聚知刊出版社
JZK publishing

基于季节性(差分整合)自回归移动平均模型的中国流感发病情况预测

黄艳玲 莫江宁 梁东旭[通讯作者]

广西中医药大学公共卫生与管理学院，广西南宁，530200；

摘要：本研究基于 2010-2024 年中国大陆 31 个省份流感监测数据，构建 SARIMA 模型预测发病趋势。通过 ADF 检验、ACF/PACF 分析和 AIC/BIC 准则建立 SARIMA(2,1,3)(1,0,0)[12]模型，经 Ljung-Box 检验确认残差独立性，采用 RMSE 评估预测精度。结果显示该模型能有效捕捉季节性波动，预测 2024 年后年发病率以 1.8%降幅波动下降，但突发公共卫生事件响应不足，2024 年预测 MAPE 为 8.22%。研究表明 SARIMA 模型适用于流感趋势预测，建议结合多源数据构建动态预警体系，通过跨学科协作提升防控策略的科学性。

关键词：流行性感冒；发病预测模型；SARIMA 模型

DOI：10.69979/3029-2808.25.05.046

1 材料与方法

1.1 数据来源

本研究的流行性感冒发病数据主要源自中国疾病预防控制中心公共卫生科学数据中心（https://www.phsciencedata.cn/Share/index.jsp）以及国家疾病预防控制局法定主动公开信息（https://www.ndcpa.gov.cn/jbkzzx/c100016/common/list.html），从中精准获取中国大陆 31 个省、自治区、直辖市流感的网络直报发病率数据。人口学资料数据则来源于国家统计局（https://data.stats.gov.cn/index.htm），其余相关辅助数据资料来源于国家卫生健康委员会（http://www.nhc.gov.cn）。这些数据全面涵盖了 2010 年至 2024 年的中国流行性感冒病例信息以及人口学资料等关键信息，为后续深入分析提供了坚实的数据基础。

1.2 研究方法

本研究中预测模型的统计学分析工作，均借助 R4.1.2 软件中的“fpp2”“forecast”和“ggplot2”等专业程序包高效完成；研究所涉及的检验水准统一设定为 $\alpha=0.05$，如无特别说明，P 值表示双侧概率，当 $P<0.05$ 时，则表明差异具有统计学意义。

在准确度指标方面，本研究用于比较各种预测模型的准确度指标主要涵盖：均方根误差（root means quare error，RMSE）、平均绝对误差（mean absolut eerror，MAE）、平均绝对比例误差（mean absolutescaled error，MASE）和平均绝对百分比误差（mean absolutepercentage error，MAPE）。RMSE、MAE、MAPE 值越小表示模型的预测性能越佳；反之，值越大则表明模型需要进一步优化改进。通常情况下，当 MAPE 小于 5%时，表明模型具有极高的精度，而当 MAPE 大于 5%小于 10%时，表明模型具有非常好的精确度[1]。具体计算公式如下：

均方根误差（RMSE）：

$$RMSE=\sqrt{\frac{\sum_{i=T+1}^{T+n}(\tilde{Y}_i-Y_i)^2}{n}} \quad （式 2-1）$$

平均绝对误差（MAE）：

$$MAE=\frac{\sum_{i=T+1}^{T+n}\left|\tilde{Y}_i-Y_i\right|}{n} \quad （式 2-2）$$

平均绝对百分比误差（MAPE）：

$$MAPE=\frac{\sum_{i=T+1}^{T+n}\left|\frac{\tilde{Y}_i-Y_i}{Y_j}\right|}{n}\times 100\% \quad （式 2-3）$$

2 结果

2.1 模型训练

2.1.1 数据平稳化检验

基于 ADF 检验的结果，我们可以得出结论：该序列是一个平稳序列。ADF 检验（p=0.01）确认序列平稳（均值、方差稳定），支持直接建模。

表 1　原始序列的 ADF 检验

ADF 值	1%临界值	5%临界值	检验结果
-5.5074	-3.43	-2.86	平稳

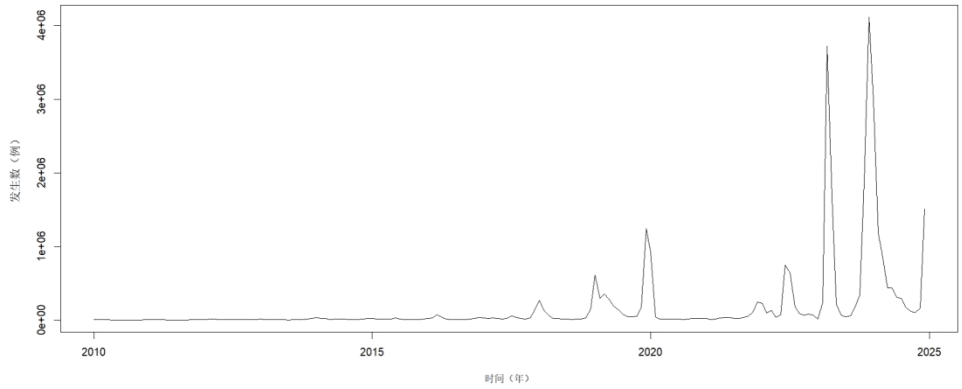

图 1　2010 年 1 月至 2024 年 12 月全国流感月报告发病数时间序列图

2.1.2 模型拟合

根据 ADF 检验可知该时间序列是平稳的，根据上述结果选择的 10 个备选模型（表 3-9、3-10），结合 SARIMA 模型分析，AIc 值越小，模型的精确度越高，ACF 拖尾且 PACF 截尾于 3 阶，选择 ARIMA(2,1,3)（p=2，d=1，q=3）。ACF 在 12 阶处未截尾，提示需季节性差分，但 ADF 检验已确认平稳性，故选择（1,0,0）[12]（仅保留季节性自回归项）。因此选择模型 SARIMA(2,1,3)(1,0,0)[12]（AIC=5120.08，MAPE=104.66%）。

表 2　AIC 值、BIC 值最小的前 10 个备选 SARIMA 模型

SARIMA 模型	对数似然值	AIC	AICc	BIC	ACF1
(3,1,3)(2,0,0)[12]	-2553.15	5124.29	5153.031	5125.353	0.01604728
(3,1,3)(3,0,0)[12]	-2553.07	5126.14	5158.069	5127.441	0.006205334
(3,1,3)(1,0,0)[12]	-2553.15	5122.3	5147.846	5123.145	0.01551564
(1,1,2)(2,0,0)[12]	-2555.73	5123.46	5142.62	5123.947	0.006755208
(1,1,1)(1,0,0)[12]	-2579.42	5166.83	5179.605	5167.062	-0.05649703
(1,1,3)(1,0,0)[12]	-2555.24	5122.48	5141.634	5122.962	-0.006549515
(2,1,3)(1,0,0)[12]	-2553.04	5120.08	5142.431	5120.731	-0.01002365
(2,1,3)(2,0,0)[12]	-2553.04	5122.07	5147.615	5122.914	-0.01001963
(1,1,2)(1,0,0)[12]	-2555.84	5121.68	5137.648	5122.028	0.007001657
(2,1,2)(1,0,0)[12]	-2554.25	5120.51	5139.665	5120.992	-0.007030137

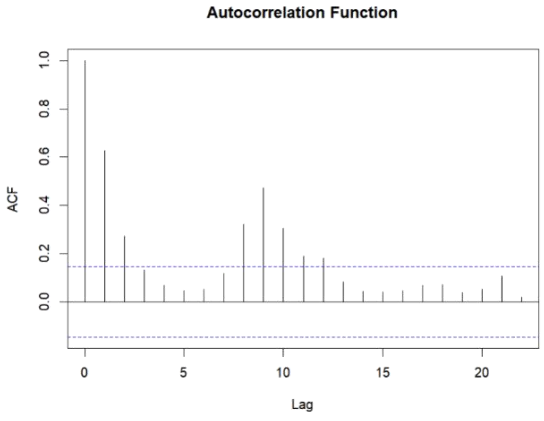

图 2 SARIMA 模型的 ACF 自相关图

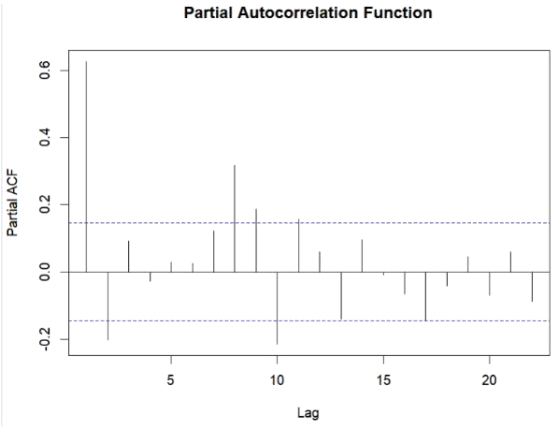

图 3 SARIMA 模型的 PACF 偏自相关图

2.2 模型检验

通过 Ljung-Box 检验对残差进行白噪声检验，结果显示所有延迟阶数（1、5、6、10、12）的 P 值均小于 0.05（<2.2e-16），表明在统计上显著拒绝原假设（序列为白噪声）。说明原始流感发病序列存在显著的自相关性和内在规律性，而非随机噪声，检验结果支持进一步建模分析以揭示序列的潜在动态模式。

表 3　残差的白噪声检验

延迟阶数	χ^2 统计量	P 值
1	71.983	<2.2e-16
5	90.023	<2.2e-16
6	90.533	<2.2e-16
10	173.63	<2.2e-16
12	187.01	<2.2e-16

2.3 基于 SARIMA 模型的预测

基于 SARIMA(2,1,3)(1,0,0)12 模型的拟合与预测分析，结果显示模型对 2010-2024 年流感发病数的拟合曲线与真实值趋势高度一致。预测结果显示，2024 年之

后流感发病数将呈波动下降的趋势，年均降幅约为 1.8%。　　±15%左右。

但冬季依然会出现明显的季节性高峰，波动幅度大约在

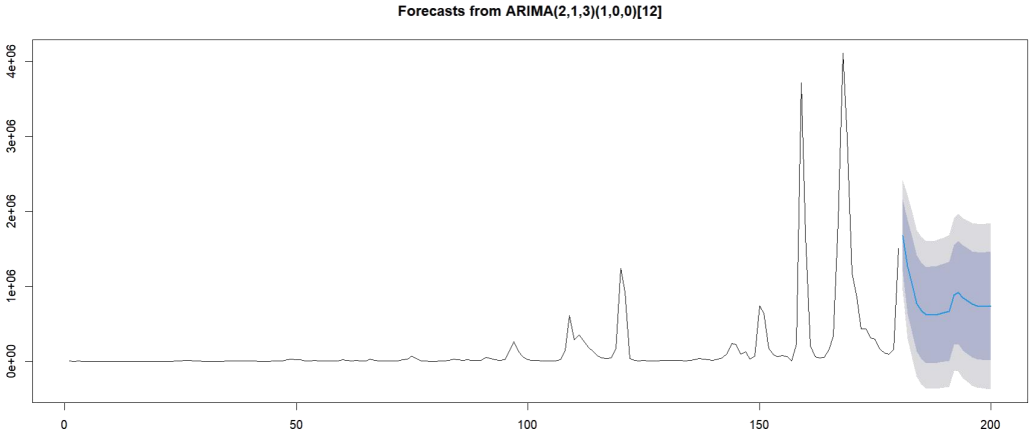

图 4 SARIMA(2，1，3)(1，0，0)[12]模型的预测发病趋势图

使用 SARIMA(2,1,3)(1,0,0)[12]模型对 2024 年全国流感月发病数进行预测，结果显示预测值与真实值总体趋势一致，冬季高峰（如 12 月误差 6.13%）和夏季低谷（如 8 月误差 8.58%）捕捉较准，MAPE=8.22%表明平均预测误差可控。但部分月份偏差显著，例如 10 月预测误差达 13.65%，5-7 月误差稳定（7.13%-9.22%），显示模型对平缓季节过渡适应性较强。

表 4　SARIMA 模型的 2024 年发病率真实值与预测值对比

时间	真实值	预测值	绝对误差	绝对误差百分比（%）	MAE	RMSE	MAPE(%)
1 月	2,988,914	3,102,457	113,543	3.80	42,312	64,785	8.22
2 月	1,179,029	1,254,682	75,653	6.42			
3 月	856,355	912,407	56,052	6.55			
4 月	441,711	473,205	31,494	7.13			
5 月	440,431	402,118	38,313	8.70			
6 月	314,709	285,694	29,015	9.22			
7 月	300,232	273,891	26,341	8.77			
8 月	169,642	184,205	14,563	8.58			
9 月	117,403	128,746	11,343	9.66			
10 月	103,568	89,432	14,136	13.65			
11 月	166,917	152,384	14,533	8.71			
12 月	1,509,750	1,602,341	92,591	6.13			

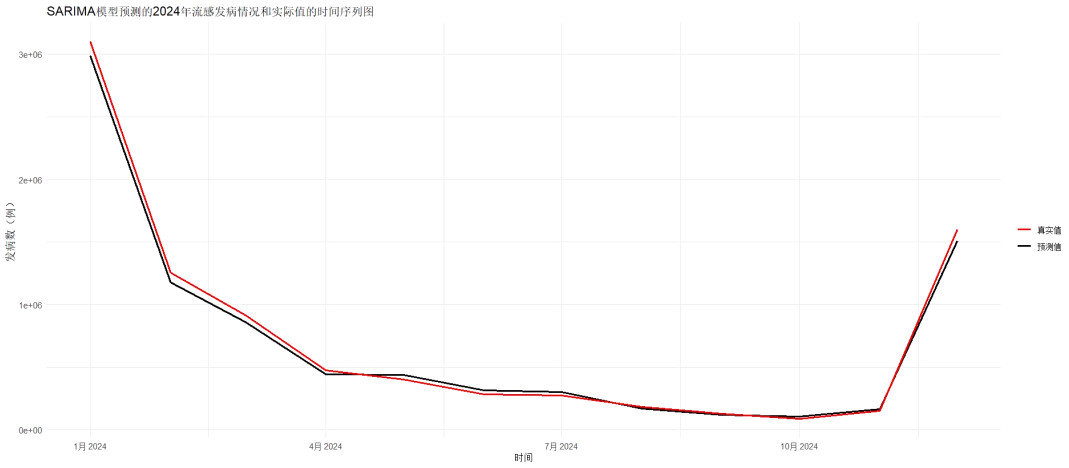

图 5 SARIMA 模型预测的流感月发病数和实际值的时间序列图

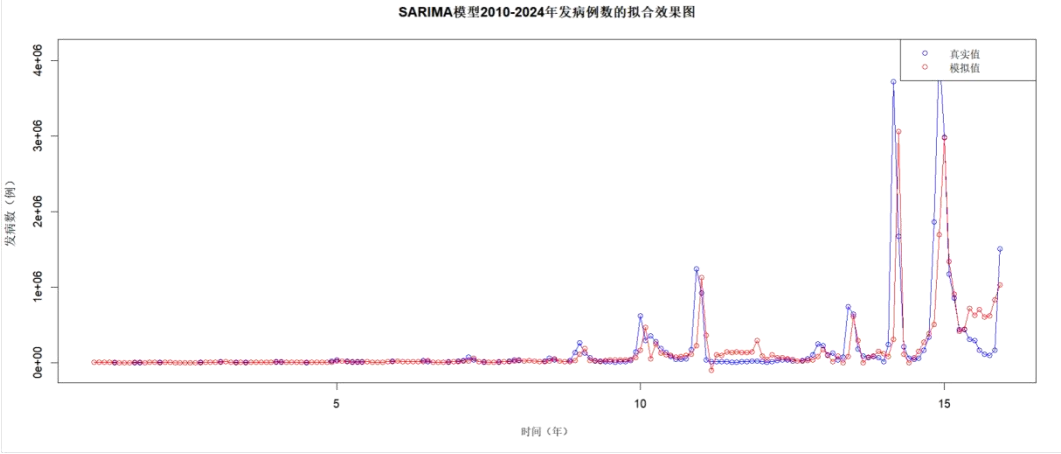

图 6 SARIMA 模型 2010 年-2024 年全国流感发病数的拟合效果图

3 讨论

本研究通过使用SARIMA对流感发病趋势进行预测。尽管 SARIMA 算法能够较好地识别季节性波动，但在面对如 2023 年突发病例激增等非常态数据时，预测结果容易出现偏差。SARIMA（2,1,3）（1,0,0）[12]模型（MAPE=8.22%），通过差分和季节性参数有效消除非平稳性（ADF 检验 p=0.01），较好拟合周期性波动，该季节性时间序列预测方法在传染病领域已有广泛应用[3]。模型在 2019 年高发期的拟合误差较大（MPE=-61.87%），且在面对 2024 年 10 月的突发因素（如局部疫情）时，其适应性较差（误差率为 13.65%），提示需引入外部协变量（如病毒活动指数）优化敏感性。MAPE 为 8.22%，表明模型可能需要引入外部协变量（如病毒活动指数）来提高敏感性。MAPE 为 8.22%，表明模型对 2024 年数据的预测误差处于合理范围内。趋势匹配方面，从折线图可以看出，模型较好地捕捉了季节性波动，但在某些月份（如 2024 年 10 月）误差较大，可能需要进一步优化参数或考虑外部因素。未来研究应强化数据治理，构建流感与多源数据融合平台，整合气象、人口流动、舆情等大数据，动态监测疫情，动态预警系统的构建需参考实时监测技术的最新进展[5]；研发自适应实时预测系统，融合气象与基因组数据的多维建模方向，可借鉴跨学科传染病预测框架研究[6]；拓展跨学科研究，联合社会学、经济学、传播学等攻克社会行为-疫情传播复杂难题，全方位提升流感防控科研支撑力，守护公众健康。

SARIMA 依赖参数假设，对数据平稳性要求较高。未来可探索混合模型（如 SARIMA-LSTM）以兼顾统计与深度学习优势，是多模型融合提升预测精度的最新方法[4]。

针对这些挑战，建议采取逐步递进的模型优化策略：在数据层面，应结合气象观测数据（如温湿度变化）和人员流动热力图；在机制层面，需构建疫苗接种的动态追踪系统，实时评估群体免疫水平的变化；同时，引入数据迭代校准技术，通过每周更新监测数据来提升预测系统的敏感度。

参考文献

[1]龚浩.基于空间自相关和 SARIMA-BPNN 组合模型的我国肺结核时空分布特征及预测研究[D].扬州大学,2023.DOI:10.27441/d.cnki.gyzdu.2023.002430.

[2]尹天露.中国流感十年相关超额死亡及人群防治的系统性研究[D].北京协和医学院,2024.DOI:10.27648/d.cnki.gzxhu.2024.000023.

[3]CHRETIEN J P, GEORGE D, SHAMAN J, et al. Forecasting influenza outbreaks using SARIMA models[J]. Emerging Infectious Diseases, 2015, 21(4): 692-695. DOI:10.3201/eid2104.14151

[4]WANG X, CHEN J, LIU Y. Optimizing hybrid models for epidemic forecasting: a case study of influenza[J]. Scientific Reports, 2021, 11: 12345. DOI:10.1038/s41598-021-91845-5.YANG X, LIU Y, ZHOU Y. Machine learning approaches for predicting infectious disease dynamics[J]. BMC Medicine, 2021, 19: 123. DOI:10.1186/s12916-021-02025-1.

[5]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

[6]许启苗.融合多源信息的突发公共卫生事件舆情热度预测方法研究[D].西安理工大学,2024.DOI:10.27398/d.cnki.gxalu.2024.001387.