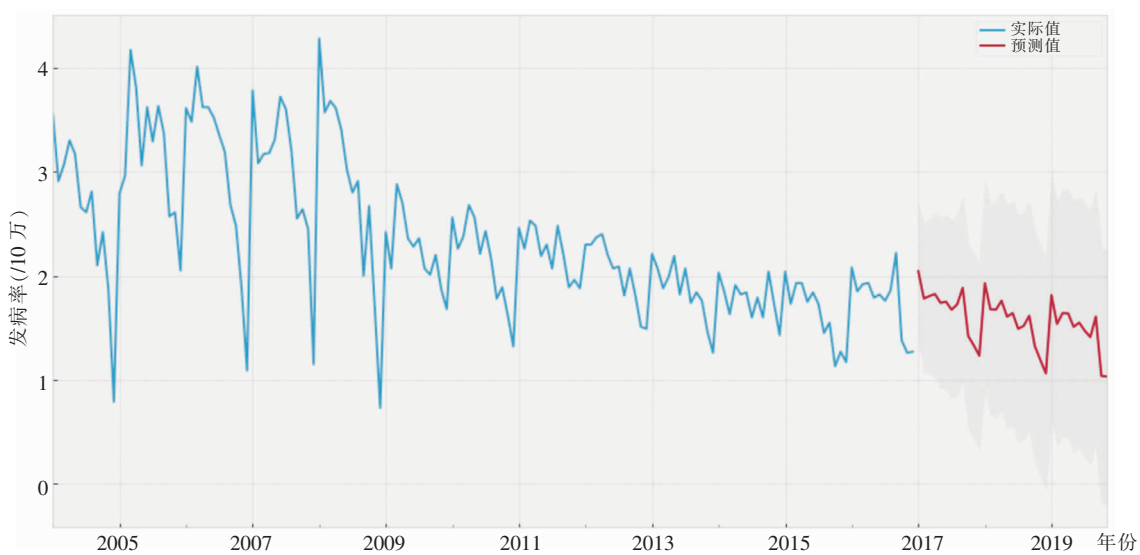


**图 8**　SARIMA 对 2017 年 1 月—2019 年 12 月天津市结核病月发病率预测结果

**Figure 8**　Monthly incidence of TB in Tianjin City predicted by SARIMA from January 2017 to December 2019

**表 4**　SARIMA 对 2019 年 7—12 月天津市结核病月发病的预测值

**Table 4**　Predictive value of SARIMA for monthly incidence of TB in Tianjin City from July to December 2019

| 月份 | 预测发病人数 | 95%*CI* |
|---|---|---|
| 7 月 | 293 | 221～356 |
| 8 月 | 304 | 237～369 |
| 9 月 | 283 | 212～345 |
| 10 月 | 252 | 173～331 |
| 11 月 | 234 | 140～329 |
| 12 月 | 229 | 134～323 |

## 3　讨论

随着深度学习和数据挖掘技术的日渐成熟,Python 语言风靡全球。在信息爆炸和多学科融合的大数据时代,Python 语言作为一门通用语言在统计学应用中的地位也愈加重要,相较于 R 语言,Python 语言在前期数据收集、处理、建模和运行速度等方面显示出卓越性能,因此本研究使用 Python 语言建立 ARIMA 时间序列预测模型。

目前,有多种模型可用于传染病发病率的预测,如 GM(1,1)模型[19]、马尔可夫链预测模型[20]、ARIMA 模型[13,21-22]等。其中 ARIMA 模型作为最

经典的预测模型之一，可将时间序列分解为趋势性成分、季节性成分和随机干扰，并对噪声进行分析处理，预测精度较高，是传染病时间序列预测模型中最重要的手段[23]。胡晓媛等[24]应用 SARIMA 模型对全国肺结核月发病率进行预测，预测 MAE 值为 0.416992。本研究 MAE 值为 0.169，较之稍高。秘玉清等[25]应用 SARIMA 模型预测山东省结核病发病趋势，得出发病率将呈现周期性上升的结论。鉴于对天津市肺结核月发病率研究较少，本研究对天津市结核病月发病率的流行病学趋势作了描述性研究，并预测其未来发病趋势，可为天津市相关部门制定防控措施提供理论依据。

本研究基于 2004 年 1 月—2016 年 12 月天津市结核病月发病率资料，将时间序列分解为趋势性、季节性和随机噪声三部分，分析结核病的发病趋势和季节性，最终确定 $SARIMA(1,1,1)\times(3,1,1)_{12}$ 为天津市结核病月发病率的最终模型。首先，利用 2004 年 1 月—2015 年 12 月数据建立最优模型，结果显示 $SARIMA(1,1,1)\times(3,1,1)_{12}$ 可较准确地拟合实际月发病率，残差为白噪声序列，说明建模良好。然后将 2016 年 1—12 月结核病月发病率进行回代预测，实际值均在预测值 95%*CI* 内，说明模型适用于对天津市未来结核病月发病率的预测。最后，将 $SARIMA(1,1,1)\times(3,1,1)_{12}$ 模型应用于对 2017 年 1 月—2019 年 12 月结核病月发病率的预测，可用预测值来估计未来结核病的流行强度，若 2017 年 1 月—2019 年 12 月实际发病人数在预测发病人数的 95%*CI* 内，表明当月的结核病疫情基本正常；若发现实际发病人数处于预测发病人数的 95%*CI* 外，则提示结核病疫情有可能发生异常[15]。

本研究将 2016 年 1—12 月发病率进行回代预测时，出现 8—10 月份发病率相对误差和其他月份相差较大的情况，其原因：一方面，可能是由于 ARIMA 模型作为经典的线性模型在处理具有非线性特点的时间序列问题上表现出一定的局限性；另一方面，结核病的发生发展具有多因素性，未来考虑将除时间因素以外的其他混杂因素列入模型，以提高其预测精度。

时间序列具有混沌现象，存在内在随机性和不规则有序性[11]。因此，对时间序列的预测在尽可能抓取线性成分的基础上还应更多的关注非线性成分。而 ARIMA 模型作为一种线性模型在处理非线性成分的问题上具有一定的不足，容易使预测精度降低。目前，关于基于误差补偿思想和相空间重构思想的 ARIMA 混合模型已经崭露头角[26-27]。然而，对于 ARIMA-SVM、ARIMA-LSTM 等混合模型虽预测效果较单纯 ARIMA 模型较好，但解释性差。在 ARIMA 模型建模过程中，残差应通过白噪声检验才能判断建模是否合理，而对残差进行相空间重构则需要残差具有混沌性，而非随机序列，以上矛盾之处此类混合模型无法给出合理的解释。另外，关于将 ARIMA 模型和各类神经网络相结合的混合模型也层出不穷，研究结果显示其预测精度相较于单纯 ARIMA 模型高，但由于神经网络背后数学理论仍处于“黑箱子”状态，其混合模型预测的准确性和重复性仍有待探讨。神经网络模型的训练效果和预测精度随着数据量的增大和数据维度的增高而增大，基于时间序列数据短期预测的特点，将神经网络运用于时间序列中，其理论有待完善。ARIMA 模型常用于短期预测，因此，其预测结果需要随着数据量的更新而不断更新，如何将 ARIMA 和其他各种预测模型如灰色模型、隐马尔可夫链等有效结合，通过充分提取时间序列的线性和非线性成分，控制随机误差，提高预测的准确度和精确度，是今后研究的方向。

## [参 考 文 献]


[1] 全国第五次结核病流行病学抽样调查技术指导组，全国第五次结核病流行病学抽样调查办公室. 2010 年全国第五次结核病流行病学抽样调查报告[J]. 中国防痨杂志，2012，34(8)：485－508.

[2] Fogel N. Tuberculosis：a disease without boundaries[J]. Tuberculosis (Edinb)，2015，95(5)：527－531.

[3] WHO. Tuberculosis fack sheet，2018[EB/OL]. (2018－06－16)[2019－08－10]. http://www.who.int/mediacentre/factsheets/fs104/en.

[4] 钟球，蒋莉，周琳，等. 广东省结核病发病趋势的时间序列分析[J]. 中国防痨杂志，2010，32(9)：515－519.

[5] 牟瑾，谢旭，李媛，等. 将 ARIMA 模型应用于深圳市 1980～2007 年重点法定传染病预测分析[J]. 预防医学论坛，2009，15(11)：1051－1052，1055.

[6] 原梅，张治国，豆智慧，等. 北京市昌平区肺结核发病数 ARIMA 模型预测[J]. 疾病监测，2015，30(12)：1045－1049.

[7] 祝永志，荆静. 基于 Python 语言的中文分词技术的研究[J]. 通信技术，2019，52(7)：1612－1619.

[8] 谢克武. 大数据环境下基于 python 的网络爬虫技术[J]. 电子制作，2017(9)：44－45.

[9] 郭丽蓉. 基于 Python 的网络爬虫程序设计[J]. 电子技术与软件工程，2017(23)：248－249.

[10] 胡建利，梁祁，吴莹，等. 季节时间序列模型在菌痢发病预测中的应用[J]. 中国卫生统计，2012，29(1)：34－36，39.

[11] 李超. 混沌时序黄金期货价格预测研究[D]. 广州：暨南大学，2018.

[12] Chadsuthi S，Iamsirithaworn S，Triampo W，et al. Modeling seasonal influenza transmission and its association with climate factors in Thailand using time-series and ARIMAX analyses[J]. Comput Math Methods Med，2015，2015：436495.

[13] Liu L，Luan RS，Yin F，et al. Predicting the incidence of hand，foot and mouth disease in Sichuan Province，China using the ARIMA model[J]. Epidemiol Infect，2016，144(1)：144－151.

[14] Wang T，Liu J，Zhou Y，et al. Prevalence of hemorrhagic fever with renal syndrome in Yiyuan County，China，2005－2014[J]. BMC Infect Dis，2016，16：69.

[15] 李晓蓉，庞学文，于燕明，等. ARIMA 模型在天津市结核病发病预测中的应用[J]. 实用预防医学，2018，25(12)：1536－1538.

[16] 刘天，姚梦雷，黄继贵，等. 组合预测模型在丙型病毒性肝炎发病率预测中的应用[J]. 中国疫苗和免疫，2018，24(6)：675－680.

[17] 张正斌，段琼红，李月华，等. ARIMA 模型在武汉市结核病疫情预测中的应用[J]. 公共卫生与预防医学，2017，28(3)：27－30.

[18] Wang KW，Deng C，Li JP，et al. Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network[J]. Epidemiol Infect，2017，145(6)：1118－1129.

[19] 王雅文，沈忠周，杨银. GM(1，1)模型在我国梅毒发病率预测中的应用[J]. 实用预防医学，2019，26(9)：1069－1071，1079.

[20] 李文华，杨亚涛，曾年华. 马尔可夫链在南方某部队肺结核发病趋势预测分析中的应用[J]. 华南国防医学杂志，2017，31(5)：339－341.

[21] Wangdi K，Singhasivanon P，Silawan T，et al. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses：a case study in endemic districts of Bhutan[J]. Malar J，2010，9：251.

[22] Wang C，Li Y，Feng W，et al. Epidemiological features and forecast model analysis for the morbidity of influenza in Ningbo，China，2006－2014[J]. Int J Environ Res Public Health，2017，14(6)：559.

[23] 孟蕾，王玉明. ARIMA 模型在肺结核发病预测中的应用[J]. 中国卫生统计，2010，27(5)：507－509.

[24] 胡晓媛，孙庆文，王玲玲，等. 基于乘积 SARIMA 模型的肺结核发病率预测[J]. 第二军医大学学报，2016，37(8)：969－974.

[25] 秘玉清，张继萍，殷延玲，等. 基于 ARIMA 模型的山东省肺结核发病趋势预测[J]. 中国卫生统计，2018，35(6)：879－881.

[26] Mei WJ，Xu P，Liu RC，et al. Stock price prediction based on ARIMA-SVM model[C]//2018 International Conference on Big Data and Artificial Intelligence(ICBDAI 2018)，Ningbo，Zhejiang，China，2018－12－21，2018：55－61.

[27] 程俊. 基于 ARIMA-LSTM 混合模型的机械传动件制造企业销售预测方法研究与应用[D]. 成都：电子科技大学，2018.

（本文编辑：文细毛）

**本文引用格式**：张晓卉，姚婷婷，陈阳，等. 基于 Python 语言的 ARIMA 模型在天津市结核病发病率预测中的应用[J]. 中国感染控制杂志，2020，19(7)：634－642. DOI：10.12138/j.issn.1671－9638.20205807.

***Cite this article as***：ZHANG Xiao-hui，YAO Ting-ting，CHEN Yang，et al. Application of ARIMA model in predicting the incidence of tuberculosis in Tianjin City based on Python language[J]. Chin J Infect Control，2020，19(7)：634－642. DOI：10.12138/j.issn.1671－9638.20205807.