

· 论 著 ·

基于 ARIMA 模型的 2010—2020 年云南省 HIV/AIDS 发病率预测分析*

陈雪梅

(红河卫生职业学院, 云南 蒙自 661100)

[摘要] **目的** 构建自回归移动平均 (ARIMA) 模型, 以预测该地区的人类免疫缺陷病毒/获得性免疫缺陷综合征 (HIV/AIDS) 发病率, 并对模型的预测效能进行评估。**方法** 对云南省 2010 年 1 月至 2020 年 12 月 HIV/AIDS 的月发病率数据建立 ARIMA 模型。通过比较分析, 选择 AIDS 和 HIV 发病率的最优拟合模型, 并对 2020 年全年的月发病率进行了预测, 分析模型预测的准确度。**结果** 在 2010—2020 年, 云南省的 AIDS 发病率保持稳定, 而 HIV 发病率则呈逐年下降趋势。经平稳化处理、单位根检验 (ADF) 和模型筛选, 确定 $ARIMA(2, 0, 2)(1, 0, 2)_{[12]}$ 为 AIDS 发病率的最优拟合模型, $ARIMA(2, 1, 1)(2, 1, 0)_{[12]}$ 为 HIV 发病率的最优拟合模型, 模型拟合优度检验显示 R^2 分别为 0.668 和 0.737, Ljung-Box 统计量分别为 12.97 ($P > 0.05$) 和 14.89 ($P > 0.05$), 贝叶斯信息准则 (BIC) 值分别为 -3.07 和 -3.08, 平均绝对百分比误差 (MAPE) 分别为 16.41 和 11.29。模型残差的自相关函数图 (ACF) 和偏自相关函数图 (PACF) 均在 95% CI 范围内, 模型曲线的预测值与实际值的趋势一致, 预测值与实际值接近。**结论** ARIMA 模型对 AIDS 和 HIV 发病率的预测效果良好, 该模型可作为短期预测和分析的有效工具, 为相关部门及时采取有效的 HIV/AIDS 防控措施提供科学的决策支持。

[关键词] 人类免疫缺陷病毒/获得性免疫缺陷综合征; 自回归移动平均模型; 时间序列分析; 疾病预测; 发病率

DOI: 10.3969/j.issn.1009-5519.2025.01.003 中图法分类号: R181.5⁺⁵
文章编号: 1009-5519(2025)01-0011-07 文献标识码: A

Prediction of HIV/AIDS incidence in Yunnan from 2010 to 2020 based on ARIMA model*

CHEN Xuemei

(Honghe Health Vocational College, Mengzi, Yunnan 661100, China)

[Abstract] **Objective** To construct an ARIMA (autoregressive moving average) model to predict the incidence of human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) in this region and evaluate the predictive efficacy of the model. **Methods** The ARIMA model was established for the monthly incidence data of HIV/AIDS in Yunnan province from January 2010 to December 2020. Through comparative analysis, the optimal fitting model of the incidence of AIDS and HIV was selected, and the monthly incidence of the whole year of 2020 was predicted, and the accuracy of the prediction model was analyzed. **Results** From 2010 to 2020, the incidence of AIDS in Yunnan province remained stable, while the incidence of HIV showed a downward trend year by year. After stabilizing treatment, unit root test (ADF) and model screening, $ARIMA(2, 0, 2)(1, 0, 2)_{[12]}$ was determined to be the best fitting model for AIDS incidence, and $ARIMA(2, 1, 1)(2, 1, 0)_{[12]}$ was determined to be the best fitting model for HIV incidence. Goodness-of-fit tests revealed that the R^2 were 0.668 and 0.737, Ljung-Box statistics were 12.97 ($P > 0.05$) and 14.89 ($P > 0.05$), Bayesian Information Criterion (BIC) were -3.07 and -3.08, and mean absolute percentage error (MAPE) were 16.41 and 11.29, respectively. The autocorrelation function graph (ACF) and the partial autocorrelation function graph (PACF) of the model residuals are both within the 95% CI range. The predicted value of the model curve was consistent with the actual value, and the predicted value was close to the actual value. **Conclusion** The ARIMA model has a good effect on the prediction of the incidence of AIDS and HIV. It can be used as an effective tool for short-term predic-

* 基金项目: 云南省教育厅科学研究基金项目 (024J1747)。
作者简介: 陈雪梅 (1992—), 博士研究生在读, 讲师, 主要从事流行病与卫生统计学研究。

tion and analysis, and provide scientific decision-making support for relevant departments to take effective AIDS prevention and control measures in time.

[Key words] HIV/AIDS; Auto-regression integrated moving average model; Time series analysis; Disease prediction; Incidence rate

云南省作为中国获得性免疫缺陷综合征(艾滋病, AIDS)的重点流行地区, AIDS 的发病率虽然一直处于下降趋势, 但在全国范围内仍处于较高水平, 降低云南省 AIDS 的发病率是中国实现“2030 年全球消除艾滋病”及“3 个 95%”目标^[1]的一个重要环节。因此, 对 AIDS 的发病情况及时进行预测, 提前掌握 AIDS 疫情的变化趋势, 可以为采取有效的防控措施提供重要参考依据, 并在 AIDS 防治及遏制疫情扩散方面发挥重要作用。自回归移动平均(ARIMA)模型被广泛应用于一些常见疾病的预警预测, 并取得了良好的预测效果^[2-5]。本研究拟通过在 2010—2020 年云南省人类免疫缺陷病毒(HIV)/AIDS 发病率时间序列的基础上建立 ARIMA 模型, 依据时间序列对云南省 HIV/AIDS 的发病率进行预测, 以期实现对其发病率的有效预测, 及时、有效地开展精准防控, 以达到有效降低 HIV/AIDS 发病的目的。

1 资料与方法

1.1 资料来源 数据来源于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/>)最新公布的云南省 2010—2020 年艾滋病相关数据, 包括 AIDS、HIV 发病率共 11 年即 132 个月的数据, 其中 2010 年 1 月至 2019 年 12 月的数据用于建立模型, 2020 年 1—12 月的数据用于验证模型预测效果。

根据 AIDS 患者所处疾病分期, 将艾滋病的发病情况分为 2 类, 一类为感染者尚处于 HIV 期被诊断发现的, 纳入计算 HIV 发病率; 另一类为感染者已经进入 AIDS 期才被诊断发现的情况, 纳入计算 AIDS 发病率, HIV/AIDS 的发病率构成 2 组时间序列数据。

1.2 方法

1.2.1 序列的平稳化 将 2010—2019 年数据整理成时间序列, 绘制原始序列图, 采用单位根检验(ADF)识别序列平稳性, 对非平稳序列进行差分后判定时间序列是否具有平稳性, 确定 ARIMA 模型的 d 值(非季节性差分阶数)和 D 值(季节性差分阶数)。

1.2.2 模型定阶 通过自相关分析分别得出自相关函数图(ACF)和偏自相关函数图(PACF), 由 PACF 图确定自回归阶数 p 值, 由 ACF 图确定移动平均阶数 q 值。

1.2.3 ARIMA 模型构建和选择 采用 ARIMA(p, d, q)(P, D, Q)模型进行模型构建和参数估计, 根据

R^2 , 贝叶斯信息准则(BIC)、Ljung-Box 值及其 P 值和平均绝对百分比误差(MAPE)进行模型诊断, 选择最优模型。

1.2.4 模型拟合及预测 应用选择的最优模型对 2020 年 1—12 月云南省 AIDS 和 HIV 的发病数据进行验证分析, 判定模型的适用性并进行预测。

1.3 统计学处理 使用 Excel2023 对 2010—2020 年 AIDS 和 HIV 的发病率进行整理, 使用 SPSS26.0 软件, 对时间序列进行数据序列的平稳化、自相关检验及 ARIMA 模型的建立、筛选和筛选最优模型进行预测及数据整理。检验水准 $\alpha=0.05$ 。

2 结果

2.1 云南省 HIV/AIDS 的发病率和死亡率基本情况 经时间序列图分析显示, 2010—2019 年 AIDS 发病率呈一定上升趋势, 于 2012 年有明显上升, 随后呈较平稳的波动, 具有一定周期性; HIV 发病率呈现一定下降趋势, 但不同月份的 HIV 发病率有明显的波动(图 1)。

2.2 序列平稳化 考虑时间序列存在非平稳序列的问题, 无法满足时间序列 ARIMA 模型建模的条件, 对 2 组数据分别进行平稳性检验。经 ADF 检验得出, AIDS 发病率为非平稳序列, $t=-2.95, P=0.315$, 不拒绝原假设, 序列不平稳。经一阶差分后 ADF 检验 $t=-6.35, P<0.01$, 序列平稳, 初步确定 AIDS 发病率的参数 $d=1$ 且 $D=1$ 。HIV 发病率为平稳序列, 经一阶差分后数据平稳状态更明显, 可初步确定参数 d 和 D 值可为 0 或 1($P<0.01$)。见表 1。故一阶差分后各组时间序列经 ADF 检验为平稳序列, 见图 2, 可进行 ARIMA 模型构建。

2.3 ARIMA 模型构建和选择

2.3.1 确定模型参数 对 HIV/AIDS 的发病率进行一阶差分自相关分析, 得出 2 组时间序列的 ACF 和 PACF(图 3)。结果显示, AIDS 发病率 ACF 滞后二阶处截尾, PACF 在滞后二阶处截尾, 初步判定 $p=2, q=2$ 。HIV 发病率的 ACF 在滞后一阶截尾, PACF 拖尾, 初步判定 $p=0, q=1$ 。

2.3.2 模型构建与选择 纳入序列平稳化及自相关分析的 p, d, q 值, 构建 ARIMA(p, d, q)(P, D, Q)模型进行拟合, 根据 R^2 、Ljung-Box 统计量、BIC 值、MAPE 和 P 值选择最优函数, 确定最终模型(表 2)。得出 AIDS 发病率的最优模型 ARIMA(2, 0, 2)(1, 0, 2)_[12], HIV 发病率的最优模型为 ARIMA(2, 1, 1)(2,

1,0)_[12],结果见表 2。Ljung-Box 统计量的 P 值均大于 0.05,模型残差为白噪声,模型提取了时间序列的

大部分信息,模型效果良好。

表 1 AIDS 和 HIV 发病率时间序列的 ADF 检验结果

时间序列	差分阶数	t	P	临界值		
				1%	5%	10%
AIDS 发病率	0	-2.95	0.15	-4.04	-3.45	-3.15
	1	-12.66	<0.000 01	-4.04	-3.45	-3.15
HIV 发病率	0	-9.83	<0.000 01	-4.04	-3.45	-3.15
	1	-8.21	<0.000 01	-4.04	-3.45	-3.15

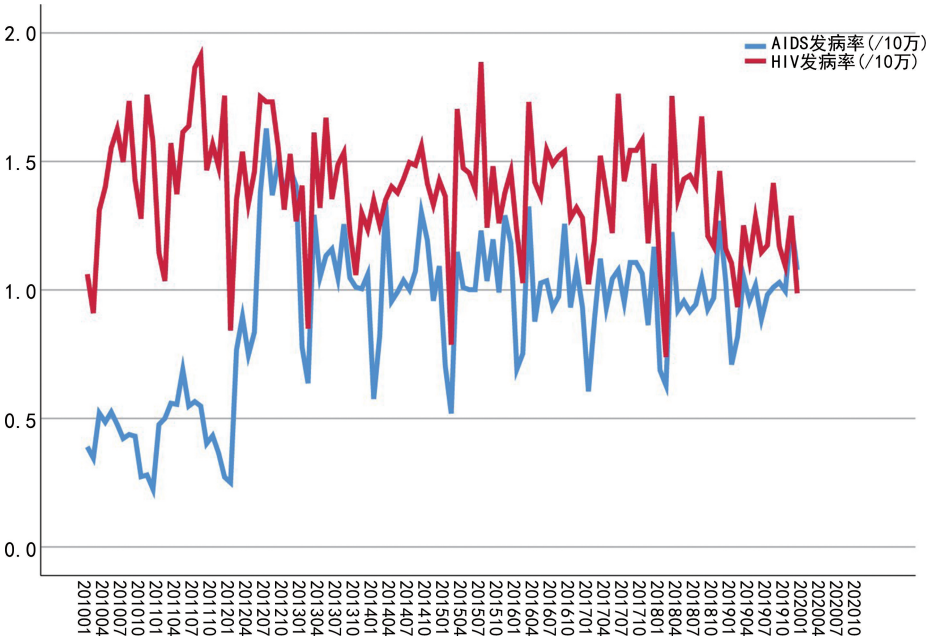


图 1 云南省 HIV/AIDS 发病率时间序列

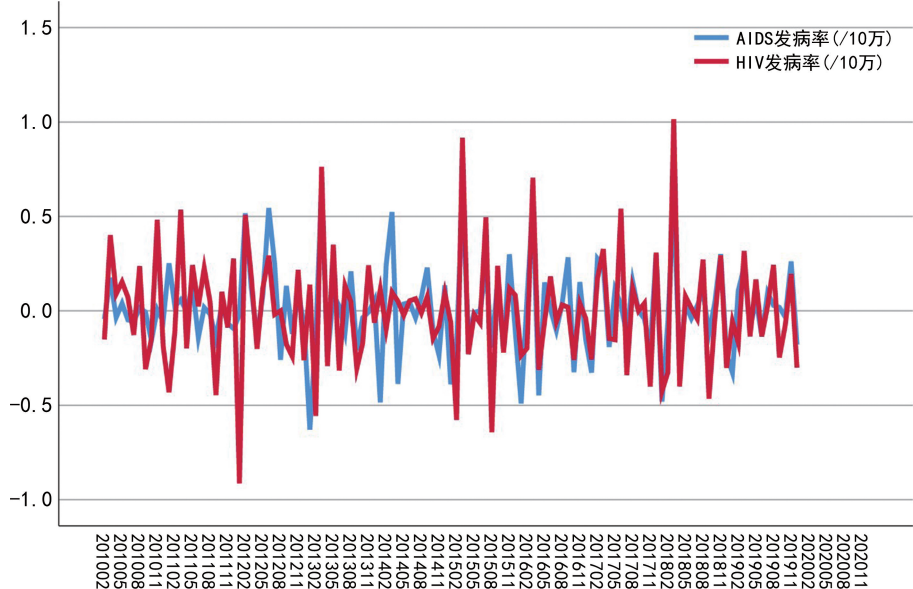


图 2 AIDS 和 HIV 发病率一阶差分后时间序列图

2.3.3 模型验证 对筛选的最优模型的各个参数进行检验,结果 $P<0.05$,模型参数差异有统计学意义(表 3)。对最优模型的残差序列进行正态性检验,得

出 2 组序列检验 $P>0.05$,残差分布均呈正态分布。残差的 ACF 和 PACF 分析结果显示,残差相关系数和偏相关系数均在 95%CI 范围内(图 4),2 组时间序

列的拟合效果均良好。经诊断,所选模型合适,可用于模型预测。

2.4 模型拟合及预测 使用 $ARIMA(2,0,2)(1,0,2)_{[12]}$ 和 $ARIMA(2,1,1)(2,1,0)_{[12]}$ 模型分别对 2020 年云南省的 AIDS 及 HIV 的发病率进行拟合预测,模型拟合预测图显示预测曲线与实际曲线的分布情况基本保持一致(图 5),模型预测值与实际值存在一定

偏差,但均在拟合值 95%CI 内,整体模型拟合效果良好。与 2020 年 1—12 月报告的发病率相比较,AIDS 报告的发病率与预测值的相对误差在 $-20.16\% \sim 26.47\%$,平均相对误差为 -2.54% ;HIV 报告的发病率与预测值的相对误差在 $-39.17\% \sim 13.16\%$,平均相对误差为 -18.23% 。2020 年云南省 HIV/AIDS 报告发病率的预测值及 95%CI 见表 4。

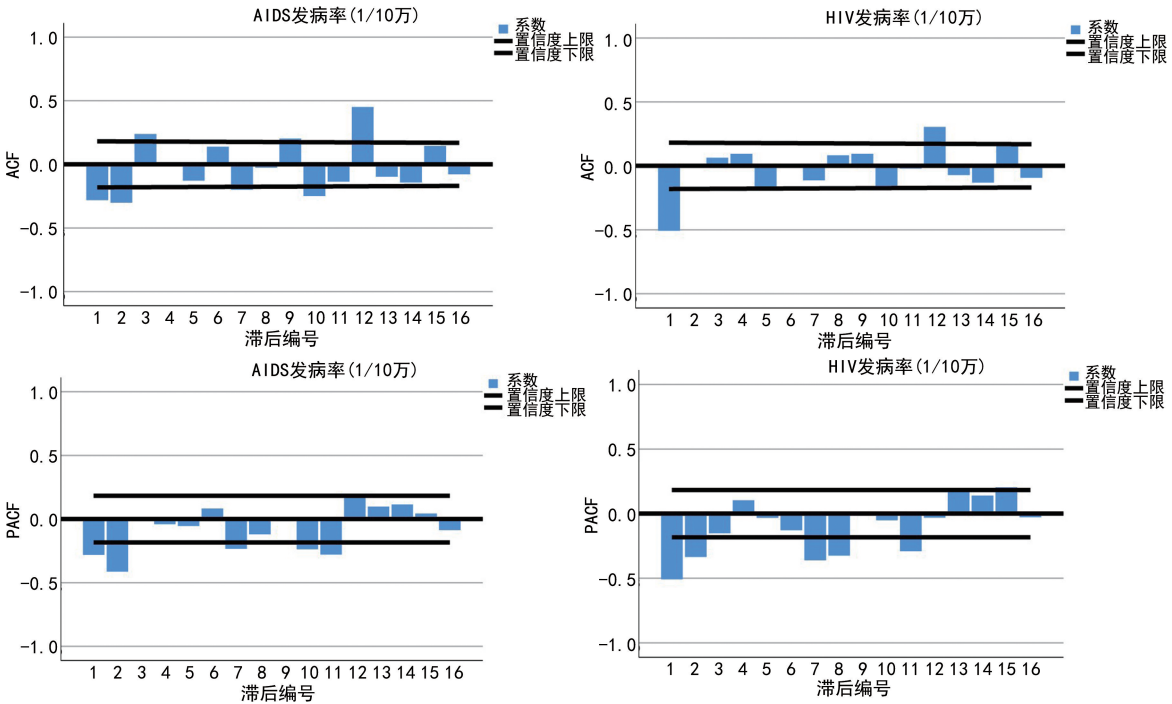


图 3 AIDS 和 HIV 发病率一阶差分后 ACF 和 PACF 图

表 2 ARIMA 模型筛选及拟合优度检验

时间序列	ARIMA 拟合模型	R^2	MAPE	BIC	Ljung-Box 统计量	
					统计量	P
AIDS 发病率	$ARIMA(2,0,1)(1,0,2)_{[12]}$	0.666	16.38	-3.11	14.36	0.278
	$ARIMA(2,0,2)(1,0,2)_{[12]}$	0.668	16.41	-3.07	12.97	0.296
	$ARIMA(2,1,2)(1,0,2)_{[12]}$	0.447	16.02	-3.13	11.27	0.421
HIV 发病率	$ARIMA(1,1,1)(2,1,0)_{[12]}$	0.737	11.34	-3.13	14.91	0.384
	$ARIMA(2,1,1)(2,1,0)_{[12]}$	0.737	11.29	-3.08	14.89	0.315
	$ARIMA(0,1,1)(2,1,0)_{[12]}$	0.707	11.61	-3.07	32.97	0.005

表 3 最优拟合 ARIMA 模型参数检验结果

模型	参数*	估算值	标准误差	t	P
AIDS 发病率 $ARIMA(2,0,2)(1,0,2)_{[12]}$	常量	-0.002	0.002	-1.070	0.287
	AR 延迟 2	0.818	0.147	5.577	0.000
	MA 延迟 1	-0.516	0.187	-2.753	0.007
	MA 延迟 2	0.397	0.106	3.730	0.000
	AR 季节性延迟 1	0.879	0.135	6.536	0.000
	MA 季节性延迟 2	0.639	0.190	3.365	0.001

续表 3 最优拟合 ARIMA 模型参数检验结果

模型	参数*	估算值	标准误差	t	P
HIV 发病率 ARIMA(2,1,1)(2,1,0) _[12]	常量	−0.002	0.002	−1.070	0.287
	AR 延迟 1	−0.376	0.133	−2.828	0.006
	MA 延迟 2	0.782	0.093	8.437	0.000
	AR 季节性延迟 1	−0.622	0.102	−6.100	0.000
	AR 季节性延迟 2	−0.323	0.109	−2.975	0.004

注：* 表示因篇幅较大，表格仅表示出具有统计学意义的参数，无统计学意义指标未展示。

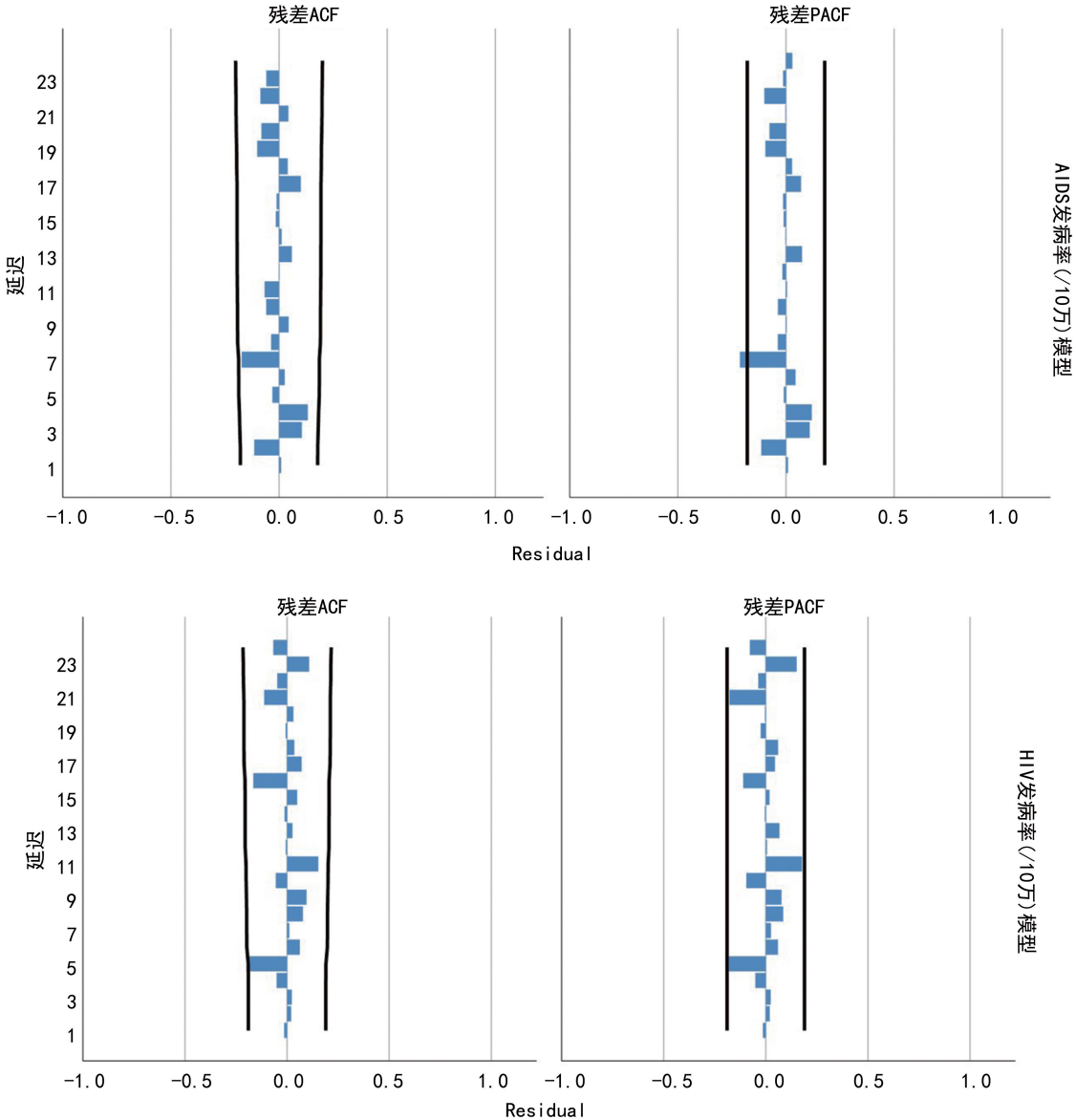


图 4 AIDS 和 HIV 发病率 ARIMA 模型残差序列的 ACF 和 PACF 图

表 4 2020 年 1—12 月 HIV/AIDS 发病率实际值与预测值比较

AIDS 发病率 (/10 万)	预测值 (95%CI)	相对误差 (%)	HIV 发病率 (/10 万)	预测值 (95%CI)	相对误差 (%)
0.73	0.80(0.45,1.15)	9.59	0.92	0.87(0.51,1.24)	−5.43
0.68	0.86(0.46,1.26)	26.47	1.20	0.73(0.36,1.10)	−39.17
1.26	1.13(0.70,1.56)	−10.32	1.88	1.26(0.89,1.64)	−32.98
1.03	0.98(0.52,1.44)	−4.85	1.32	1.04(0.66,1.42)	−21.21
0.97	1.01(0.54,1.49)	4.12	1.13	1.07(0.69,1.46)	−5.31

续表 4 2020 年 1—12 月 HIV/AIDS 发病率实际值与预测值比较

AIDS 发病率 (/10 万)	预测值 (95%CI)	相对误差 (%)	HIV 发病率 (/10 万)	预测值 (95%CI)	相对误差 (%)
1.11	0.99(0.49,1.48)	−10.81	1.32	1.20(0.81,1.59)	−9.09
1.07	0.98(0.48,1.49)	−8.41	1.50	1.08(0.69,1.48)	−28.00
0.96	1.04(0.52,1.56)	8.33	1.14	1.29(0.89,1.69)	13.16
1.29	1.03(0.50,1.56)	−20.16	1.47	1.06(0.65,1.46)	−27.89
1.07	1.00(0.46,1.53)	−6.54	1.12	1.02(0.62,1.43)	−8.93
1.20	1.11(0.56,1.65)	−7.50	1.35	1.05(0.64,1.46)	−22.22
1.16	1.04(0.49,1.59)	−10.34	1.39	0.95(0.53,1.36)	−31.65

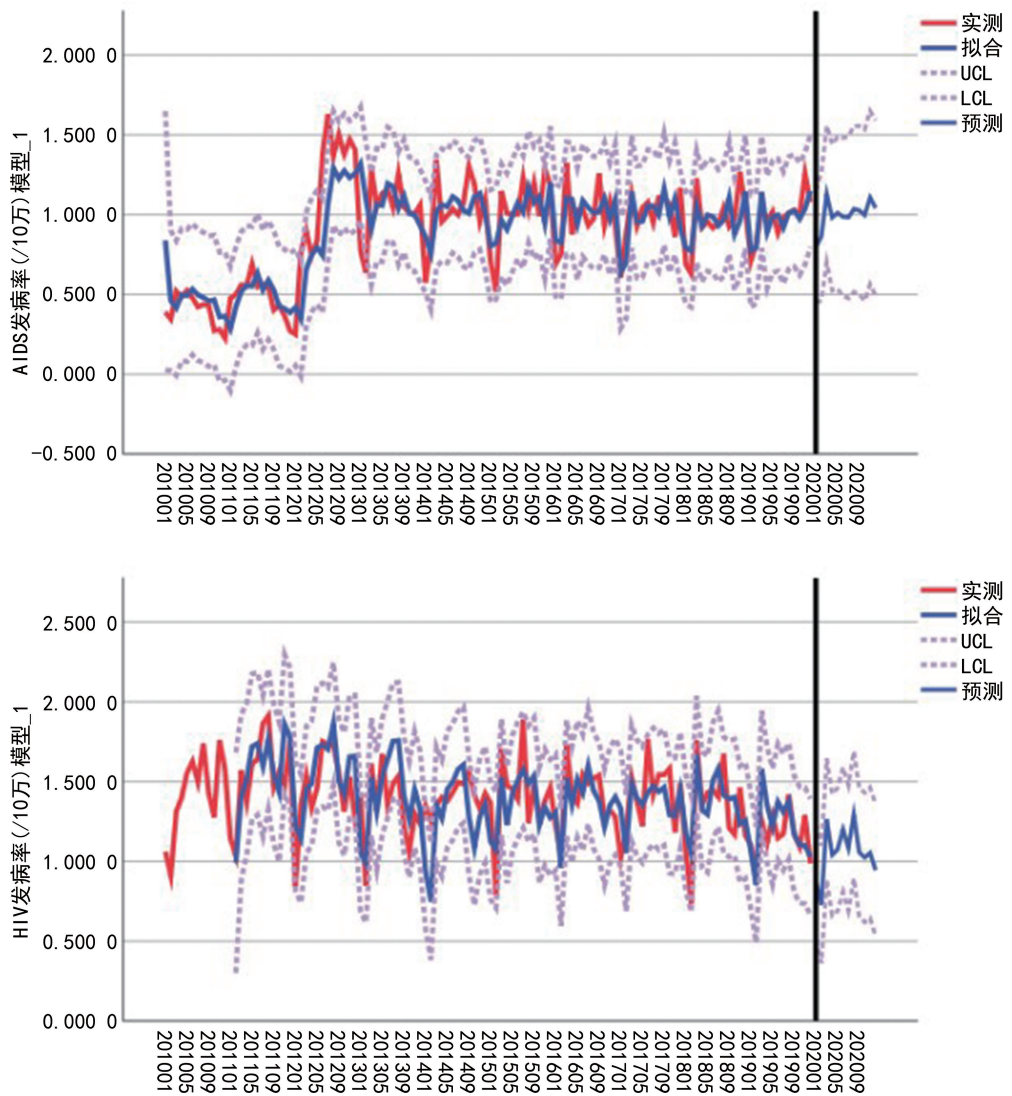


图 5 AIDS 和 HIV 发病率 ARIMA 模型拟合及预测图

3 讨 论

应用数学模型对疾病的发生、发展、死亡的时间序列进行流行状况及发展趋势分析,做出科学准确的预测,揭示疾病发生、发展规律和预警,能帮助卫生部门全面有效地掌控疾病发展趋势,快速做出科学正确的疾病防控决策提供依据。近年来,对疾病预测预警模型的研究众多,常见的疾病预测模型有 ARIMA 模型、神经网络模型(GRNN、LSTM 模型等)、传播动力

学(SEIR)模型,甚至多种混合模型^[6-9]。本研究使用的 ARIMA 模型能够将疾病传播中多种影响因素考虑在内并转换成时间变量进行拟合,克服了随机干扰的问题,保证预测效果且简便有效,被广泛应用于时间序列的预测分析。目前 ARIMA 模型在急性传染病^[7]、慢性病^[10]、地方病^[11]等发病率预测的领域中均有应用,包括在新型冠状病毒感染疫情^[8]和猴痘疫情^[12]的拟合和预测中均取得显著的效果。

研究结果显示,依据 R^2 、Ljung-Box 统计量、BIC 值、MAPE 和 P 值等模型诊断指标 2010—2020 年云南省 AIDS 和 HIV 发病率的最优模型分别为 $ARIMA(2, 0, 2)(1, 0, 2)_{[12]}$ 和 $ARIMA(2, 1, 1)(2, 1, 0)_{[12]}$, $P > 0.05$, 大部分发病率相关信息已被模型提取, 模型拟合效果良好。AIDS 发病率预测值的相对误差在 $-20.16\% \sim 26.47\%$, 平均相对误差仅为 -2.54% , 曲线的变化趋势与实际值基本一致, 可见模型的预测精度较好, 进一步证实了该 $ARIMA(2, 0, 2)(1, 0, 2)_{[12]}$ 可用于云南省 AIDS 发病率的预测。HIV 发病的预测值相对误差在 $-39.17\% \sim 13.16\%$, 平均相对误差为 -18.23% , 曲线变化趋势与实际值基本一致, $ARIMA(2, 1, 1)(2, 1, 0)_{[12]}$ 的拟合效果整体较好, 但相比 AIDS 发病率的拟合模型, 平均相对误差有所增加, 拟合效果有所下降。经 HIV 发病率 $ARIMA(2, 1, 1)(2, 1, 0)_{[12]}$ 拟合的预测值与实际值比较发现, 预测值低于实际值, 这可能与 2020 年新型冠状病毒感染疫情防控期间实施社会隔离的过程中, 减少了高危人群 HIV 干预服务的机会, 增加了 HIV 感染的风险^[13] 有关, 使预测模型出现波动。

AIDS 发病率是 HIV 感染者在进入 AIDS 相关疾病出现的时期才被发现存在 HIV 感染的情况, 是晚发现病例中的一种^[14]。AIDS 的晚发现是降低 AIDS 抗病毒治疗效果和增加 AIDS 死亡率的重要原因^[15]。模型拟合预测图显示, 云南省 AIDS 的发病率基本保持平稳状态, 而 HIV 的发病率处于下降趋势, 说明云南省 AIDS 的晚发现情况变化不明显, 在一定程度上阻碍“3 个 95%”的实现, 降低 AIDS 发病率和晚发现比例仍是云南省 AIDS 防控的一个重要问题, 应采取积极有效的措施降低 AIDS 的发病率, 降低晚发现的比例。

综上所述, ARIMA 模型在 HIV 发病率和 AIDS 发病率的预测中均具有良好效果, 可作为一个有效的工具, 对云南省 HIV/AIDS 发病率时间序列进行短期地预测和动态分析, 以此发现 AIDS 流行过程中存在的挑战和问题, 及时采取有效的防控措施。需要注意的是, 在 ARIMA 模型的拟合过程中, 要做好时间序列的平稳性检验和差分, 充分提取模型的有效信息进行数据的预测, 全面客观地解读预测结果, 并结合实际情况, 将有助于提出更科学、合理、有效的防控措施。

参考文献

[1] 中华人民共和国国家卫生健康委员会. 消除艾滋病、梅毒和乙肝母婴传播行动计划(2022—2025 年)[J]. 中国病毒病杂志, 2023, 13(2): 107.

[2] 苗子阳, 冯亚琴, 翟光富, 等. ARIMA 乘积季节模型在食源性疾病月发病率预测中的应用[J]. 现代预防医学, 2021, 48(23): 4225-4228.

[3] 白桦. 新冠肺炎疫情预测模型的比较研究[D]. 郑州: 河南财经政法大学, 2021.

[4] 陈亮, 连巧龄, 刘美增, 等. ARIMA 模型在预测福建省新报告 HIV/AIDS 病例中的适用性[J]. 中国艾滋病性病, 2021, 27(12): 1353-1357.

[5] 王仕鸿, 令垚, 杨子华, 等. 基于时间序列模型的中国 2020—2029 年慢性肾病发病和患病情况预测研究[J]. 中国慢性病预防与控制, 2023, 31(11): 801-806.

[6] WEI W D, WANG G, TAO X, et al. Time series prediction for the epidemic trends of monkeypox using the ARIMA, exponential smoothing, GM(1, 1) and LSTM deep learning methods[J]. J Gen Virol, 2023, 104(4): 001839.

[7] GAO J, ZHOU C, LIANG H W, et al. Monkeypox outbreaks in the context of the COVID-19 pandemic: network and clustering analyses of global risks and modified SEIR prediction of epidemic trends [J]. Front Public Health, 2023, 11: 1052946.

[8] SAH S, SURENDIRAN B, DHANALAKSHMI R, et al. Forecasting COVID-19 pandemic using prophet, ARIMA, and hybrid stacked LSTM-GRU models in India[J]. Comput Math Methods Med, 2022: 1556025.

[9] WEI W D, JIANG J J, GAO L, et al. A new hybrid model using an autoregressive integrated moving average and a generalized regression neural network for the incidence of tuberculosis in Heng county, China[J]. Am J Trop Med Hyg, 2017, 97(3): 799-805.

[10] 李子悦, 方珈文, 林凯程. 1990—2019 年中国归因于高体质指数的 2 型糖尿病疾病负担分析与预测研究[J]. 中国全科医学, 2024, 27(9): 1126-1133.

[11] 王明君, 张金梅, 姜泓, 等. 定量预测模型在地方病防治中的应用及展望[J]. 中国地方病防治, 2023, 38(6): 461-464.

[12] QURESHI M, KHAN S, BANTAN R, et al. Modeling and forecasting monkeypox cases using stochastic models [J]. J Clin Med, 2022, 11(21): 6555.

[13] 李怡君. 新冠肺炎疫情下山东省 MSM 的 HIV 感染状况及高危行为研究[D]. 济南: 山东大学, 2022.

[14] MAHMOUD M, BALLOUZ T, LAHOUD C, et al. Late presentations and missed opportunities among newly diagnosed HIV patients presenting to a specialty clinic in Lebanon[J]. Sci Rep, 2024, 14(1): 8296.

[15] 赵塔娜, 曾茵茹, 李文瑞, 等. 我国 HIV 感染者/AIDS 患者晚发现发生率及影响因素的 meta 分析[J]. 职业与健康, 2023, 39(12): 1682-1685.

(收稿日期: 2024-05-21 修回日期: 2024-11-03)