2018–2023 年柳州市流感样病例发病时间序列分析及趋势预测

杨　俊

广西柳州市疾病预防控制中心　广西柳州　545007

摘　要：目的 建立时间序列分析 ARIMA（autoregressive integrated moving average）模型，对柳州市流感样病例（influenza-illness,ILI）的发病趋势进行预测，以期为柳州地区流感防控工作提供科学数据支撑。方法 获取柳州市 2018 年第 1 周 -2022 年第 52 周的 ILI 发病资料，构建 ARIMA 模型，对柳州市流感发病情况进行预测和评价。结果 基于 2018-2022 年的 ILI 月发病数，建立季节性 ARIMA 模型，预测 2023 年的 ILI 发病数，模型表现较为良好。结论 柳州市流感样病例的发病趋势短期预测可以运用季节性 ARIMA 模型，而长期预测须不断加入新的数据以调整模型参数。

关键词：柳州市；流感样病例；时间序列分析；趋势预测

流行性感冒，简称为流感，是由流感病毒主要通过呼吸道传播，并由其引起的急性传染病，该病具有如下特征：传染性强、潜伏期短、传播速度快、人群普遍易感等。由于流感病毒存在自身的变异性，如果出现新的亚型，极易引发流感暴发或流行，最终导致人群发病率和死亡率的增涨。全球近 100 年共发生了多次流感大流行，巨大的健康损害和经济负担给人类造成。虽然流感监测网络我国在 20 世纪初就在全国各地建立了，但基于 ILI 监测数据建立的预测预警模型，及时甄别流感暴发疫情，是流感防控的关键技术手段之一。近年常用于预测流感流行趋势的技术方法之一是时间序列分析的 ARIMA 模型 [1]。Jenkins 和 Box 二者在 20 世纪中期提出 ARIMA 模型，目前已普遍应用于传染病的预测预警研究工作中。有研究提示，本地区流感具有明显规律的季节性流行，建立季节性 ARIMA 模型可以预测评估当地的 ILI 发病数。所以本研究拟建立了季节性时间序列模型，同时短期预测柳州市每月 ILI 发病数，为当地流感的防控工作提供科学依据。

1. 资料与方法

1.1 资料来源

由柳州市疾病预防控制中心提供 2018 年第 1 周至 2022 年第 52 周的 ILI 监测数据资料，数据资料来源于流感监测网络的柳州市哨点医院。其中发热（体温≥ 38℃），且伴咳嗽或咽痛之一者为 ILI。人口学数据来源于柳州市统计局。

1.2 统计学分析

本研究对柳州市周 ILI 数使用 Excel2016 软件进行整理，应用 SPSSPRO 软件进行 ARIMA 模型预测及评估最终模型。

2. 结果

2.1 柳州市人口学信息及 ILI 的流行趋势

2018 年柳州市全市人口数 3904713，报告 ILI 数为 57690；2019 年柳州市全市人口数 3935184，报告 ILI 数为 42083；2020 年柳州市全市人口数 3948880，报告 ILI 数为 20118；2021 年柳州市全市人口数 3967935，报告 ILI 数为 23621；2022 年柳州市全市人口数 3984745，报告 ILI 数为 36207。依据 2018–2022 年柳州市流感监测哨点每周的 ILI 发病数，绘制出时序图，ILI 呈现不同年份的流行高峰时间不同，2018–2019 年的流行高峰均在冬季，2020–2021 年流行高峰不明显，2022 年 ILI 高发于夏季和冬季。见图 1。

2.2 应用 SPSSPRO 软件进行预测及最终模型的评估结果

2.2.1 序列分解图

下图 2 展示了从 SPSSPRO 软件里面利用原始数据分解出来的季节性数据，初步判断序列存在季节性效应。

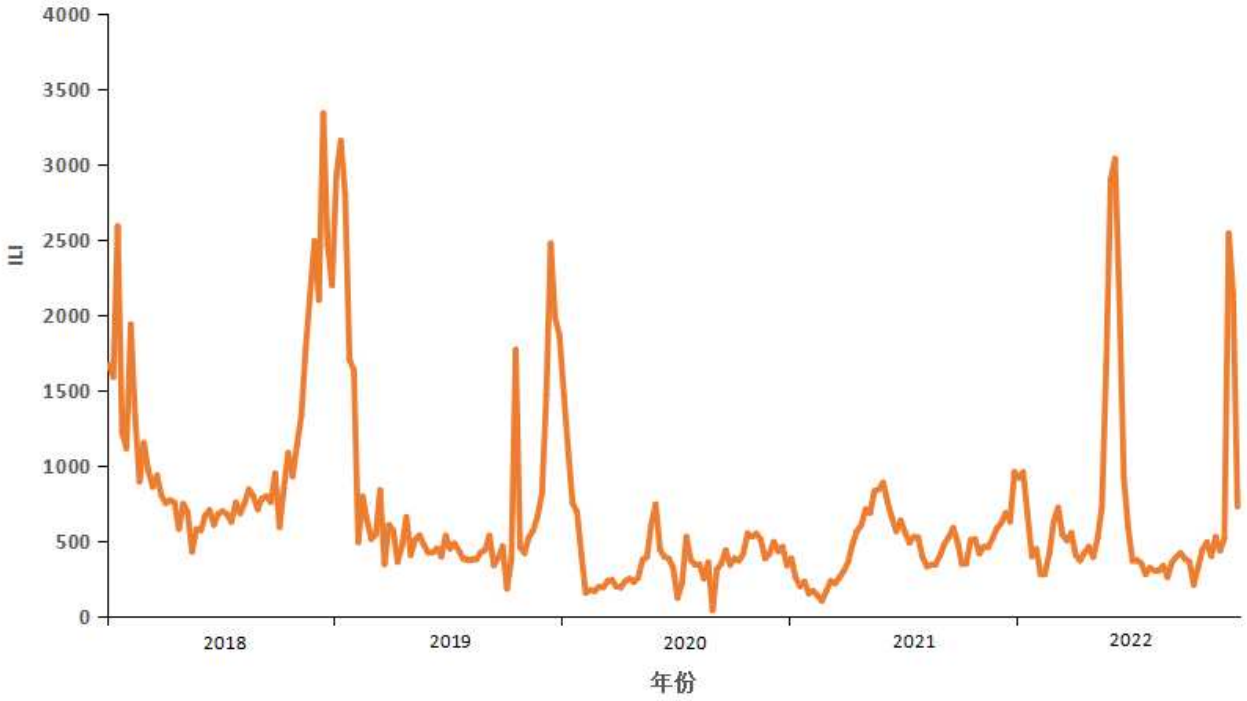

图 1 2018–2022 年柳州市流感样病例时序图

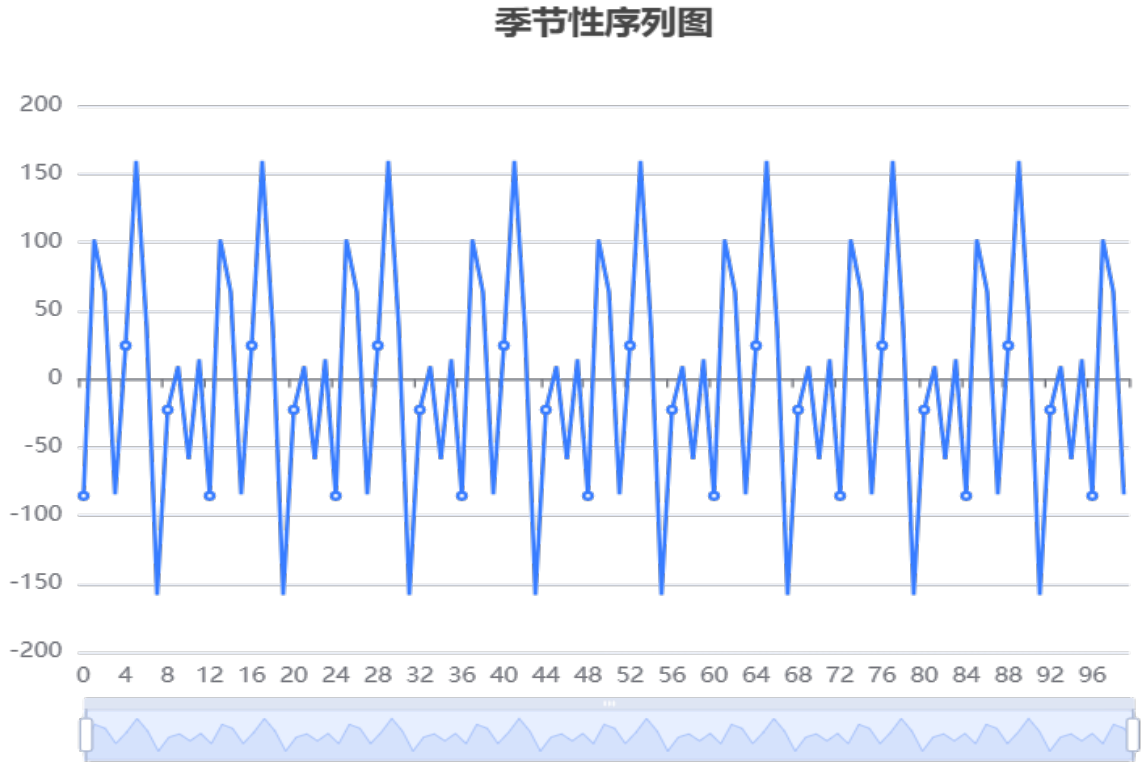

图 2 季节性序列图

2.2.2 ADF 检验

该序列检验的结果显示，该序列为平稳时间序列，原因是变量 ILI 的 1 阶差分序列，统计学显著性 P 值 <0.05，水平上存在显著性，原假设被拒绝。

表 1 ADF 检验

变量	序列	t	P	AIC	临界值		
					1%	5%	10%
ILI	原序列	−2.336	0.161	1267.796	−3.499	−2.892	−2.583
	1 阶差分	−13.627	0.000***	1255.98	−3.499	−2.892	−2.583
	1 阶差分 −1 阶季节差分	−6.851	0.000***	1132.627	−3.51	−2.896	−2.585
	2 阶差分	−6.769	0.000***	1257.88	−3.504	−2.894	−2.584
	2 阶差分 −1 阶季节差分	−10.194	0.000***	1130.534	−3.512	−2.897	−2.586

注：***、**、* 分别代表 1%、5%、10% 的显著性水平

2.2.3 最佳差分序列图

下图 3 展示了原始数据 1 阶差分后的时序图。

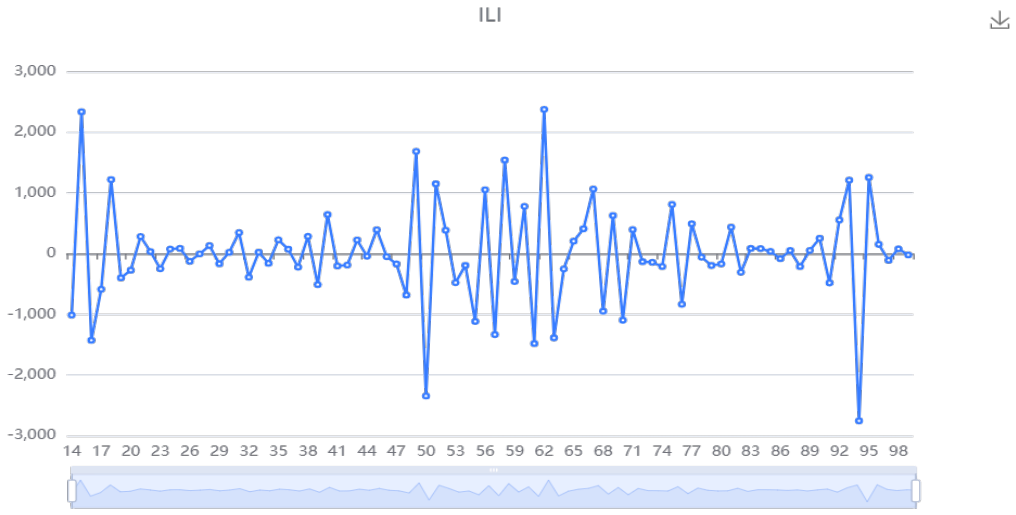

图 3 ILI 最佳差分序列图

2.2.4 最终差分数据自相关图（ACF）和最终差分数据偏自相关图（PACF）

下图 4 与图 5 分别展示了自相关图 (ACF) 和偏自相关图 (PACF)，包括系数，置信上限和置信下限。

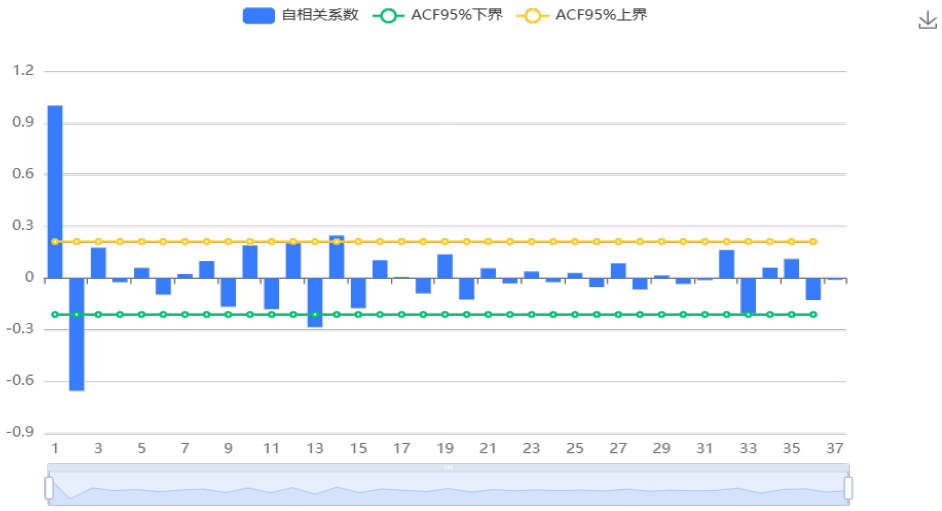

图 4 最终差分数据自相关图（ACF）

图 5 最终差分数据偏自相关图（PACF）

2.2.5 模型评价结果

最优的参数被系统自动寻找出来，模型结果显示是 SARIMAX(4,0,1)×(0,0,0,12)。基于变量 ILI，残差 Q 统计量结果分析从中可以获得：模型表现较为良好。下表格展示本次模型检验结果。根据信息准则 AIC 和 BIC 值用于多次分析模型对比（越低越好）。R^2 代表时间序列的拟合程度，越接近 1 效果越好。见表 2。

表 2 模型评价表

SARIMA 模型 (4, 0, 1)x(0, 0, 0, 12)		
项	符号	值
样本数量	N	100
Q 统计量	Q6（p 值）	0.913
	Q12（p 值）	0.952
	Q18（p 值）	0.992
	Q24（p 值）	1
	Q30（p 值）	1
信息准则	AIC	1468.914
	BIC	1487.15
拟合优度	R^2	0.728

2.2.6 模型参数表

下表格展示本次模型参数结果，包括模型的系数、标准差，t 统计量结果等，用于分析模型公式。见表 3。

表 3 模型参数表

项	系数	标准误	z	P	95% 置信下限	95% 置信上限
intercept	54.79	35.122	1.56	0.119	−14.048	123.628
ar.L1	1.257	0.209	6.027	0.000***	0.848	1.666
ar.L2	−0.221	0.148	−1.487	0.137	−0.511	0.07
ar.L3	0.255	0.175	1.456	0.145	−0.088	0.597
ar.L4	−0.349	0.09	−3.874	0.000***	−0.525	−0.172
ma.L1	−0.707	0.241	−2.929	0.003***	−1.18	−0.234
sigma2	119437.229	13650.813	8.749	0.000***	92682.127	146192.331

注：***、**、* 分别代表 1%、5%、10% 的显著性水平

2.2.7 时间序列图

下图 6 表示了该时间序列模型的真实值、拟合值、预测值。

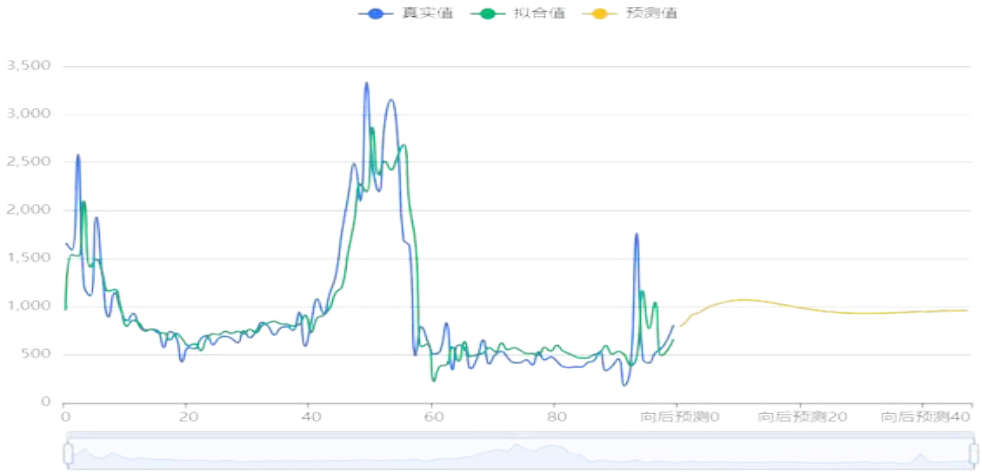

图 6 时间序列图

2.2.8 时间序列预测表

下表 4 显示了时间序列模型最近 48 期数据预测情况。人工对比预测值与真实值相对接近。

表 4 时间序列预测表

阶数（时间）	预测结果
1	790.9191779938403
2	840.8571336533482
3	914.0945896967519
4	936.5527763201612
5	969.0469681331774
6	1006.185672281706
7	1025.9004209654593
8	1042.9403807623107
9	1058.1443657728375
10	1065.5749616548055
11	1069.0300571830142
12	1069.6662989516724
13	1066.2962415046788
14	1060.2084640416629
15	1052.2555220802767
16	1042.5195967025206
17	1031.6579673287779
18	1020.2467550275119
19	1008.5890655304877
20	997.077610120743
21	986.056838579116
22	975.7495439827046
23	966.3541915588221
24	958.0219694733571
25	950.8358422090835
26	944.8396403113613
27	940.0394834184167
28	936.4018204745585
29	933.8653596057075
30	932.3467213282518
31	931.7439199529376
32	931.9432142017156
33	932.824233078135
34	934.2638252328163
35	936.1402986646056
36	938.3368191474879
37	940.7439411028729
38	943.2617776226718
39	945.8015579043479
40	948.2865664719106
41	950.6526507944943
42	952.8482924413679
43	954.8342808239652
44	956.583089164599
45	958.0780030643236
46	959.3120546620098
47	960.2868296547704
48	961.0111981599207

3. 讨论

本研究对柳州市 ILI 的发病趋势进行了短期预测，是第一次应用了时间序列分析的 ARIMA 模型，为当地的流感防控工作提供了基础依据。本研究建立了季节性 ARIMA 模型，根据柳州市 2018 年 1 月 –2022 年 12 月的 ILI 发病数，对进

行短期预测了 2023 的 ILI 发病数，将模型分析后的预测值和实际观测值进行对比发现，预测曲线整体动态趋势与实际值曲线基本吻合，且实际值均落在了预测值的 95% 置信区间内，能说明该模型能够一定程度短期预测柳州市流感样病例的发病趋势。田竞等 [2] 建立季节性差分自回归滑动平均模型（SARIMA）模型，进行参数估计和预测北京市房山区流感样病例趋势；祝小平等 [3] 构建月度发病数 ARIMA 时间序列模型，分析四川地区流感趋势；赵俊等 [4] 利用 ARIMA 模型预测新疆地区流感样病例占门急诊病例的百分比，这些研究结论与本研究有相同之处。ILI 监测属于症状检测的范围，是体现流感流行趋势的重要指标，对流感疫情的发生有预测预警作用。本研究对柳州市 2018 年至 2022 年 ILI 病例数进行研究分析得出，当地 ILI 数在研究期间呈现一定的季节性，呈现冬季或冬夏季高发，并呈现一定上升的趋势，这种流行趋势与某些地区的研究结果不同，比如银川市的研究发现，当地流感高峰在 12 月至次年 1 月，最终选择最佳模型 ARIMA（0，2，0）（0，2，0）预测当地流感流行发病趋势。新疆的一项研究发现，2012–2014 年新疆每周的 ILI% 呈冬季高发，具有明显的季节性，并最终建立季节性 ARIMA（1，0，1）（0，1，0）模型预测新疆 ILI%。气候条件的不同可能是造成这种差异的主要原因，银川和新疆均属于温带大陆性气候，而柳州地处于亚热带季风气候区，在该气候条件下流感的季节性变化不显著，一般情况下可出现秋冬季和夏季两个高峰。目前，传染病预测预警领域被引入了越来越多的数理统计方法，从而成为传染病防控的重要技术手段。ARIMA 模型具有短期预测效果好、操作简单（可以通过软件快速分析）、不需要考虑复杂的外部因素的影响等特点，已成为传染病防控领域最为常用的预测预警模型之一 [5–10]。

但受模型方法自身的局限性（如数据波动较大、模型对于季节性变化的敏感性、分析软件自身局限性等）以及传染病突发事件会影响预测精度，该模型仅适用于短期预测，未来可将每周的 ILI 监测数据及时加入序列，同时在有条件的情况下继续收集相关数据，如气象数据、人口流动及人群免疫力等因素，重新建模，以确保该模型的准确性。也可考虑采用其他模型比如神经网络模型、SIR 模型等方法，对柳州市 ILI 的发病数进行更精准的预测。

综上，ARIMA 模型在流感样病例预测中具有一定的适用性，而季节性 ARIMA 模型可以较好地对柳州市流感样病例的发病趋势进行短期阶段预测。

参考文献：

[1] 林梦宣，陈辉，宋宏彬等 . 基于互联网大数据的传染病预测预警研究进展 [J]. 中国公共卫生 ,2021,37(10):1478–1482.

[2] 田竞，韩晓畅，任雅楠，等 .2017–2021 年北京市房山区流感样病例病原学监测与预测预警分析 [J]. 预防医学论坛 ,2023,29(05):326–332.

[3] 祝小平，刘伦光，陈秀伟，等 .2010–2018 年四川省流行性感冒时空流行特征分析及其短期预测 [J]. 预防医学情报杂志 ,2020,36(09):1097–1102.

[4] 赵俊，刘万里，李新兰，等 .ARIMA 模型在新疆流感样病例占门急诊病例百分比预测中的应用 [J]. 职业与健康 ,2016,32(21):2954–2957.

[5] 赵金华，龙江，马永成等 . 高原地区青海省流感病毒活动规律的时间序列分析及预测模型研究 [J]. 医学动物防制 ,2021,37(11):1030–1034.

[6] 钱晨嗣，姜晨彦，夏寒等 . 上海市流感样病例就诊百分比时间序列分析和预测模型研究 [J]. 上海预防医学 ,2023,35(02):116–121.

[7] 余艳妮，廖青，聂绍发等 . 岳阳市流感样病例发病趋势的时间序列分析 [J]. 中国社会医学杂志 ,2020,37(06):650–653.

[8] 闫祥祥 . 使用 ARIMA 模型预测公园绿地面积 [J]. 计算机科学 ,2020,47(S2):531–534+556.

[9]Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software].

[10] 陈丽丽 .2012–2021 年湖南省流感流行特征、气象因素相关性及预测预警研究 [D]. 中南大学 ,2023.

作者简介：

杨俊（1988—），男，汉族，广西柳州人。学历：硕士研究生，柳州市疾病预防控制中心，主管医师，主要从事传染性疾病控制工作。研究方向：传染性疾病控制。

基金项目：

广西壮族自治区卫生健康委员会自筹经费科研课题（项目编号：Z20211223）