Modeling and Simulation 建模与仿真, 2022, 11(6), 1595-1604
Published Online November 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.116150

Hans汉斯

ARIMA模型在预测新型冠状病毒传播中的应用

孟得新，马新然，杨雨欣，杨　洁，许　韬，王立群，严彦文

中国石油大学(北京)理学院，北京

收稿日期：2022年9月30日；录用日期：2022年11月17日；发布日期：2022年11月24日

摘　要

目前，新冠病毒仍在全球肆虐。为了更好地模拟病毒的传播情况，本文基于疫情数据，应用ARIMA模型对巴基斯坦的新冠疫情趋势进行预测，对模型的相关性质进行检验。最后分析了该模型在传染病预测中的应用价值，为新冠疫情预测提供实践经验。

关键词

ARIMA模型，新冠肺炎，传染病预测

Application of ARIMA Model in Transmission Prediction of the COVID-19

Dexin Meng, Xinran Ma, Yuxin Yang, Jie Yang, Tao Xu, Liqun Wang, Yanwen Yan

College of Science, China University of Petroleum (Beijing), Beijing

Received: Sep. 30th, 2022; accepted: Nov. 17th, 2022; published: Nov. 24th, 2022

Abstract

At present, the COVID-19 is still rampant in the world. In order to better simulate the spread of the virus, based on the epidemic data of Pakistan, the ARIMA model is used to predict the trend of the COVID-19 in this paper, and the relevant properties of the model are tested. Finally, the application value of the model in the prediction of infectious diseases is analyzed, which provides practical experience for the prediction of the COVID-19.

文章引用: 孟得新, 马新然, 杨雨欣, 杨洁, 许韬, 王立群, 严彦文. ARIMA 模型在预测新型冠状病毒传播中的应用[J]. 建模与仿真, 2022, 11(6): 1595-1604. DOI: 10.12677/mos.2022.116150

Keywords

ARIMA Model, COVID-19, Prediction of Infectious Diseases

1. 模型背景

1.1. 问题选择

从 2019 年冬天开始，新型冠状病毒引起的肺炎疫情首先在中国爆发，然后在全球蔓延，对整个世界的人类安全和经济产生了巨大影响。世卫组织报告显示，新冠肺炎的全球死亡率约为 6.4%。虽然一些地区疫情得到了有效控制，但全球疫情仍然非常严重。合理有效的数学建模可以为政府的防疫决策、疾病预防和控制提供很大的帮助。

为更好预测疫情的未来发展状况，需要对各类预测模型的优劣进行研究探讨。目前，对新冠疫情的预测主要是应用传染病传播动力学模型，如：SEIR 模型、SIR 模型等，但传染病动力学模型需要对各种模型参数及其未来的变化趋势有了解，而这些参数很难获取，时间序列模型[1] [2] [3] [4]则只需要有病例数的历史序列就可以构建病例数的预测模型，所以我们尝试利用时间序列模型来进行预测。

1.2. 区域选择

巴基斯坦位于南亚次大陆西北部，南濒阿拉伯海，东接印度，东北邻中国，西北与阿富汗交界，西邻伊朗，是我国重要的战略合作伙伴。我国的新型冠状病毒肺炎疫情虽然呈现多点散发态势，但总体已经得到了有效控制，而巴基斯坦每天仍有新增确诊病例，未来疫情发展趋势是大家关注的问题[5]。

本文基于 2021 年 1 月 15 日至 3 月 15 日的巴基斯坦疫情发展相关的时间序列数据，通过数据处理、曲线拟合、参数估计等过程，反复试验，建立 ARIMA 时间序列预测模型[2] [3] [4] [5]，对巴基斯坦的新冠疫情趋势进行短期预测，并对模型的合理性进行检验。以此来探讨 ARIMA 时间序列模型在传染病未来发展走向的预测中的应用方法与研究价值，为新冠疫情预测提供实践经验。

2. 模型理论

2.1. 差分

差分，一般用于以时间为统计维度的分析中，反应了离散量之间的一种变化，它可以减轻数据之间的不规则波动，使其波动曲线更平稳。时间距离为 1 的两个序列值，做减法运算，可以得到 1 阶差分，迭代下去，可以得到 p 阶差分，一般记为：$\nabla^p x_t = \nabla^{p-1} x_{t-1}$。

2.2. 白噪声的检验

下面来检验序列是否为纯随机序列，也即是进行白噪声检验。理想的结果是观察值序列为非白噪声序列，这样我们就可以使用平稳时间序列对模型进行预测；而残差序列应具有完全随机性，理想的结果是白噪声。

假设条件：H_0：$\rho_1 = \rho_2 = \cdots = \rho_m = 0, \forall m \geq 1$；

H_1：至少存在某个 $\rho_k \neq 0, \forall m \geq 1, k \leq m$。

构造统计量：

$$Q = n\sum_{k=1}^{m} \hat{\rho}_k^2 \sim \chi^2(m).$$

此时可以比较统计量和上分位点 $\chi_{1-\alpha}^2(m)$，来做出拒绝或者接受原假设的判断；或者计算统计量的 P 值，当 P 值小于 α 时，能够在 $1-\alpha$ 的置信水平下拒绝原假设，该序列不是白噪声；否则就接受原假设，序列是白噪声。如果样本数量较小，可以使用 LB 统计量：

$$\mathrm{LB} = n(n+2)\sum_{k-1}^{m}\left(\frac{\hat{\rho}_k^2}{n-k}\right) \sim \chi^2(m).$$

2.3. ARIMA 模型

本文选用的主要模型是求和自回归移动平均(ARIMA(p,d,q))模型，该模型能够将非平稳时间序列转化为差分平稳序列，应用差分平稳序列，我们可以进行 ARIMA 拟合，最后进行预测。

ARIMA(p,d,q)模型的基本形式：

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, s \neq t, \\ E(x_s\varepsilon_t) = 0, \forall s < t \end{cases}$$

其中，$\nabla^d = (1-B)^d$；$\Phi(B) = 1-\varphi_1 B - \cdots - \varphi_p B^p$，$\Theta(B) = 1-\theta_1 B - \cdots - \theta_p B^p$ 分别是平稳可逆 ARMA(p,q)模型的自回归系数多项式和移动平滑多项式；$\{\varepsilon_t\}$ 是均值为 0 的白噪声序列。

ARIMA 模型的实质就是差分运算与 ARMA 模型的组合。这意味着任何非平稳序列如果能通过适当阶数的差分运算变得平稳，就可以对差分后的序列进行 ARMA 拟合。

ARIMA 模型建模过程：

1) 获得观察值序列；
2) 平稳性检验：若序列不平稳，检验差分后的序列是否平稳；
3) 白噪声检验：通过平稳性检验的序列若不是白噪声序列就可以拟合 ARMA 模型；
4) 利用该模型预测未来趋势。

2.4. 残差自相关检验

下面来检验模型的拟合效果。如果残差序列满足

$$E(\varepsilon_t, \varepsilon_{t-j}) = 0, \forall j \geq 1,$$

则说明是白噪声序列，因此模型拟合效果较好。下一步就可以充分提取序列中的相关信息，不需要修正拟合模型；反之，如果残差序列满足

$$E(\varepsilon_t, \varepsilon_{t-j}) \neq 0, \forall j \geq 1,$$

则说明自相关性显著，模型拟合效果不好，需要修正模型再次拟合。具体操作如下：

假设条件：H_0：$E(\varepsilon_t, \varepsilon_{t-1}) = 0$，即 $\rho = 0$；H_1：$E(\varepsilon_t, \varepsilon_{t-1}) \neq 0$，即 $\rho \neq 0$

DW 统计量：$\mathrm{DW} = \dfrac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=2}^{n}\varepsilon_t^2} \approx 2\left(1 - \dfrac{\sum_{t=2}^{n}\varepsilon_t\varepsilon_{t-1}}{\sum_{t=1}^{n}\varepsilon_t^2}\right)$

自相关系数：$\rho = \dfrac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2}$

即 $\mathrm{DW} \approx 2(1-\rho)$

当 $0 < \rho \le 1$ 时，序列正相关，且 $\rho \to 1$ 时，$\mathrm{DW} \to 0$；$\rho \to 0$ 时，$\mathrm{DW} \to 2$。当 $-1 < \rho \le 0$ 时，序列负相关，且 $\rho \to -1$ 时，$\mathrm{DW} \to 4$；$\rho \to 0$ 时，$\mathrm{DW} \to 2$。DW 值越接近于 2，ρ 值越小，拟合效果越好。

2.5. 参数的显著性检验

为了使模型更精简，需要对参数做显著性检验。如果某个参数所对应的自变量对因变量的影响不明显，经过显著性检验之后就可以把对应的自变量剔除，进而实现对模型的精简。

做假设检验：

$$\mathrm{H}_0 : \beta_j = 0 \leftrightarrow \mathrm{H}_1 : \beta_j \ne 0, \forall 1 \le j \le m.$$

构造 t 检验统计量：

$$T = \sqrt{n-m}\,\frac{\hat{\beta}_j}{\sqrt{a_{ij} Q(\tilde{\beta})}} \sim t(n-m).$$

代入样本值，如果满足

$$|T| \ge t_{1-\alpha}(n-m),$$

则拒绝原假设；或者计算检验统计量的 P 值，当 P 值小于 α 时，也能够在 $1-\alpha$ 的置信水平下拒绝原假设，则这个参数效果显著，不能剔除。反之，如果这个参数的效果不显著，可以在模型中剔除该参数，重新拟合模型即可。

3. 模型建立与问题求解

3.1. 获得观察值序列

从网易疫情实时数据平台获取巴基斯坦地区 2020 年 3 月至 2021 年 6 年新冠肺炎确诊人数数据，对该数据进行初步筛选分析后将 2021 年 1 月 15 日至 3 月 15 日这 60 组确诊人数数据作为研究对象。

3.2. 判断序列平稳性

确诊人数序列时序如图 1 所示。

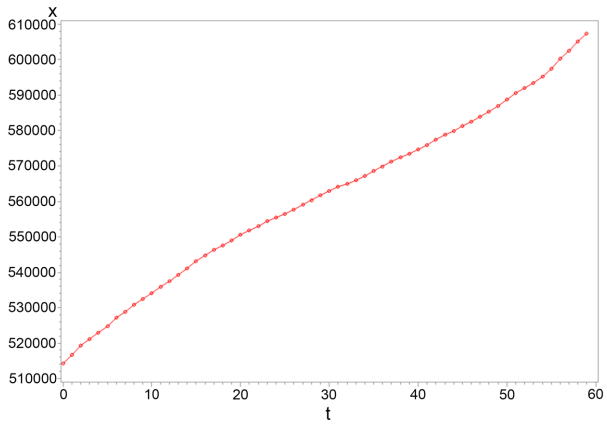

Figure 1. Time series of confirmed infection
图 1. 确诊人数时序图

从图 1 中可以看出该序列有显著的递增趋势，为典型的非平稳序列。可以考虑对该序列进行差分，获得平稳序列，建立ARIMA模型。

3.3. 对原序列进行差分运算

首先尝试对原序列进行一阶差分，一阶差分后序列时序如图 2 所示：

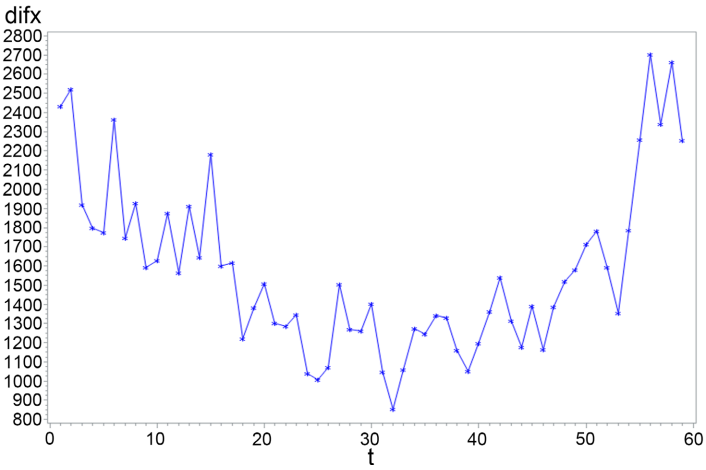

Figure 2. First order difference time series
图 2. 一阶差分时序图

由图 2 我们发现，差分后序列波动较大，仍不平稳，可以进一步对一阶差分序列再次差分得到二阶差分时序图 3。

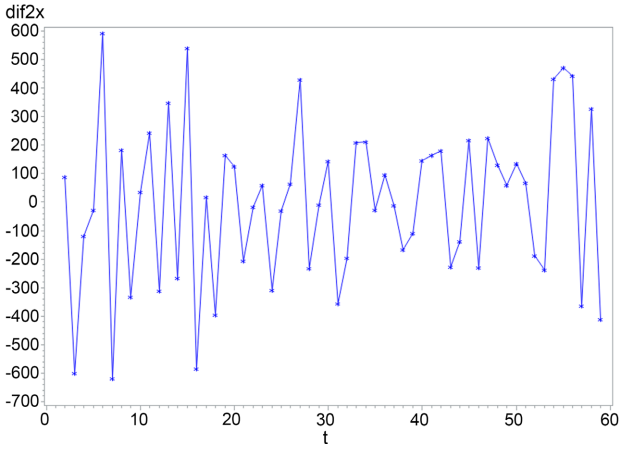

Figure 3. First order difference time series
图 3. 一阶差分时序图

图 3 显示，二阶差分后序列在均值附近比较平稳地波动，为了进一步确定该时序的平稳性，考察二阶差分后序列的自相关图 4。

自相关图 4 显示序列具有较强的短期相关性，所以可以初步认为二阶差分后序列平稳。

3.4. 对平稳的二阶差分序列进行白噪声检验

白噪声检验结果如表 1 所示。

自相关

滞后	协方差	相关性	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	标准误差
0	79253.906	1.00000	|**********|	0
1	-26219.003	-.33082	*******| . |	0.131306
2	11239.362	0.14181	. |*** . |	0.144967
3	-22999.498	-.29020	*******| . |	0.147339
4	7128.807	0.08995	. |** . |	0.156885
5	-3369.282	-.04251	. *| . |	0.157771
6	-2209.189	-.02787	. *| . |	0.157969
7	21278.715	0.26849	. |***** . |	0.158054
8	-7229.089	-.09121	. **| . |	0.165731
9	10878.614	0.13726	. |*** . |	0.166594
10	-13692.752	-.17277	. ***| . |	0.168533
11	272.128	0.00343	. | . |	0.171559
12	-11718.833	-.14786	. ***| . |	0.171560
13	7982.840	0.10072	. |** . |	0.173744
14	12639.267	0.15948	. |*** . |	0.174748

"." 标记两个标准误差

Figure 4. Autocorrelogram
图 4. 自相关图

Table 1. White noise autocorrelation test
表 1. 白噪声自相关检验表格

白噪声的自相关检查									
至滞后	卡方	自由度	Pr > 卡方	自相关					
6	13.95	6	0.0302	−0.331	0.142	−0.290	0.090	−0.043	−0.028
12	24.61	12	0.0168	0.268	−0.091	0.137	−0.173	0.003	−0.148

取显著性水平 $\alpha = 0.05$，从表 1 可以发现，不管是延迟 6 阶还是 12 阶，计算出来的 P 值均小于 0.05，分别为 0.0302 和 0.0168，因此在 $\alpha = 0.05$ 的显著性水平下，拒绝原假设。因此需要进一步地提取相关信息。

3.5. 对平稳非白噪声差分序列拟合 ARIMA 模型

观察图 5 中的二阶差分后序列的自相关图，发现序列的自相关系数具有 1 阶截尾的特性，下面来计算序列的偏自相关性。

观察偏自相关图 5，可以发现序列的截尾性不显著，因此在拟合二阶差分后序列的时候，可以应用 MA(q)模型。

3.6. ARIMA 模型定阶

为了按照最小信息量准则确定一个相对最优的模型，利用 SAS 软件分析数据，得到最小信息量表 2，结果如下：

偏自相关		
滞后	相关性	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.33082	|*******| . |
2	0.03635	| . |* . |
3	-0.26194	| *****| . |
4	-0.09630	| . **| . |
5	-0.02183	| . | . |
6	-0.13956	| . ***| . |
7	0.26995	| . |***** |
8	0.07940	| . |** . |
9	0.11636	| . |** . |
10	0.07185	| . |* . |
11	-0.07890	| . **| . |
12	-0.14045	| . ***| . |
13	-0.01705	| . | . |
14	0.11542	| . |** . |

Figure 5. Partialautocorrelogram
图 5. 偏自相关图

Table 2. Minimum information
表 2. 最小信息量表格

Minimum Information Criterion						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	10.94285	10.92593	10.96785	10.96612	10.91432	10.88622
AR 1	10.89763	10.96706	11.00043	11.00089	10.98433	10.93492
AR 2	10.93903	11.00875	10.99937	11.04563	10.97991	10.98544
AR 3	10.90604	10.97474	11.03039	11.09621	11.03556	11.00962
AR 4	10.90676	10.96291	11.01945	11.07318	11.02092	10.97937
AR 5	10.85219	10.91391	10.97512	11.03322	10.97982	11.04553

根据最小信息量准则，AR(5)模型的结果最优，可以用来拟合差分序列。

Table 3. Conditional least squares estimation and significance test
表 3. 条件最小二乘估计以及显著性检验表格

条件最小二乘估计					
参数	估计	标准误差	t 值	近似 Pr > |t|	滞后
MU	1.43908	20.17253	0.07	0.9434	0
AR1,1	−0.34258	0.13914	−2.46	0.0172	1
AR1,2	−0.04303	0.14911	−0.29	0.7740	2
AR1,3	−0.31251	0.14203	−2.20	0.0322	3
AR1,4	−0.07688	0.15280	−0.50	0.6170	4
AR1,5	0.0060323	0.15481	0.04	0.9691	5

由条件最小二乘估计以及显著性检验表 3 可以得到拟合模型为：

$$\left(1+0.34258B+0.04303B^2+0.31251B^3+0.07688B^4-0.00603B^5\right)\nabla x_t=\varepsilon_t$$

3.7. ARIMA 模型预测

首先，我们不对参数进行剔除与检验，直接利用该组参数所对应的 AR(5)模型进行序列预测，得到的预测值如表 4 所示：

Table 4. Model AR(5) prediction
表 4. AR(5)模型预测表格

以下变量的预测：x				
观测	预测	标准误差	95%置信限	
61	609917.5	267.45	609393.3	610441.7
62	612258.6	517.72	611243.8	613273.3
63	614736.5	822.40	613124.6	616348.4
64	617142.9	1111.86	614963.7	619322.1
65	619590.3	1426.91	616793.6	622387.0
66	621997.2	1762.16	618543.5	625451.0
67	624429.9	2131.49	620252.3	628607.6

为了探究模型是否可得到一定程度的简化，下面对模型不显著的参数进行剔除，只留下通过显著性检验的变量，同时重新估计参数值，得到表 5：

Table 5. Parameter estimation and significance test of sparse coefficient model
表 5. 疏系数模型参数估计与显著性检验表格

条件最小二乘估计					
参数	估计	标准误差	t 值	近似 Pr > |t|	滞后
AR1,1	−0.31160	0.12376	−2.52	0.0147	1
AR1,2	−0.28172	0.12722	−2.21	0.0309	3

即最终得到的拟合模型为：

$$\left(1+0.3116B+0.28172B^3\right)\nabla x_t=\varepsilon_t$$

为了判断该模型是否显著，是否还有其他可提取的信息，我们来检验残差序列是否为白噪声序列，结果见表 6。从表中可以看出，P 值均大于显著性水平 0.05，因此该模型是显著的，残差序列具有纯随机性，因此可以应用该模型来进行预测。

Table 6. Residual autocorrelation test
表 6. 残差自相关检验表格

残差的自相关检查									
至滞后	卡方	自由度	Pr > 卡方	自相关					
6	0.42	4	0.9808	−0.031	−0.026	−0.056	0.038	−0.004	0.019

Continued

12	8.77	10	0.5538	0.316	−0.003	0.071	−0.088	−0.067	−0.063
18	17.90	16	0.3297	0.073	0.262	0.153	−0.093	−0.087	−0.014
24	29.23	22	0.1383	−0.036	−0.017	0.232	0.132	−0.180	−0.109

利用此拟合模型对确诊人数序列进行 7 期(一周)预测，预测值以及 95%的预测区间如表 7 所示：

Table 7. Sparse coefficient prediction
表 7. 疏系数预测表格

以下变量的预测：x				
观测	预测	标准误差	95%置信限	
61	609936.3	258.66	609429.4	610443.3
62	612256.1	507.58	611261.2	613250.9
63	614742.5	816.77	613141.7	616343.4
64	617112.2	1117.29	614922.3	619302.0
65	619564.3	1455.44	616711.7	622416.9
66	621943.8	1813.01	618390.3	625497.2
67	624378.8	2207.27	620052.6	628705.0

我们将参数剔除前与剔除后的预测值与真实值做对比，并分别计算偏差平方和，以对比剔除前后模型预测效果的优劣，具体结果见表 8：

Table 8. Comparison of predicted values
表 8. 预测值对比结果

以下变量的预测：x			
观测	剔除前	剔除后	真实值
61	609917.5	609936.3	609,964
62	612258.6	612256.1	612,315
63	614736.5	614742.5	615,810
64	617142.9	617112.2	619,259
65	619590.3	619564.3	623,135
66	621997.2	621943.8	626,802
67	624429.9	624378.8	630,471
偏差平方和	7.7782e+07	7.9219e+07	-
AIC	818.5811	811.0018	-

剔除前后两种模型的预测图像如图 6 所示：

对比可知，剔除前的模型虽然 AIC 信息量较大，但其拟合的偏差平方和较小，其预测效果相对于剔除后的疏系数模型较为准确。

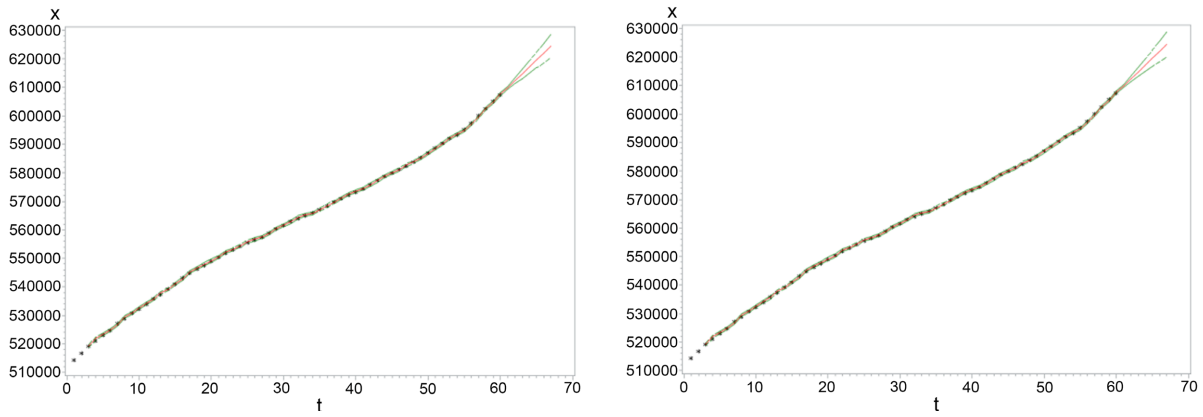

Figure 6. Predictionresults of AR (5) and sparse coefficient models
图 6. AR(5)以及疏系数模型预测图像

4. 模型评价与推广

4.1. 模型评价

本文针对巴基斯坦近两个月新冠肺炎确诊人数非线性非平稳的数据，基于 ARIMA 模型建立时间序列模型进行预测。首先通过差分得到平稳时间序列模型，基于理论知识与 SAS 软件的应用对模型的平稳性、参数及残差相关性进行检验。然后用该模型进行短期预测。将预测值与真实值对比发现在只有传染病随时间变化的时间序列历史数据的前提条件下，ARIMA 时间序列模型对传染病的未来发展趋势具有较好的预测效果。

但由于时间序列模型对短期预测具有较好的效果这种局限性以及该模型是在传染病的传播发展模式未发生根本性变化的情况下对传染病发展趋势所做的预测，通常难以保证长期预测结果的准确性，该模型只适用于对短期确诊人数的预测。

4.2. 模型推广

在模型后续的改进和优化时可以考虑及时将新的数据加入时间序列，将这种建模及预测过程全程自动化并与病例监测报告系统进行集成，实现实时动态建模和预测，可为今后开展疫情监测提供便捷的手段。

基金项目

中国石油大学(北京)教育教学改革项目(编号 XM10720210153、XM10720200035、YJS2020032，YJS2020033)。

参考文献

[1] 王燕. 应用时间序列分析[M]. 北京: 中国人民大学出版社, 2005: 147-152.

[2] 杨真真, 谢艳秋, 靳旭东, 庄桂敏. 基于 ARIMA 时间序列模型的传染病发展趋势预测——以 COVID-19 为例[J]. 中国科技信息, 2021(Z1): 70-72.

[3] 李荞每, 成丽波. 基于小波分析的时间序列 ARIMA 模型预测方法[J]. 沈阳师范大学学报(自然科学版), 2021, 39(1): 49-53.

[4] 景楠, 胡怡, 韩喜双. 基于 ARIMA 与 LSTM 的新冠肺炎网络关注度趋势研究[J]. 中国安全科学学报, 2020, 30(12): 37-42.

[5] 温亮, 黄清臻, 王志刚, 等. 运用 ARIMA 模型预测巴基斯坦新型冠状病毒肺炎疫情发展趋势的结果分析[J]. 解放军预防医学杂志, 2020, 38(8): 96-99+102.