

基于 SARIMA – LSTM 模型的肾综合征出血热发病率预测研究^{*}

唐诗诗¹ 李宇轩¹ 唐圣晟¹ 刘庆华² 周 毅¹

(¹ 中山大学中山医学院 广州 510080 ² 中山大学附属第一医院肾内科 广州 510080)

〔摘要〕 目的/意义 探究前沿技术在肾综合征出血热发病率预测中的应用，梳理、组合多种时序分析方法，评价并筛选最佳模型。方法/过程 利用 2004—2020 年全国肾综合征出血热发病率数据，基于统计学方法的 SARIMA、STL – ARIMA、TBATS 模型，基于神经网络的 NNAR、LSTM 模型，基于 3 种加权方式的 SARIMA – LSTM 组合模型进行预测，运用 RMSE、MAE、MAPE 综合评价模型效果。结果/结论 SARIMA、LSTM 在单一模型中较优；SARIMA – LSTM 组合模型效果相较单一模型均有提升；基于误差倒数法的 SARIMA – LSTM 组合模型为最优模型。本研究有望为肾综合征出血热发病预警系统模型设计提供技术支持与参考。

〔关键词〕 肾综合征出血热；传染病监测预警；统计学模型；机器学习；SARIMA – LSTM 模型
〔中图分类号〕 R – 058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673 – 6036.2024.08.012

Study on the Prediction of Incidence of Hemorrhagic Fever with Renal Syndrome Based on the SARIMA – LSTM Model
TANG Shishi¹, LI Yuxuan¹, TANG Shengsheng¹, LIU Qinghua², ZHOU Yi¹
¹ Zhongshan School of Medicine, Sun Yat – sen University, Guangzhou 510080, China; ² Department of Nephrology, The First Affiliated Hospital, Sun Yat – sen University, Guangzhou 510080, China

〔Abstract〕 **Purpose/Significance** To investigate the application of cutting – edge technologies in predicting the incidence of hemorrhagic fever with renal syndrome (HFRS), to compile and integrate various time – series analysis methods, evaluate and select the optimal model. **Method/Process** By utilizing national HFRS incidence data from 2004 to 2020, the effectiveness of models is predicted based on statistical methods: SARIMA, STL – ARIMA and TBATS, neural network approaches: NNAR, LSTM and combined models of SARIMA – LSTM with 3 different weighting schemes. The performance of these models is comprehensively assessed using RMSE, MAE and MAPE. **Result/Conclusion** The SARIMA and LSTM models are identified as the superior individual models. The combined SARIMA – LSTM model demonstrates enhanced performance compared to individual models. The SARIMA – LSTM model optimized using the reciprocal of error method is deemed the optimal model. The optimal model is expected to provide technical support and references for the early warning system model design of HFRS.

〔Keywords〕 hemorrhagic fever with renal syndrome (HFRS); infectious disease surveillance and early warning; statistical model; machine learning; SARIMA – LSTM model

〔修回日期〕 2024 – 07 – 22
〔作者简介〕 唐诗诗，硕士研究生，发表论文 1 篇；通信作者：周毅，博士，教授，博士生导师。
〔基金项目〕 国家重点研发计划项目（项目编号：2022YFC3601600）；广东省自然科学基金项目（项目编号：2024A1515011989）；中山大学高校基本科研业务费项目（项目编号：24xkj025）。

1 引言

传染病暴发构成重大公共安全挑战，对人类社会造成深远影响。随着信息技术的发展，以数字化和网络化为标志的现代信息技术成为突发公共卫生事件应急管理现代化与智能化的基础^[1]。传染病预测模型通过分析疾病特性与传播模式，基于历史数据预测未来趋势，有助于了解疾病的传播机制，并为防控措施提供理论依据^[2]。肾综合征出血热(hemorrhagic fever with renal syndrome, HFRS)是严重影响我国公共卫生的乙类传染病，近年来全国所有省市均有病例报告^[3-4]，对全社会持续构成威胁，且其呈现复杂的地区分布特征，加剧了疫情防控的复杂性。因此，开展更为精准的 HFRS 发病率预警模型研究显得尤为迫切与重要。

国内 HFRS 预测模型研究多针对局域地区，且算法模型创新和疾病种类相对较少，多数研究仅使用 2~3 种模型预测方法^[5]，在多模型比较、机器学习组合模型构建方面探索相对有限，难以满足高效、精确预测需求。本研究创新地应用组合模型，为 HFRS 防控提供有力支持，并为传染病预警系统的优化及公共卫生政策制定提供科学依据。

2 资料与方法

2.1 资料来源

资料来源于公共卫生科学数据中心 (<http://www.phsciencedata.cn/>)，包括全国 HFRS 按地区分布的发病数、死亡数、发病率等数据，整合 2004—2020 年全国 HFRS 逐月发病率资料。以 2004 年 1 月—2018 年 12 月全国 HFRS 发病率数据作为训练数据，以 2019 年 1 月—2020 年 12 月数据作为测试数据。选择均方根误差 (root mean square error, RMSE)、平均绝对误差 (mean absolute error, MAE)、平均绝对百分比误差 (mean absolute percentage error, MAPE) 评估模型预测效果。

2.2 方法

主要工具为 R 语言和 Python 语言，实现季节性

自回归积分滑动平均 (seasonal auto-regressive integrated moving average, SARIMA)、STL (seasonal-trend decomposition procedure based on loess)、TBATS (trigonometric seasonality, Box-Cox transformation, ARMA errors, trend and seasonal components)、神经网络自回归 (neural network autoregression, NNAR) 和长短时记忆网络 (long short-term memory network, LSTM) 模型的建立，基于预测效果组合建立 SARIMA-LSTM 模型，最后综合 3 种损失函数评价模型效果。研究技术路线框架，见图 1。

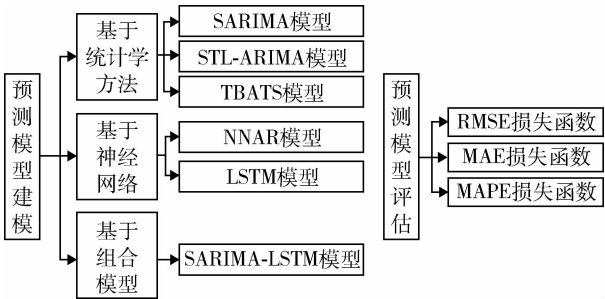


图 1 研究技术路线框架

2.2.1 SARIMA 模型 SARIMA 模型一般形式为 $SARIMA(p, d, q)(P, D, Q)_s$ 。本研究建模思路：第 1 步，原始序列平稳性检验，非平稳则进行差分运算；第 2 步，若原始序列或差分后序列平稳，则展开白噪声检验；第 3 步，模型为非白噪声序列并且具有季节性，则考虑拟合 SARIMA 模型，进行模型定阶、参数估计和择优；第 4 步，拟合模型残差的白噪声检验；第 5 步，选择最佳模型展开 HFRS 发病率预测。

2.2.2 STL-ARIMA 模型 针对 HFRS 发病率的预测，还有一种在时间序列预测领域中常见的思路，即“先分解，后预测”。STL 是一种通用且稳健的时间序列分解方法^[6]，将时序数据分解为季节、趋势和残差 3 个分量，分别展开分析和预测。首先通过 STL 分解数据，再利用自回归差分移动平均 (auto-regressive integrated moving average, ARIMA) 模型建模预测，最终将 3 个分解项的预测结果加总为最终预测结果。

2.2.3 TBATS 模型 针对 HFRS 发病率时间序列的预测，TBATS 模型包含傅里叶三角函数季节性、Box-Cox 转换、ARMA 残差、趋势项和季节项。相

比 STL 和 ARIMA 模型, TBATS 具有处理多种复杂季节性模式时间序列的优势^[7], 对传统的指数平滑模型进行了改进, 求解复杂周期性时间序列问题^[8], 能够对季节性因素进行不同程度的均衡。

2.2.4 NNAR 模型 可以视为描述复杂非线性关系和函数形式的神经元或节点网络, 具有动态特性和高容错性能^[9], 基于神经网络框架, 主要原理是利用过去观测值为输入, 预测未来观测值, 通常由多个前馈神经网络层组成, 以是否包含隐藏层区分为简单神经网络(等价线性回归)和含隐藏层的神经网络, 通过添加隐藏层, 神经网络则变得非线性。

2.2.5 LSTM 模型 近年来, 循环神经网络(recurrent neural network, RNN)常用于时间序列的分析与预测, 能够获得很高的效率^[10]。1997 年 LSTM 模型首次被提出^[11], 主要解决梯度消失问题。LSTM 是基于门控制单元结构的深度学习模型, 可以用于多种时序预测问题, 本研究利用 LSTM 对 HFRS 发病率数据开展单变量时间序列预测。

2.2.6 SARIMA-LSTM 模型 SARIMA 模型可以有效捕捉序列长、短期趋势和周期特征, LSTM 模型则适用于处理数据集中复杂的非线性和动态特征, 因此利用 SARIMA 和 LSTM 的组合模型预测 HFRS 发病率, 可以同时处理线性和非线性的分量, 有效提升模型预测能力。选择并联模式组合模型, 其流程, 见图 2。

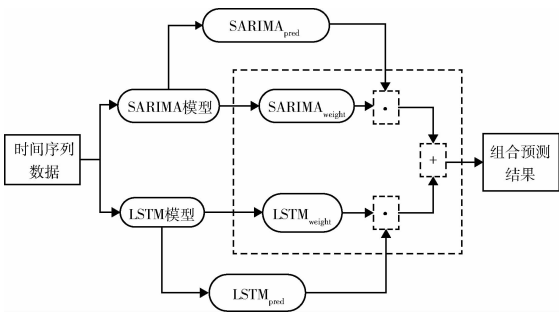


图2 SARIMA-LSTM 并联组合模型流程

注: SARIMA_{pred} 和 LSTM_{pred} 分别表示 SARIMA 和 LSTM 模型的预测值; SARIMA_{weight} 和 LSTM_{weight} 分别表示模型组合时对应的权重。

作为定结构组合预测方法, 其中单一预测模型的种类以及用于模型组合的权重均不变, 通用表达

式, 见公式 (1)。其中, w_i 为第 i 个预测模型的权重 ($\sum_{i=1}^n w_i = 1$), $f_i(t)$ 为第 i 个预测模型在 t 时刻的预测值, n 为参与组合的模型个数, Hybrid_{pred} [t] 为加权不同模型得到的组合模型在 t 时刻的预测值。

$$\text{Hybrid}_{\text{pred}}[t] = \sum_{i=1}^n w_i f_i(t), i = 1, 2, \dots, n \tag{1}$$

模型权重的确定方法有以下 3 种, 均满足 n 为参与组合的模型个数, w_i 为第 i 个预测模型的权重, 且 $\sum_{i=1}^n w_i = 1$ 的条件^[12]。一是等权平均法, 即各预测模型的权重相等, 见公式 (2)。二是误差倒数法, 利用各预测模型相对误差的倒数确定权重, 见公式 (3)。三是 Critic 客观权重赋值法, 是一种基于训练数据的集成学习方法, 根据预测结果与真实标签间的误差大小, 分配组合权重, 见公式 (4)。

$$w_i = \frac{1}{n}, i = 1, 2, \dots, n \tag{2}$$

$$w_i = \frac{\varepsilon_i^{-1}}{\sum_{j=1}^n \varepsilon_j^{-1}}, i = 1, 2, \dots, n \tag{3}$$

$$w_i = \frac{\text{RMSE}_i^{-1}}{\sum_{j=1}^n \text{RMSE}_j^{-1}}, i = 1, 2, \dots, n \tag{4}$$

3 结果

3.1 HFRS 发病率概况分析

使用加法模型对原始数据进行时间序列分解, 见图 3。

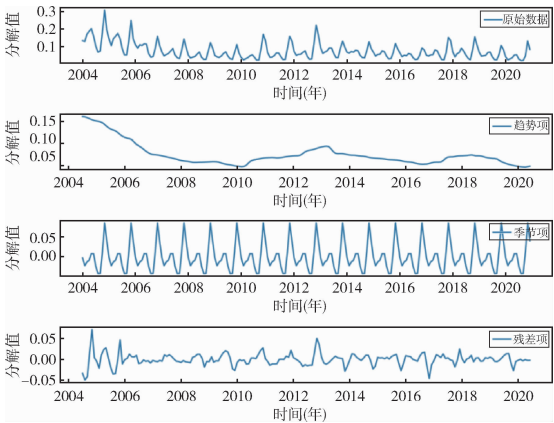


图3 2004—2020 年全国 HFRS 发病率时间序列分解

由季节项分解得知 HFRS 呈明显的季节性，且以 12 个月为周期，这与大多数乙类传染病的规律相似；每年有两个高峰期，第 2 个高峰的发病率明显高于第 1 个，也即 HFRS 常见发病于冬春季节。

3.2 SARIMA 模型建立

利用 ADF 和 KPSS 两种统计方法检验序列平稳性，结合两种方法得知序列非平稳。由 Ljung – Box 检验得到序列为非白噪声，考虑其季节性，对数据进行以 12 个月为周期的季节差分。根据序列特征和滞后项，并结合 SARIMA 模型相关参数取值一般

不超过 5 的规律^[13]，利用 R 中 auto. arima () 函数，在默认区间内自动搜索参数，设定搜索的 p 、 q 、 P 、 Q 的最大值为 5，设定 stepwise = TRUE，通过逐步法选择模型，搜索区间内的每一对由 KPSS 检验确认的 d 和 D 都将使用 ARIMA (p , d , q) (P , D , Q)。模型拟合给定的时间序列数据，并使用信息准则确定最佳模型。得到 HFRS 发病率数据的最优模型为 SARIMA (2, 1, 1) (1, 1, 0)₁₂。其中最佳模型对应的 AIC = - 1 017. 626，Ljung – Box 检验的 P 值为 0. 906，残差为白噪声。得到拟合模型诊断图，见图 4。

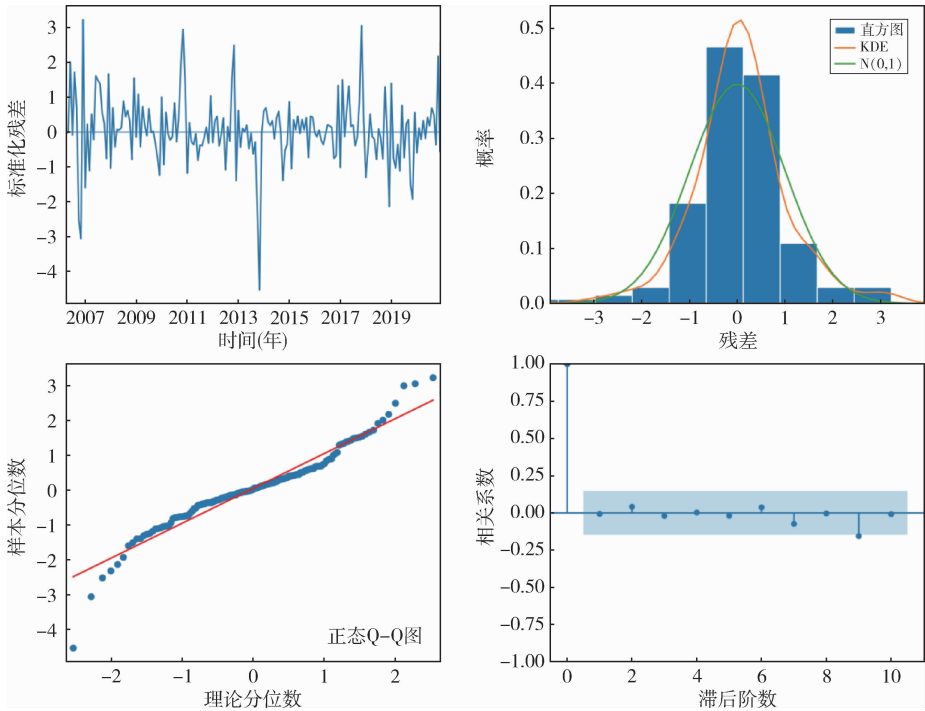


图 4 SARIMA 模型诊断

模型拟合得到的序列残差随着时间的变化基本稳定；核密度估计 (kernel density estimation, KDE) 曲线走势紧随标准正态分布；正态 Q – Q (quantile – quantile) 图也印证了残差是正态分布的；最后残差序列自相关图显示其不存在自相关。该模型具有较好的拟合效果，可用于进一步预测。真实值与预测值的比较，见图 5。

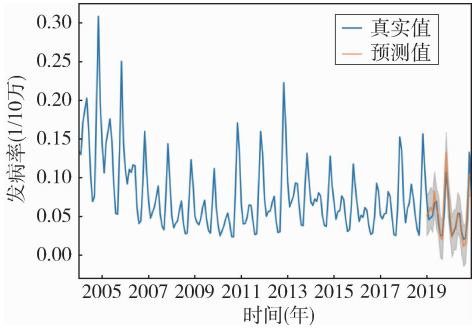


图 5 测试集上 SARIMA 模型真实值与预测值比较

3.3 STL-ARIMA 模型建立

使用 R 中 stlf () 函数实现 STL-ARIMA，利用网格搜索法确认 s.window 的最优取值为 5，自动选参函数在默认区间搜索得到最佳模型为 ARIMA (2, 1, 1)，AIC = -1 230.45，残差序列 Ljung - Box 检验的 P 值为 0.773，为白噪声。真实值与预测值的比较，见图 6。

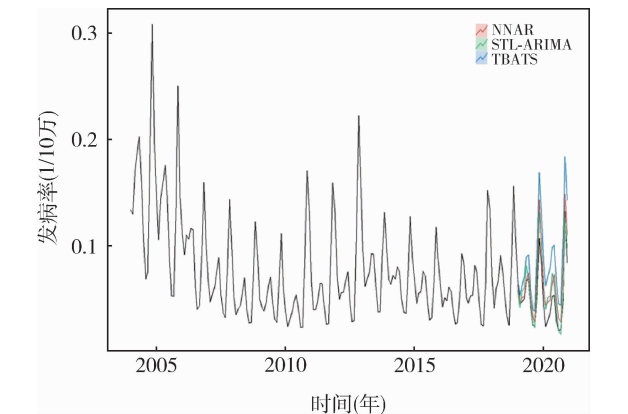


图 6 测试集上 STL-ARIMA、TBATS、NNAR 模型真实值与预测值比较

3.4 TBATS 模型建立

利用 R 中 tbats () 函数建立 TBATS 模型，根据 AIC 最小化原则自动确定参数得到：TBATS (0.295, {3, 0}, 1, { <12, 5 > })，其中 0.295 为 Box - Cox 变换的参数，ARMA 模型的 $p = 3$ 、 $q = 0$ ，季节周期长度为 12，傅里叶级数为 5，模型对应的 AIC = -737.60，残差序列 Ljung - Box 检验的 P 值为 0.884，为白噪声。真实值与预测值的比较，见前文图 6。

3.5 NNAR 模型建立

对于 NNAR 模型，由于 HFRS 发病率数据具有季节性，因此增加同一个季节的最后观测值为输入。利用 R 中 nnetar () 函数建模，通过交叉验证法得到 λ 为 0.6，函数拟合得到最佳模型为 NNAR (4, 1, 15)₁₂。残差序列 Ljung - Box 检验的 P 值为 0.155，为白噪声。真实值与预测值的比较，见前文图 6。

3.6 LSTM 模型建立

利用 Python 构建 LSTM 模型，主要建模步骤如下。第 1 步，数据准备：将数据以 12 个为一组切分，形成多个时间窗口，将这些时间窗口作为输入数据，下一个时间点的数据则作为输出，正则化处理并划分训练集、测试集。第 2 步，模型构建：形成 [样本，时间步长，特征数] 格式的数据，为减少计算复杂度设置时间步长为 1，特征为周期 12。第 3 步，模型训练：关键参数对训练集 RMSE 的影响，见表 1，通过比较确定最佳参数，其中，隐藏层中设置 6 个神经元、批量大小设置为 1、设置优化器为 Adam，模型的损失函数为均方误差。第 4 步，模型预测：利用已训练模型对测试数据展开预测。第 5 步，误差计算：将预测数据反归一化，即转换为同一单位，计算相应误差指标。真实值与预测值的比较，见图 7。

表 1 关键参数对预测准确率的影响			
隐藏层神经元	批量大小	优化器	RMSE
5	1	Adam	0.010 64
6	1	Adam	0.010 30
7	1	Adam	0.011 16
6	1	Adam	0.010 30
	16	Adam	0.015 79
	32	Adam	0.016 15
	64	Adam	0.023 19
6	1	SGD	0.017 55
		Adam	0.010 30
		Adagrad	0.029 68
		RMSprop	0.011 09

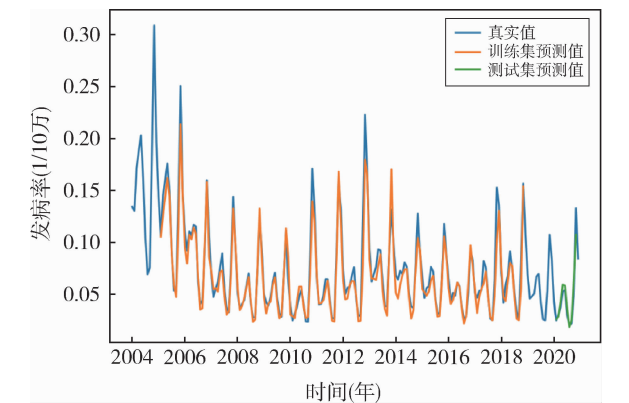


图 7 测试集上 LSTM 模型真实值与预测值比较

3.7 SARIMA – LSTM 模型建立

利用 Python 构建组合模型，基于等权平均法、误差倒数法和 Critic 客观权重赋值法的组合模型分别命名为并联组合 1、并联组合 2 和并联组合 3，3 种并联组合模型的权重，见表 2。单一模型以及 SARIMA – LSTM 组合模型的预测结果，见图 8。

表 2 3 种并联组合模型权重

组合类别	模型	权重
并联组合 1	SARIMA	0.50
	LSTM	0.50
并联组合 2	SARIMA	0.47
	LSTM	0.53
并联组合 3	SARIMA	0.45
	LSTM	0.55

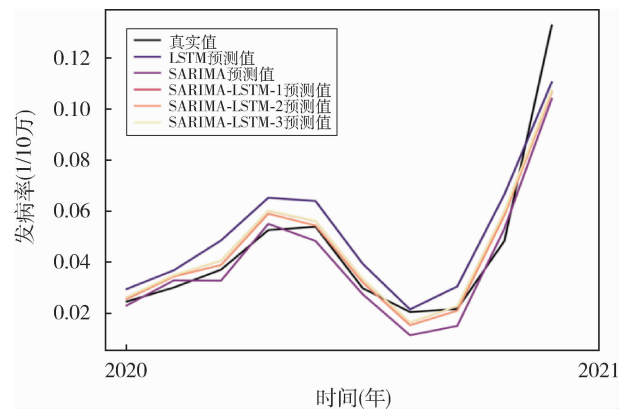


图 8 测试集上 SARIMA – LSTM 组合模型真实值与预测值比较

3.8 预测模型比较与评价

分别计算各模型在测试集上的 RMSE、MAE 和 MAPE 指标值，评价各模型预测效果，见表 3。比较 HFRS 预测模型效果时，通常 RMSE < 0.01 认为预测效果较优，RMSE 越小，预测效果越好；Pao H T^[14] 提出 MAPE 为 5% ~ 10% 时预测模型高度准确，为 10% ~ 20% 时被认为较好，为 20% ~ 50% 时表示模型可接受；3 项指标均满足值越小，预测效果越好。通过以上比较规则，综合图像以及表 3 中的指标值可知，对于单一模型，STL – ARIMA、TBATS、

NNAR 模型的效果中等；SARIMA 和 LSTM 模型在 HFRS 发病率预测的各项指标都较优，其中 LSTM 已经满足较优模型的指标范围。对于 SARIMA – LSTM 并联组合模型，3 种组合模型预测的各项指标相较单一模型都有一定提升，其中达到最佳效果的是并联组合 2，即以误差倒数法确定权重的 SARIMA – LSTM 并联组合模型。

表 3 预测模型在测试集上的预测效果评价

模型	RMSE	MAE	MAPE (%)
SARIMA	0.012 24	0.009 81	18.43
STL – ARIMA	0.012 89	0.010 69	22.69
TBATS	0.037 20	0.033 14	37.29
NNAR	0.016 60	0.013 57	30.27
LSTM	0.009 98	0.007 05	14.08
并联组合 1	0.009 41	0.005 22	9.55
并联组合 2	0.009 40	0.005 19	9.32
并联组合 3	0.009 42	0.005 20	9.43

4 讨论

鉴于中国在国际公共卫生领域扮演着重要角色，且传染病是中国的主要疾病负担之一^[15]，紧紧围绕多渠道全维度传染病风险自动识别与早期智能探测预警^[16]，加强传染病监测数据的时序预测模型研究，对于实时监控疫情动态并迅速实施防控举措至关重要。本研究聚焦探寻 HFRS 发病率预测的最佳模型，比较基于传统统计学方法的 SARIMA、STL – ARIMA 和 TBATS 模型，基于神经网络的 NNAR、LSTM 模型，以及 SARIMA – LSTM 组合模型在测试集上的预测效果。

实验结果显示，单一模型中 LSTM 和 SARIMA 模型的 3 项指标值均较低，相较既有研究^[17–18] 实验结果，本研究由于训练数据增加以及多种参数优化操作，预测误差更低，均在 0.01 附近；3 种方法得到的组合模型相较单一模型预测效果均有提升，RMSE 均小于 0.01，以及 MAPE 均小于 10%，属于高度准确的预测模型，最佳组合模型为基于误差倒数法的 SARIMA – LSTM 模型。SARIMA 与 LSTM 都是较好的 HFRS 发病率预测模型，并且二者的组合能够在单一模型基础上进一步降低预测误差。

5 结语

通过建立 SARIMA - LSTM 模型,可以同时利用 SARIMA 对周期、趋势的敏感性以及 LSTM 的非线性序列预测能力,实现预测精度的提升。SARIMA - LSTM 组合模型以及这种建模思路可以拓展至更多传染病的发病预测。本研究仍然存在一定局限性:一是数据受限于公开资源,可能一定程度上影响样本的充分性和时效性;二是优化的串联组合、非线性组合^[19]、集成学习^[20]等组合建模技术有待进一步探索;三是未深入分析地域性流行异质性及气象、环境等外部因素对 HFRS 的动态影响。未来将进一步拓宽数据来源,考虑多维度数据,纳入多种外部因素,以优化 HFRS 预测模型。进一步强化疫情的早期识别与反应能力,为医疗卫生资源配置的科学决策提供坚实支撑,从根本上助力传染病的防控与大流行的阻断。

作者贡献: 唐诗诗负责实验设计、数据收集与分析、论文撰写;李宇轩负责协助数据分析、论文修订;唐圣晟负责论文修订;刘庆华负责提供临床专业指导、论文审核;周毅负责提供指导、论文审核。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- SHAPIRO J S, MOSTASHARI F, HRIPCSAK G, et al. Using health information exchange to improve public health [J]. American journal of public health, 2011, 101 (4): 616 - 623.
- BRACHER J, HELD L. Endemic - epidemic models with discrete - time serial interval distributions for infectious disease prediction [J]. International journal of forecasting, 2022, 38 (3): 1221 - 1233.
- TIAN H, HU S, CAZELLES B, et al. Urbanization prolongs hantavirus epidemics in cities [J]. Proceedings of the national academy of sciences, 2018, 115 (18): 4707 - 4712.
- 邓雪飞, 杜珊珊, 黄晓霞, 等. 2004—2021 年中国肾综合征出血热报告病例流行病学特征分析 [J]. 疾病监测, 2023, 38 (1): 70 - 74.
- 何琪乐, 张瑾瑶, 吴卓存, 等. 基于互联网数据的传染病预测模型研究进展 [J]. 医学信息学杂志, 2024, 45 (2): 32 - 37.

- CLEVELAND R B, CLEVELAND W S, MCRAE J E, et al. STL: a seasonal - trend decomposition [J]. Journal of official statistics, 1990, 6 (1): 3 - 73.
- XIAO Y, LI Y, LI Y, et al. Estimating the long - term epidemiological trends and seasonality of hemorrhagic fever with renal syndrome in China [J]. Infection and drug resistance, 2021, 14 (1): 3849 - 3862.
- ABOTALEB M, MAKAROVSKIKH T, YONAR A, et al. State of the art in wind speed in England using BATS, TBATS, Holt's Linear and ARIMA model: wind speed in England using BATS, TBATS, Holt's Linear and ARIMA model [J]. Mausam, 2022, 73 (1): 129 - 138.
- YU G, FENG H, FENG S, et al. Forecasting hand - foot - and - mouth disease cases using wavelet - based SARIMA - NNAR hybrid model [J]. Plos one, 2021, 16 (2): e0246673.
- TORRES J F, MARTÍNEZ - ÁLVAREZ F, TRONCOSO A. A deep LSTM network for the Spanish electricity consumption forecasting [J]. Neural computing and applications, 2022, 34 (13): 10533 - 10545.
- HOCHREITER S, SCHMIDHUBER J. Long short - term memory [J]. Neural computation, 1997, 9 (8): 1735 - 1780.
- 杨艳, 黄晴, 龙思, 等. 基于 ARIMA - LSTM 的货运量组合预测方法研究 [J]. 交通科学与工程, 2022, 38 (2): 102 - 108.
- HYNDMAN R J, KHANDAKAR Y. Automatic time series forecasting: the forecast package for R [J]. Journal of statistical software, 2008, 27 (3): 1 - 22.
- PAO H T. Forecasting energy consumption in Taiwan using hybrid nonlinear models [J]. Energy, 2009, 34 (10): 1438 - 1446.
- JIA W, WAN Y, LI Y, et al. Integrating multiple data sources and learning models to predict infectious diseases in China [J]. AMIA summits on translational science proceedings, 2019 (1): 680.
- 赵嘉, 苏雪梅, 赵自雄, 等. 新形势下疾病预防控制信息化需求及应用场景分析 [J]. 医学信息学杂志, 2024, 45 (2): 65 - 70.
- 刘天, 姚梦雷, 侯清波, 等. 7 种时间序列模型对全国肾综合征出血热发病率预测效果比较 [J]. 中国媒介生物学及控制杂志, 2022, 33 (4): 548 - 554.
- 刘天, 姚梦雷, 侯清波, 等. SARIMA、GAM 和 LSTM 在肾综合征出血热预测中的应用效果比较 [J]. 中华地方病学杂志, 2022, 41 (9): 709 - 714.
- 陈裕友, 王琴, 胡静匀, 等. 基于非线性组合优化模型的抗癌药物 QSAR 预测 [J]. 数学的实践与认识, 2022, 52 (10): 184 - 190.
- 王荣政, 廖贤艺, 陈湘萍, 等. 基于集成学习融合模型的血糖预测 [J]. 医学信息学杂志, 2019, 40 (1): 59 - 62, 84.