Statistics and Application 统计学与应用, 2025, 14(2), 8-21
Published Online February 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.142030

Hans汉斯

基于SARIMA-LSTM模型的中国肺结核传染病预测研究

王晓琴[1]，杨　震[2]，包　城[2]，郭松柏[1]，许传青[1*]

[1]北京建筑大学理学院，北京
[2]北京市昌平区结核病研究所，北京

收稿日期：2025年1月17日；录用日期：2025年2月11日；发布日期：2025年2月20日

摘　要

背景：中国是结核病高负担国家之一，尽管肺结核新发病例数逐年下降，但每年新增感染者的数量一直处于较高水平且肺结核感染者的诊断率较低。目的：选择更精准的预测肺结核的发病情况模型，为肺结核的防控和预警提供科学依据。方法：建立SARIMA和LSTM模型，运用加权组合的方法构建SARIMA-LSTM组合模型，使用平均绝对误差(MAE)、均方根误差(RMSE)和平均绝对误差百分比(MAPE)三个评价指标比较模型的预测性能，确定最佳预测模型，并使用该模型对肺结核发病趋势进行预测。结果：SARIMA模型、LSTM模型和SARIMA-LSTM组合模型的平均绝对百分比误差(MAPE)分别为17.95、14.62、8.49，组合模型的MAPE比SARIMA模型降低了52.70%，比LSTM模型降低了41.89%。结论：SARIMA-LSTM组合模型的拟合效果更好，预测误差在三个模型中最低。该组合模型能发挥单一模型的优势，相比两种单一模型提升了预测的准确性。

关键词

肺结核，SARIMA，LSTM

Prediction of Tuberculosis Infection in China Based on SARMIA-LSTM Model

Xiaoqin Wang[1], Zhen Yang[2], Cheng Bao[2], Songbai Guo[1], Chuanqing Xu[1*]

[1]College of Science, Beijing University of Civil Engineering and Architecture, Beijing
[2]Beijing Changping Institute for Tuberculosis Prevention and Treatment, Beijing

Received: Jan. 17th, 2025; accepted: Feb. 11th, 2025; published: Feb. 20th, 2025

Abstract

Background: China is one of the countries with a high burden of Tuberculosis (TB). Although the number

*通讯作者。

文章引用: 王晓琴, 杨震, 包城, 郭松柏, 许传青. 基于 SARIMA-LSTM 模型的中国肺结核传染病预测研究[J]. 统计学与应用, 2025, 14(2): 8-21. DOI: 10.12677/sa.2025.142030

of new cases of TB has been decreasing year by year, the number of newly infected people each year has been at a high level and the diagnosis rate of TB-infected people is low. Objective: To select a more accurate model to predict the incidence of tuberculosis and provide a scientific basis for the prevention and control of tuberculosis and early warning. Methods: The SARIMA and LSTM models were established, and the SARMIA-LSTM combined model was constructed by the weighted combination method. The prediction performance of the model was compared by the three evaluation indexes of Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), and the optimal prediction model was determined. The model was used to predict the trend of tuberculosis incidence. Results: The Mean Absolute Percentage Error (MAPE) of SARIMA model, LSTM model and SARMI-LSTM combined model were 17.95, 14.62 and 8.49, respectively. The MAPE of the combined model was reduced by 52.70% compared with SARIMA model and 41.89% compared with LSTM model. Conclusion: SARMIA-LSTM combined model has a better fitting effect, and the prediction error is the lowest among the three models. The combined model can give full play to the advantages of a single model and improve the accuracy of prediction compared with the two single models.

Keywords

Tuberculosis, SARIMA, LSTM

1. 引言

肺结核是由结核分枝杆菌引起的肺部感染性疾病，以呼吸道传播为主。尽管全球结核病发病率缓慢下降，但结核病在全球范围内仍是一个重要的公共卫生问题[1]。2023 年世界卫生组织《全球结核病报告》[2]显示，在 30 个结核病高负担国家中，中国结核病发病数排第 3 位，占全球发病数的 7.1%。2022 年全球结核病死亡人数为 130 万，结核病是仅次于新型冠状病毒感染的单一传染源死因，严重威胁人类公共卫生和生命健康安全。中国肺结核发病率总体呈下降趋势，由于人口基数大，中国实现“2035 年终结结核病”[3]的战略目标十分困难。

时间序列分析是一种处理时间序列数据或趋势分析的统计技术[4]，通过深入分析疾病的长期趋势，识别已有的发病规律并预测未来的趋势，以减轻医疗卫生系统的负担。ARIMA 模型通过拟合历史数据把握疾病的内在规律，对于短期预测的效果较好[5]，众多研究人员使用该模型来预测 COVID-19 [6]-[8]、手足口病[9]、流感[10]等传染病的发病趋势。SARIMA 模型主要应用于带有季节性特点的时间序列数据，在 ARIMA 模型的基础上增加了 4 个季节性参数。结核病的病例数存在季节性变化，在对时间序列数据的研究中考虑季节性因素有助于提高模型的预测效果[11]。LSTM 模型能够捕捉肺结核发病序列数据中的非线性特征，更适合处理长时间序列数据。近年来，LSTM 模型也更多地用于传染病预测，相比于其他单一模型，其预测效果更好[12]。然而单一的模型不能完全捕获时间序列的所有信息。为进一步提高预测精度，许多学者提出了如 SARIMA-NNAR [13]、ARIMA-BPNN [14]等组合模型，结果表明，组合模型的预测精度均高于单一模型。很少有研究人员采用 SARIMA-LSTM 组合模型预测肺结核的发病趋势。本研究根据肺结核历史发病数据建立预测模型，捕捉肺结核的发病规律并对未来的发病数进行预测，研究内容对于控制策略的实施具有重要的现实意义。

本文内容分为四节，第一节介绍了肺结核的背景以及预测模型的研究现状。第二节说明了数据来源，介绍了 SARIMA、LSTM 和 SARIMA-LSTM 组合模型的基本知识以及研究流程。第三节构建 SARIMA 模

型、LSTM 神经网络模型以及 SARIMA-LSTM 组合模型对肺结核发病趋势进行预测，并利用三种评价指标对模型的预测能力进行评估比较，最终确定最佳预测模型并使用该模型对肺结核未来的发病趋势进行预测。第四节是讨论部分。

2. 对象和方法

2.1. 数据来源

2012~2023 年中国肺结核逐月发病数据来自中国疾病预防控制中心[15]发布的《全国法定传染病疫情概况》。数据按照 80% (2012 年 1 月~2021 年 8 月)为训练集，20% (2021 年 9 月~2023 年 12 月)为测试集进行划分。

2.2. 研究方法

2.2.1. SARIMA 模型

ARIMA 模型被广泛应用于单变量分析的时间序列预测。模型通常表示为 ARIMA(p, d, q)形式，由自回归模型 AR(p)、移动平均模型 MA(q)和差分项 d 组合得到。SARIMA 模型[16]在 ARIMA 模型的基础上增加了 4 个季节性参数，分别是 3 个超参数(P, D, Q)和季节性周期参数 S，通常表示为 SARIMA(p, d, q) × (P, D, Q, S)。使用 SARIMA 模型的步骤如下：

1) 绘制肺结核发病时间序列图直观呈现每年发病数据的变化特征。

2) 为了检验数据的平稳性，采用 Dickey-Fuller (ADF)单位根检验；若数据不平稳，采用季节性和非季节性(d 和 D)差分将非平稳数据转换为平稳数据。

3) 根据自相关函数(ACF)和部分自相关函数(PACF)选择模型的阶数，利用贝叶斯信息准则(BIC)最小值来选择最优的模型。

4) 对所选模型的可靠性进行检验，使用 Ljung-Box 检验残差序列是否为白噪声，若是，则表明残差之间没有自相关性，即残差序列数据特征被充分提取。

5) 使用通过检验的最优模型进行预测，得到预测发病数与预测误差，绘制预测结果图。

2.2.2. LSTM 模型

LSTM 模型[17]是一种特殊的 RNN 模型，为解决 RNN 的梯度消失和梯度爆炸问题，在隐含层中增加记忆单元状态。隐含层中建立了控制单元分别为输入门、遗忘门和输出门。输入门的作用是将新的信息选择性记录到细胞状态中，遗忘门是将细胞中的信息选择性遗忘，输出门是将储存的信息带到下一个神经元中。各个门之间相互作用，提高了 LSTM 模型的信息分析能力。

使用 LSTM 模型的步骤如下：

1) 设置参数建立模型。设置随机种子为 42，定义参数，包括特征数量为 1、时间步长为 12。设置一个 LSTM 层和一个全连接层。使用激活函数 tanh 和优化器 Adam，指定损失函数为均方误差。将数据进行归一化处理，使数据取值保持在 0~1 内。

2) 训练模型。将数据划分为训练集(80%)和测试集(20%)进行模型训练。绘制 LSTM 模型的训练损失函数曲线，绘制损失值随着训练次数的变化趋势。

3) 模型预测。使用已训练好的模型进行预测，将输出的结果进行反归一化处理，得出预测结果。计算评价指标，衡量模型的预测效果。

2.2.3. SARIMA-LSTM 组合模型

单一模型存在一定局限性，无法完全提取时间序列中的全部信息。采用加权融合的方法是一种有效

的解决方案，加权组合模型结合 SARIMA 模型和 LSTM 模型的优势，从而提高肺结核发病数的时间序列预测精度，其流程见图 1。

本文引入一组权重数组 weights，为 SARIMA 模型和 LSTM 模型分配一个权重值。为了确定最优权重配置，将预测值与测试数据之间的均方误差作为损失函数，通过逐步调整权重值，寻找能使损失最小化的最佳权重组合。这一过程是迭代进行的，直到找到一组权重，使集成预测的结果最接近实际测试数据。确定最优权重配置后，将 SARIMA 模型和 LSTM 模型的预测结果按照最优权重进行加权求和，得到最终的集成预测结果，并将组合模型的预测结果与两种单一模型的预测结果进行比较。

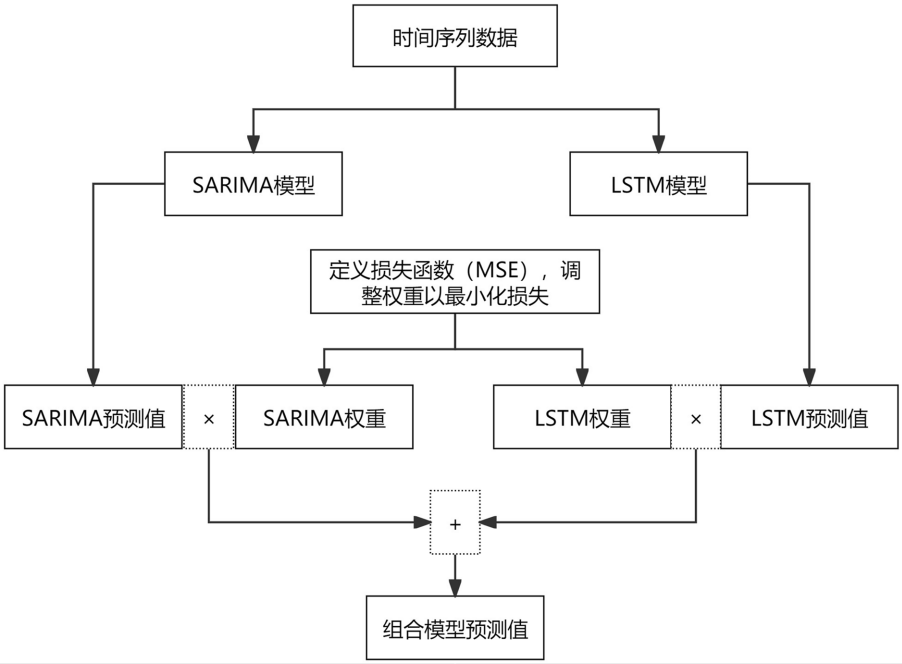

Figure 1. Flow chart of SARIMA-LSTM combination model
图 1. SARIMA-LSTM 组合模型流程图

2.2.4. 评价指标

在模型评价方面，本文选取了平均绝对误差(MAE)、均方根误差(RMSE)和平均绝对百分比误差(MAPE)三个指标[18]对模型预测肺结核发病数的性能进行评估。如公式(1)所示：

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \\ RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \\ MAPE &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \end{aligned} \tag{1}$$

其中，n 为样本个数，y_i 为肺结核发病数真实值，$\hat{y}_i$ 为模型预测值。MAE、RMSE 和 MAPE 的值越小，表示模型预测的效果越好。

3. 结果

根据 2012~2023 年中国每月的肺结核发病数绘制序列图(见图 2)，得到肺结核的发病趋势有较明显的

周期性，肺结核的发病数总体呈下降趋势，每年 3~5 月为发病高峰期。通过对 2012~2023 年每月的发病数据进行分析，得到月均发病数，见表 1。每年的 3~5 月肺结核的月均发病数较高，每年的 10~12 月和 2 月肺结核的月均发病数相对较低。

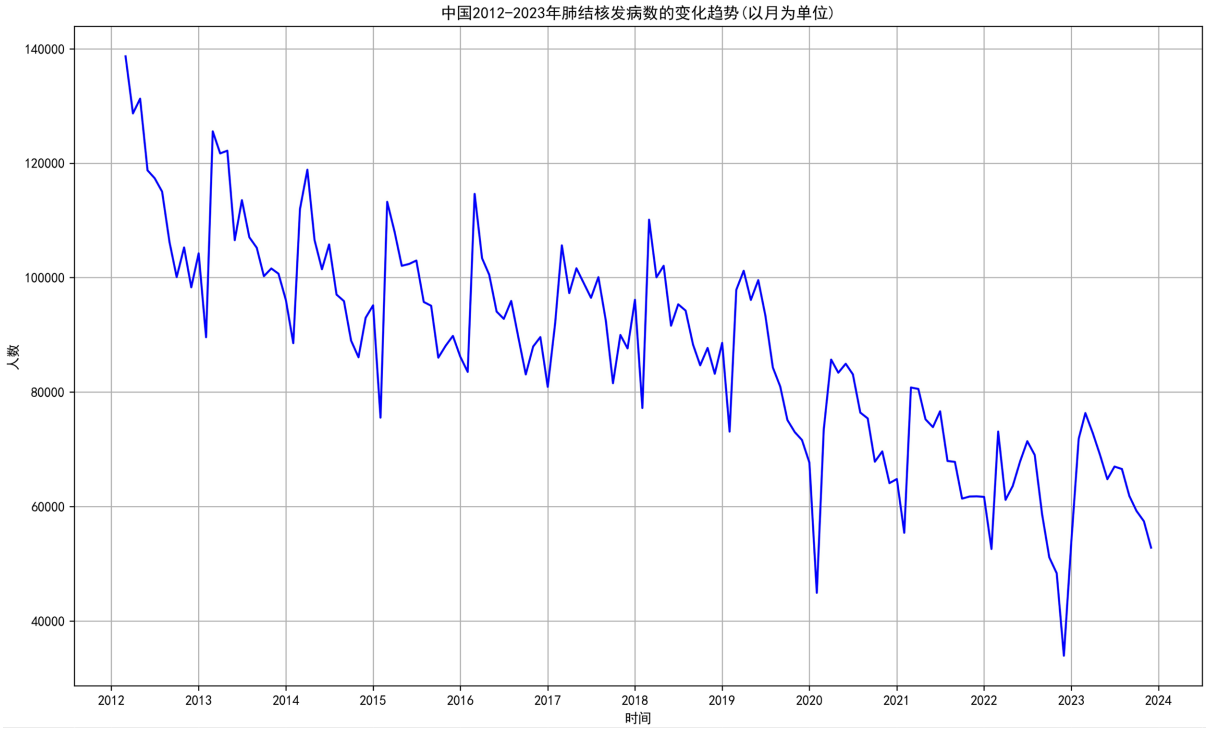

Figure 2. Trends in the incidence of pulmonary tuberculosis in China from 2012 to 2023 (in months)
图 2. 中国 2012~2023 年肺结核发病数的变化趋势(以月为单位)

Table 1. Monthly average incidence of pulmonary tuberculosis from 2012 to 2023
表 1. 2012~2023 年肺结核的月均发病数

月份	月均发病数	月份	月均发病数
1 月	83,285	7 月	94,491
2 月	78,963	8 月	91,116
3 月	104,401	9 月	86,472
4 月	100,676	10 月	79,985
5 月	98,367	11 月	82,119
6 月	94,171	12 月	79,327

3.1. SARIMA 模型结果

对肺结核发病数的原始时间序列进行 STL (Seasonal and Trend decomposition using Loess)分解，见图 3，将该序列分解为趋势项、季节项以及残差项三个部分。趋势项呈下降趋势，季节项呈现每年重复出现的周期性模式，对于肺结核发病数据，存在以一年为周期的循环趋势，残差项是去除了趋势和季节效应后的随机波动。

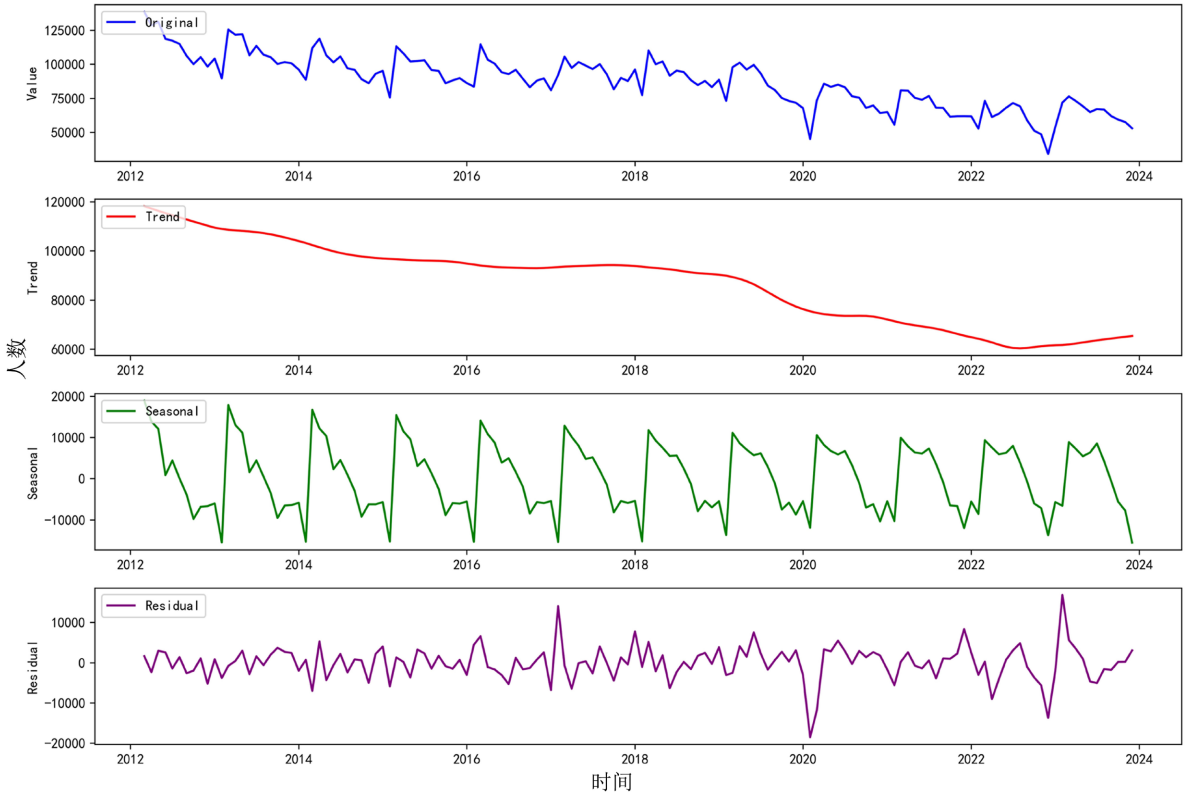

Figure 3. STL decomposition map of incidence of pulmonary tuberculosis
图 3. 肺结核发病数 STL 分解图

原始时间序列展示了明显的趋势性和季节性变化，是非平稳序列，对其进行 ADF (Augmented Dickey-Fuller)检验以确认其平稳性。检验结果显示，p 值为 0.79，远大于显著性水平 0.05，这意味着不能拒绝原假设，即该序列确实为非平稳序列。

Table 2. Results of first-order differential ADF test
表 2. 一阶差分 ADF 检验结果

ADF 检验结果
Dickey-Fuller: −5.66401810716829
Lag-order: 11
p-value: 9.240091071680634e−07
Alternative Hypothesis: Stationary

Table 3. Results of first-order differential white noise test
表 3. 一阶差分白噪声检验结果

Box-Pierce
Test Data: diff_data, Lag-order = 11, df = 1
p-value = [0.00013, 0.00065, 0.00152, 0.00192, 0.00031, 0.00071, 8e−05, 3e−05, 4e−05, 7e−05, 4e−05, 0.0]

鉴于肺结核发病数序列不平稳，首先对序列进行一阶差分，得到差分序列，序列在零值上下波动。

对一阶差分后的数据进行统计检验：ADF 检验与白噪声检验，结果见表 2 和表 3。其中，ADF 检验中 p 值小于 0.05，趋近于 0；白噪声检验得到的 p 值同样均小于 0.05，因此得到的差分序列是平稳的非白噪声序列，可以通过 ARIMA 模型建模。

一阶差分后的 ACF 图和 PACF 图见图 4，自相关系数和偏自相关系数在第二阶之后开始迅速下降并趋向于零。因此，可以考虑使用 max(p) = 2、max(q) = 2 的参数进行建模。由 PACF 图可以得到，从滞后 12 阶开始存在以 12 为周期的波动，表明原序列存在季节性趋势。

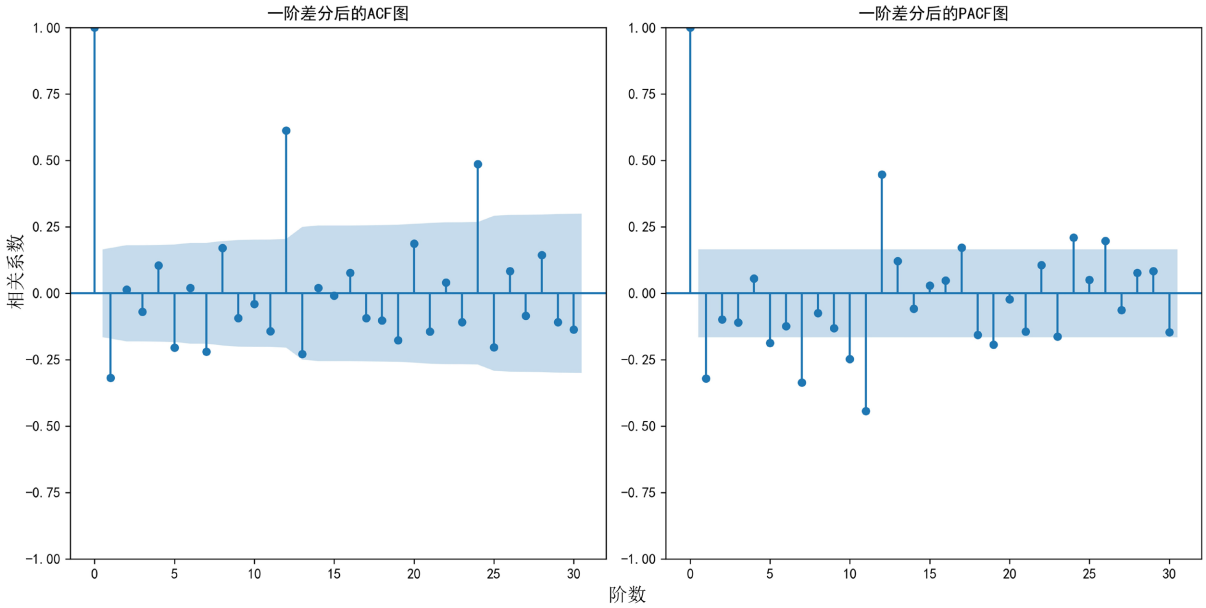

Figure 4. ACF and PACF plots after first-order differencing
图 4. 一阶差分后的 ACF 和 PACF 图

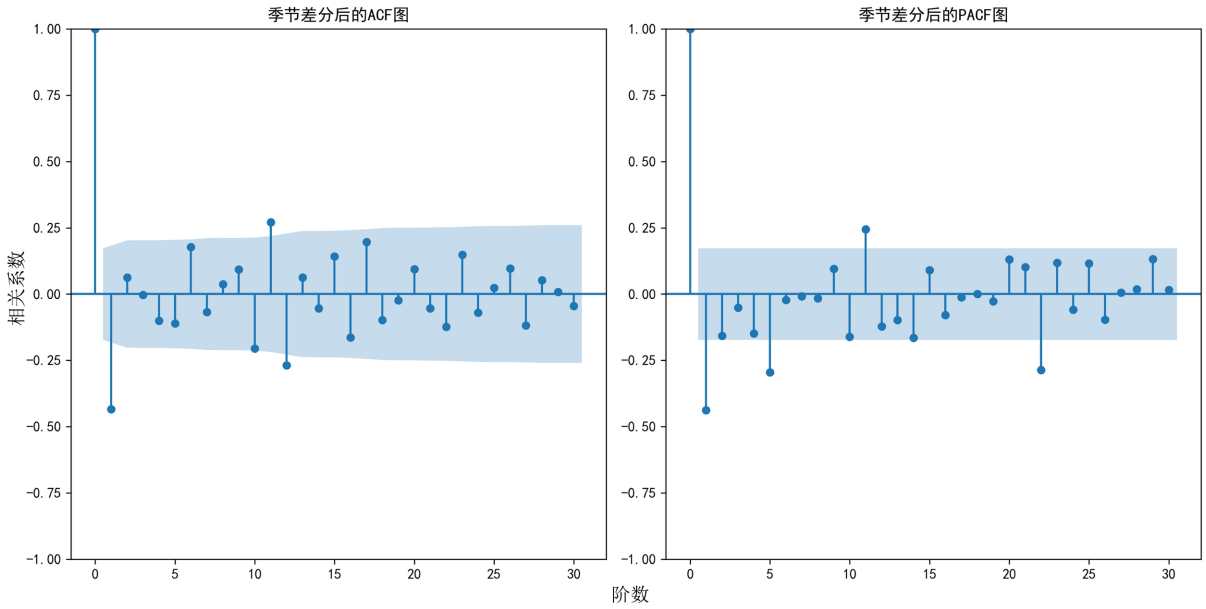

Figure 5. ACF and PACF plots after seasonal differencing
图 5. 季节阶差分后的 ACF 和 PACF 图

在处理具有季节性特征的时间序列预测问题时，SARIMA 模型具有更好的效果。因此，选择 SARIMA 模型进行建模，在一阶差分的基础上进行季节差分，得到季节差分后的 ACF 图和 PACF 图，见图 5，可得自相关系数和偏自相关系数均在第 2 阶之后衰减趋于零，可以考虑使用 MAX(P) = 2、MAX(Q) = 2 的参数进行建模。

对于模型参数的范围划定，最终的 SARIMA 模型参数应满足：max(p) = 2，max(q) = 2，MAX(P) = 2，MAX(Q) = 2，D = 1，d = 1。使用 BIC 准则进行参数寻优，BIC 值越小，选取的参数越好，得到的结果见表 4，仅列出选取参数寻优的前 10 个结果，选取最终参数为 p = 1，q = 1，P = 1，Q = 1。最终得到模型为 SARIMA(1, 1, 1) × (1, 1, 1, 12)。

Table 4. Results of parameter selection based on BIC criteria
表 4. 依据 BIC 准则选取参数的结果

Parameters (p, q, P, Q)	BIC
(1, 1, 1, 1)	2676.883140
(1, 1, 1, 2)	2679.304953
(1, 1, 2, 1)	2679.465158
(1, 2, 1, 1)	2681.079566
(2, 1, 1, 1)	2681.618812
(1, 2, 1, 2)	2683.516432
(1, 2, 2, 1)	2683.703928
(2, 2, 1, 1)	2683.779918
(2, 1, 1, 2)	2683.977101
(2, 1, 2, 1)	2684.131312

在选定模型参数后对模型进行白噪声检验，得到检验结果见表 5，得到的 p 值均大于 0.05，可知残差序列是不相关的，模型通过检验。

Table 5. Results of residual white noise test
表 5. 残差白噪声检验结果

Box-Pierce
Test Data: resid, Lag-order = 24, df = 1
p-value = [0.15031, 0.35427, 0.52418, 0.59038, 0.72363, 0.82611, 0.88548, 0.91305, 0.95013, 0.97096, 0.95533, 0.06891, 0.08751, 0.12036, 0.15937, 0.20540, 0.24745, 0.30279, 0.36108, 0.41929, 0.47971, 0.49143, 0.53965, 0.54923]

根据选定的 SARIMA 模型对肺结核的发病数序列进行拟合与预测，拟合与预测结果见图 6。其中，深色曲线表示实际数据，浅色虚线表示模型对未来 28 个月的预测值。从图中可以看出，拟合值与实际值较为接近，但仍有一定的误差。

实际值与预测值的误差结果见表 6，可以得到模型预测值与实际值之间的最大误差为 23,455，最小预测误差为 77。SARIMA 模型评价指标结果见表 7，SARIMA 模型的平均绝对误差为 5965.75，均方根误差为 8344.06，平均绝对百分比误差为 17.95%。SARIMA 模型的拟合效果欠佳，这表明 SARIMA 建模仍有改进的空间。

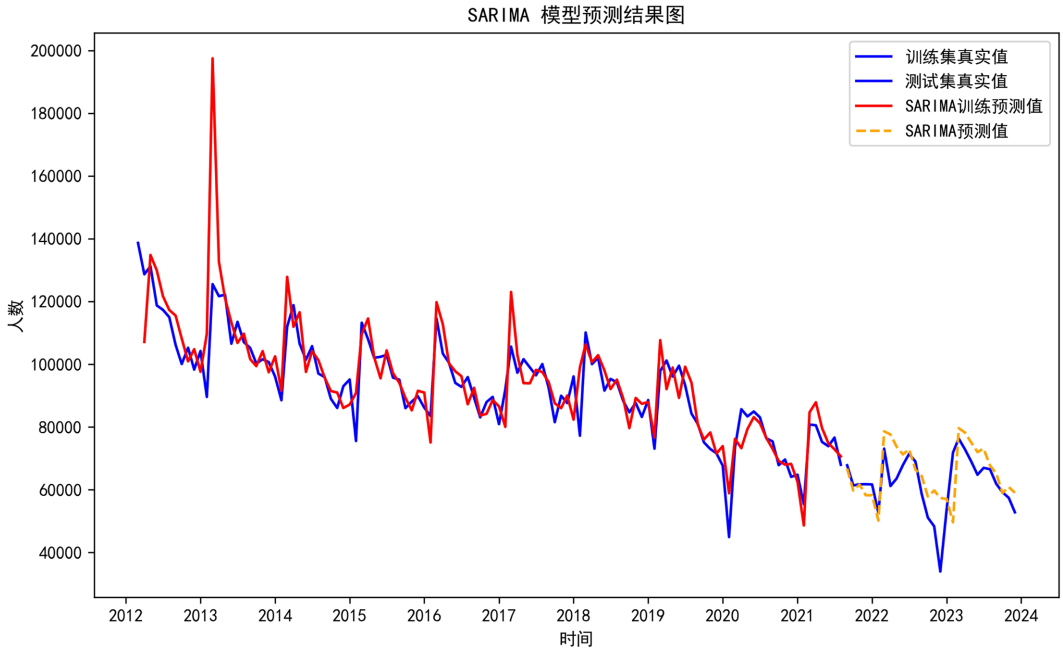

Figure 6. Fitting and prediction results of SARIMA model on the incidence of pulmonary tuberculosis from 2012 to 2023
图 6. SARIMA 模型对 2012~2023 年肺结核发病数的拟合与预测结果

3.2. LSTM 模型结果

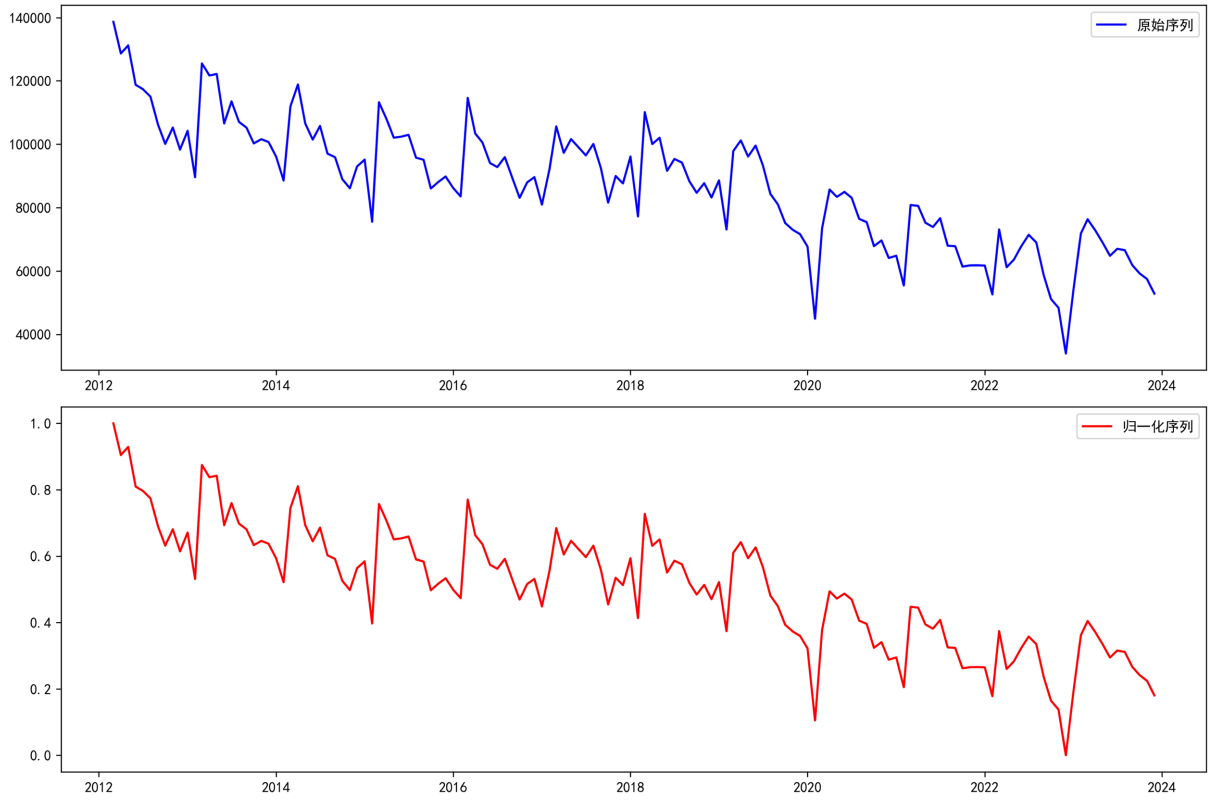

Figure 7. Normalization of the incidence sequence of pulmonary tuberculosis from 2012 to 2023
图 7. 2012~2023 年肺结核的发病数序列归一化

对肺结核发病数序列进行归一化处理，处理后的数据如图 7 所示，数据的整体趋势没有发生变化，但是数据之间的差距缩小，归一化后数据范围在 0~1 之间。

将数据进行归一化后，设置 LSTM 模型参数。将均方误差作为损失函数，神经元数设置为 200，激活函数为 tanh，优化函数为 Adam，学习率为 0.001，每次训练批量数设置为 16，训练轮数设置为 1000，时间步长设置为 12，训练得到的结果进行反归一化处理，得到 LSTM 的预测结果。在训练过程中得到的损失函数见图 8，可以发现损失函数值不断减小，最后达到收敛。

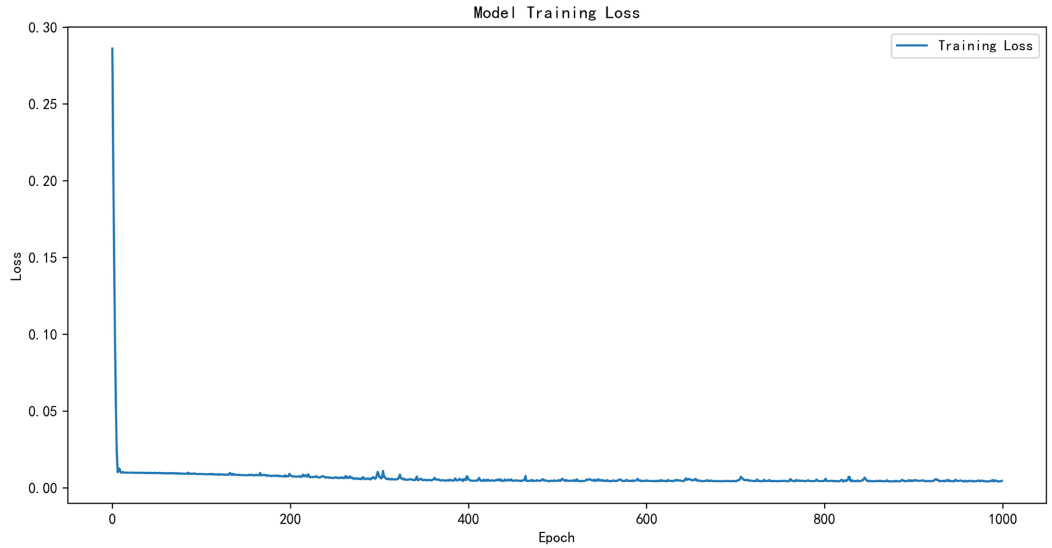

Figure 8. Loss function curve of LSTM model
图 8. LSTM 模型的损失函数曲线

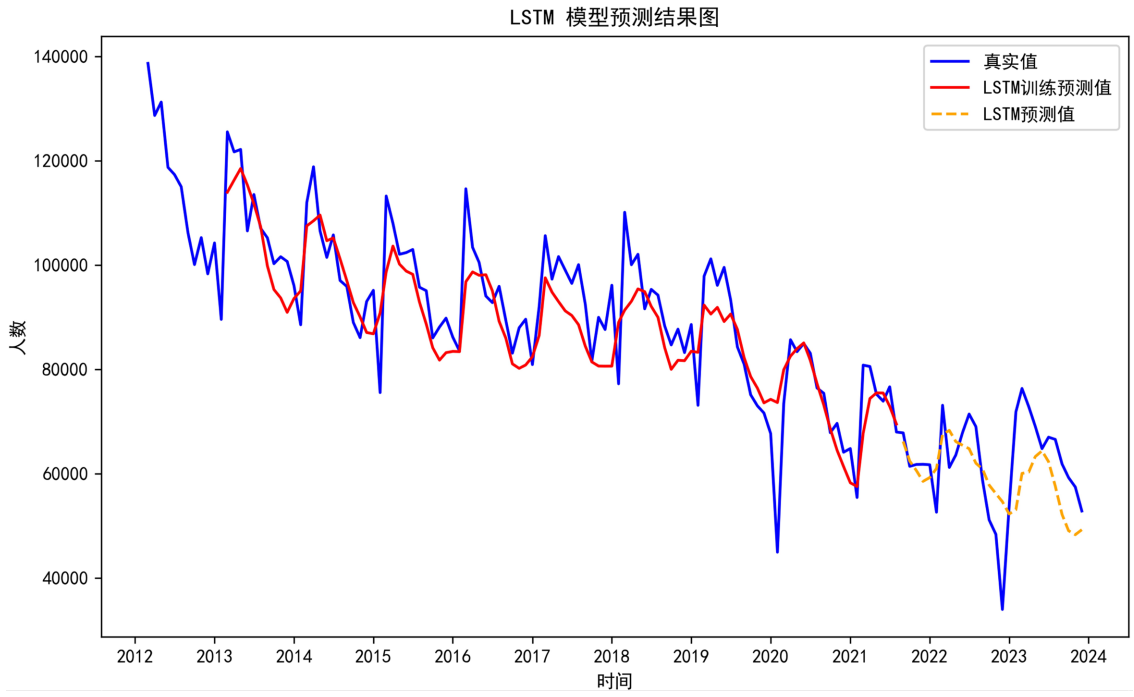

Figure 9. Fitting and prediction results of LSTM model on the incidence of pulmonary tuberculosis from 2012 to 2023
图 9. LSTM 模型对 2012~2023 年肺结核发病数的拟合与预测结果

使用训练后的模型进行预测，得到 LSTM 模型的预测结果见图 9。模型预测实际值与预测值的误差见表 6，模型预测值与真实值之间的最大误差为 21,492，最小预测误差为 29。LSTM 模型的预测效果比 SARIMA 模型更好，预测值与实际值更接近，预测误差更小。

LSTM 模型评价指标结果见表 7，LSTM 模型的平均绝对误差为 6265.93；均方根误差为 8193.73；平均绝对百分比误差为 14.62%，比 SARIMA 模型降低了 18.61%。由此可见，LSTM 模型比 SARIMA 模型的拟合效果好，预测精度仍有进一步提高的空间。

3.3. SARIMA-LSTM 模型结果

使用均方误差(Mean Squared Error, MSE)作为损失函数，其衡量的是预测值与实际观测值之间差异的平方平均数。MSE 越小，说明预测结果越接近真实值，即预测越准确。使用 Python 的 SciPy 库提供的 minimize 函数进行优化，经过一系列迭代后，minimize 函数返回一组使得损失值最小化的权重，即为最优权重。组合模型的最优权重为：SARIMA 模型的权重为 0.54823307，LSTM 模型的权重为 0.43115494。根据最优权重对预测值进行加权求和，得到 SARIMA-LSTM 组合模型的预测结果，见图 10。

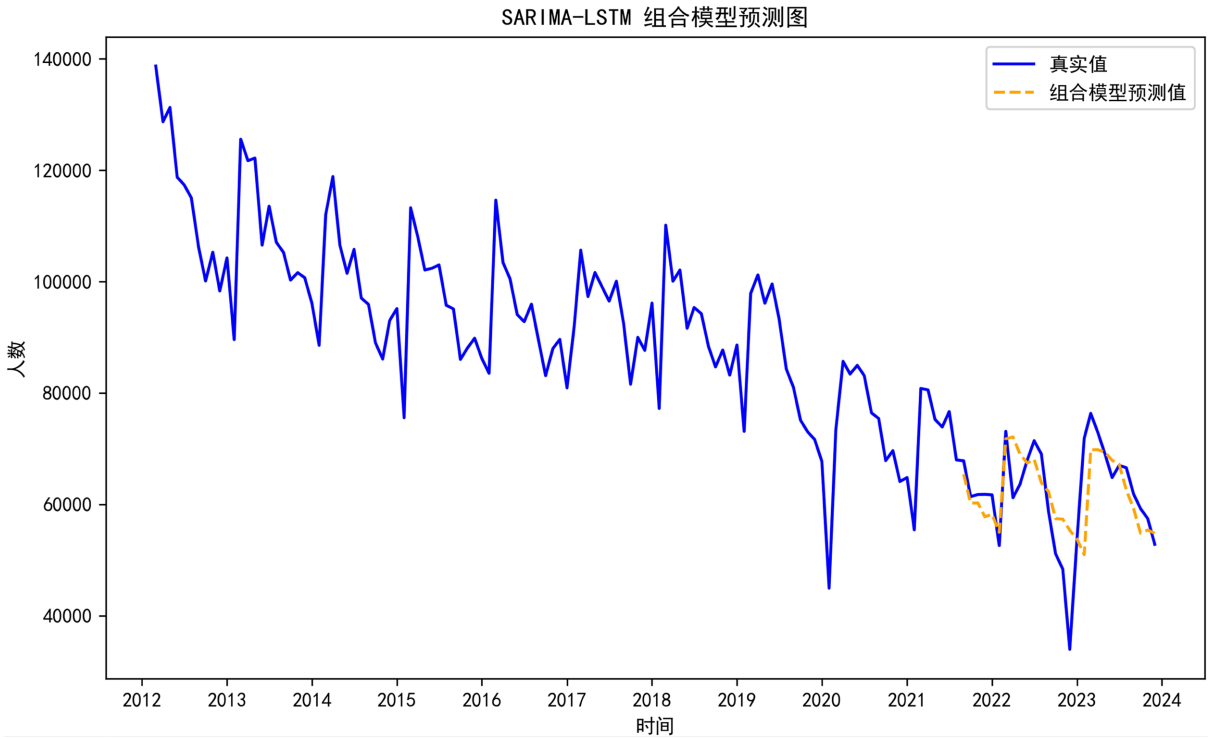

Figure 10. SARIMA-LSTM combination model prediction diagram
图 10. SARIMA-LSTM 组合模型预测图

SARIMA-LSTM 组合模型对于肺结核发病数的预测值与实际值的误差见表 6，最小误差为 165，最大误差为 21,439。比较单一模型和组合模型的预测误差，组合模型的预测误差在三个模型中最小。

SARIMA-LSTM 组合模型的评价指标结果见表 7，平均绝对误差为 4667.13，比 SARIMA 模型降低了 21.77%，比 LSTM 模型降低了 25.52%；均方根误差为 7011.52，比 SARIMA 模型降低了 15.97%，比 LSTM 模型降低了 14.43%；平均绝对百分比误差为 8.49%，比 SARIMA 模型降低了 52.70%，比 LSTM 模型降低了 41.89%。

Table 6. Comparison of prediction errors among three models
表 6. 三种模型预测误差对比

年－月	SARIMA	LSTM	SARIMA-LSTM	年－月	SARIMA	LSTM	SARIMA-LSTM
2021-09	1045	546	271	2022-11	−11409	−8612	−8698
2021-10	1655	−2106	−2418	2022-12	−23455	−21492	−21439
2021-11	77	29	−165	2023-01	−3318	699	1019
2021-12	3547	1997	1794	2023-02	22211	17752	17871
2022-01	3450	1012	822	2023-03	−3299	15126	14911
2022-02	2422	−9931	−10127	2023-04	−5383	11311	10780
2022-03	−5533	3840	4035	2023-05	−6010	4739	4178
2022-04	−16436	−8672	−8750	2023-06	−7221	−428	−880
2022-05	−10166	−3823	−3947	2023-07	−6136	3975	3664
2022-06	−3353	1376	1222	2023-08	−1186	8081	7803
2022-07	−1495	5652	5420	2023-09	−3318	8477	8231
2022-08	2574	6048	5647	2023-10	265	8877	8694
2022-09	−5690	−3133	−3551	2023-11	−3438	7840	7662
2022-10	−6725	−7551	−7842	2023-12	−6224	2321	2098

Table 7. Comparison of prediction evaluation indicators among three models
表 7. 三种模型预测评价指标对比

模型	MAE	RMSE	MAPE (%)
SARIMA	5965.75	8344.06	17.95
LSTM	6265.93	8193.73	14.62
SARIMA-LSTM	4667.13	7011.52	8.49

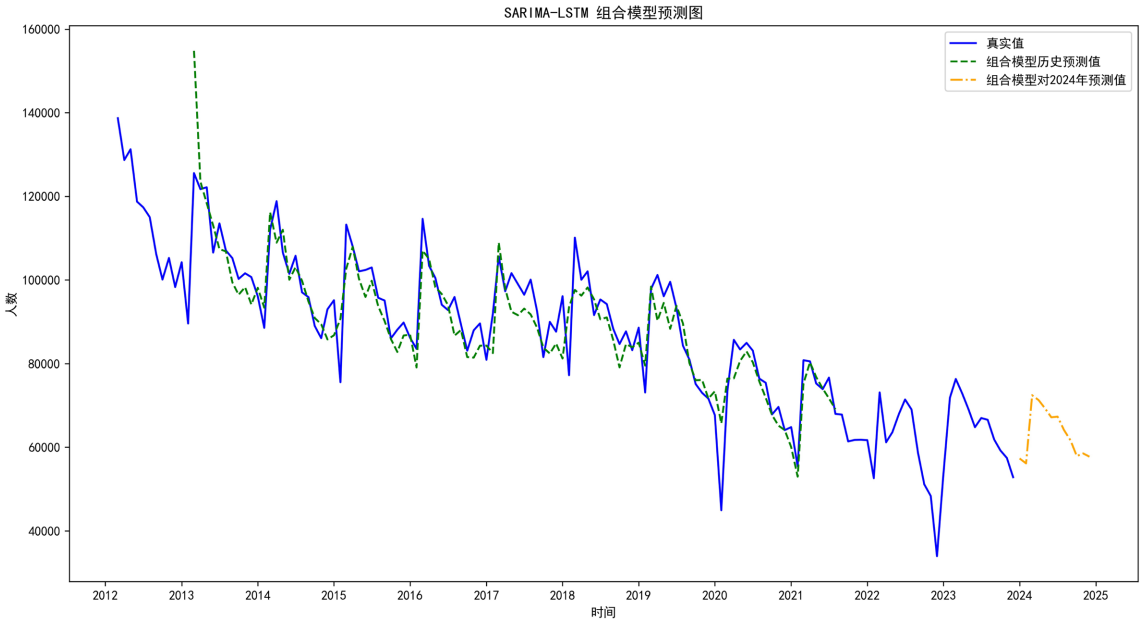

Figure 11. SARIMA-LSTM combination model for predicting the incidence of pulmonary tuberculosis in 2024
图 11. SARIMA-LSTM 组合模型对 2024 年肺结核发病数预测图

由此可见，SARIMA-LSTM 组合模型的预测精度较单一模型有所提升，更适合预测未来的发病数。根据 2012~2023 年的已有数据，对 2024 年 1~12 月肺结核发病数进行预测，得到预测图见图 11，预计在 2024 年肺结核发病会出现先增后降的趋势，在 3 月会出现肺结核发病的高峰，在下半年肺结核的发病数会缓慢降低。

4. 结论

传染病预测有助于疾病发病的早期预警，可以提升传染病防控工作的主动性，对于抑制传染病的流行和暴发具有重要意义。对于任何一个地区来说，传染病流行的影响因素错综复杂，如地理条件、气候条件、政策出台、人群流动以及其他不可控的因素。肺结核的发病具有季节性特征，11~12 月、2 月的发病数较低，3~5 月发病数高于平均水平，原因在于 3 月天气转暖，病原微生物开始大量繁殖以及春节后复工复学扩大了人口流动等。

SARIMA 模型是传染病预测中最经典的时间序列方法之一，已被证明具有较高的准确性，常作为新模型的评价基础。结果表明，最优模型 SARIMA(1, 1, 1) × (1, 1, 1, 12)的预测性能仍存在一定的缺陷，LSTM 模型也无法完全捕捉数据的趋势，单一的模型效果不够精确。针对这两种模型的预测优势，本研究提出组合模型 SARIMA-LSTM。采用权重数组 weights 决定 SARIMA 模型和 LSTM 模型在组合预测中的权重比例。通过调整权重数组的数值，可以控制不同模型在最终预测中的影响程度。这种加权的方式能够充分利用 SARIMA 模型和 LSTM 模型各自的优势，并根据其性能表现调整预测结果的权重。研究表明，SARIMA-LSTM 组合模型能够更准确地预测肺结核的发病数，在三种模型中整体预测误差最小。为了评估模型的预测效果，采用 MAE、RMSE、MAPE 三种评价指标比较了 SARIMA、LSTM 和组合模型的预测结果，组合模型的 MAE、RMSE、MAPE 值在三个模型中均为最低，表明组合模型的预测精度最高。更准确的疾病发病预测有助于分配公共卫生资源和增强早期预防意识。

本研究也有一些局限性。中国各地区的肺结核发病率存在差异，SARIMA-LSTM 组合模型的适用性和适用范围有待进一步研究。影响肺结核发病的因素较多，如地理、经济、人口等因素，本研究暂未考虑这些因素，未来我们会将更多的影响因素纳入模型，研究其预测效果，建立更准确的预测模型。

致　谢

本研究得到北京建筑大学“(2024 年)分类发展定额项目——研究生教学研究与质量提升项目”(No. 31081024005)、山东省自然科学基金——针对单细胞 RNA-seq 数据的转录组组装算法设计及在肝癌数据的应用研究(No. ZR2023QA059)的支持。感谢为本研究慷慨分享时间，提供材料和帮助的所有人。

参考文献

[1] Dheda, K., Barry, C.E. and Maartens, G. (2016) Tuberculosis. *The Lancet*, **387**, 1211-1226. https://doi.org/10.1016/s0140-6736(15)00151-8

[2] World Health Organization (2023) Global Tuberculosis Report 2023. World Health Organization.

[3] World Health Organization (2015) Implementing the End TB Strategy. World Health Organization.

[4] 徐晓岭, 王磊. 统计学[M]. 北京: 人民邮电出版社, 2015.

[5] Ariff, M.R.A., Rafdzah, Z.A., Rozita, W.M.W., *et al*. (2023) Forecasting New Tuberculosis Cases in Malaysia: A Time-Series Study Using the Autoregressive Integrated Moving Average (ARIMA) Model. *Cureus*, **15**, e44676.

[6] Munshi, R.M., Khayyat, M.M., Ben Slama, S. and Khayyat, M.M. (2024) A Deep Learning-Based Approach for Predicting COVID-19 Diagnosis. *Heliyon*, **10**, e28031. https://doi.org/10.1016/j.heliyon.2024.e28031

[7] Hong, S., Woo, S., Kim, S., Park, J., Lee, M., Kim, S., *et al*. (2024) National Prevalence of Smoking among Adolescents at Tobacco Tax Increase and COVID-19 Pandemic in South Korea, 2005-2022. *Scientific Reports*, **14**, Article No. 7823.

https://doi.org/10.1038/s41598-024-58446-4

[8] Wang, Y., Wang, L., Ma, W., Zhao, H., Han, X. and Zhao, X. (2024) Development of a Novel Dynamic Nosocomial Infection Risk Management Method for COVID-19 in Outpatient Settings. *BMC Infectious Diseases*, **24**, Article No. 214. https://doi.org/10.1186/s12879-024-09058-w

[9] Wan, Y., Song, P., Liu, J., Xu, X. and Lei, X. (2023) A Hybrid Model for Hand-Foot-Mouth Disease Prediction Based on ARI-MA-EEMD-LSTM. *BMC Infectious Diseases*, **23**, Article No. 879. https://doi.org/10.1186/s12879-023-08864-y

[10] Meng, P., Huang, J. and Kong, D. (2022) Prediction of Incidence Trend of Influenza-Like Illness in Wuhan Based on ARIMA Model. *Computational and Mathematical Methods in Medicine*, **2022**, Article ID: 6322350. https://doi.org/10.1155/2022/6322350

[11] Yang, Y., Guo, C., Liu, L., Zhang, T. and Liu, W. (2016) Seasonality Impact on the Transmission Dynamics of Tuberculosis. *Computational and Mathematical Methods in Medicine*, **2016**, Article ID: 8713924. https://doi.org/10.1155/2016/8713924

[12] Zhu, H., Chen, S., Liang, R., Feng, Y., Joldosh, A., Xie, Z., *et al*. (2023) Study of the Influence of Meteorological Factors on HFMD and Prediction Based on the LSTM Algorithm in Fuzhou, China. *BMC Infectious Diseases*, **23**, Article No. 299. https://doi.org/10.1186/s12879-023-08184-1

[13] Yadav, B.K., Srivastava, S.K., Arasu, P.T. and Singh, P. (2023) Time Series Modeling of Tuberculosis Cases in India from 2017 to 2022 Based on the SARIMA-NNAR Hybrid Model. *Canadian Journal of Infectious Diseases and Medical Microbiology*, **2023**, Article ID: 5934552. https://doi.org/10.1155/2023/5934552

[14] Hayat, C. and Soenandi, I.A. (2018) The Hybrid-Model Architectural Modelling Based on ARIMA-BPNN Methods for Building Materials Demands Forecasting. *MATEC Web of Conferences*, **204**, Article No. 02003. https://doi.org/10.1051/matecconf/201820402003

[15] http://www.chinacdc.cn

[16] Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015) Time Series Analysis: Forecasting and Control. 5th Edition, Wiley.

[17] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[18] Zhao, Z.Y., Zhai, M.M., Li, G.H., Gao, X., Song, W., Wang, X., *et al*. (2023) Study on the Prediction Effect of a Combined Model of SARIMA and LSTM Based on SSA for Influenza in Shanxi Province, China. *BMC Infectious Diseases*, **23**, Article No. 71. https://doi.org/10.1186/s12879-023-08025-1