Statistics and Application 统计学与应用, 2020, 9(6), 979-987
Published Online December 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.96103

Hans汉斯

基于ARIMA模型在美国COVID-19累计确诊人数中的应用

谢　旺[1]，牟　银[2]，宋佳硕[3]

[1]沈阳航空航天大学自动化学院，辽宁 沈阳
[2]遵义医科大学医学与科技学院护理学院，贵州 遵义
[3]沈阳航空航天大学自动化学院，辽宁 沈阳
Email: 3379822855@qq.com

收稿日期：2020年11月16日；录用日期：2020年12月14日；发布日期：2020年12月21日

摘　要

本文针对新型冠状病毒肺炎给世界人民造成的不良影响，收集2020年1月20日~2020年6月1日内以美国为主的各个国家和地区每日COVID-19累计确诊人数，建立自回归求和滑动平均(auto regressive integrated moving average, ARIMA)模型对美国累计确诊人数进行分析与预测，用SPSS25.0和MATLAB2019a拟合，结合拟合优度R^2和Q检验评价拟合效果，将后5日累计确诊人数预测值和真实值进行比较，评价该模型预测精度及预测美国未来10日累计确诊人数。结果表明，原始序列经2次差分后能较好拟合ARIMA (0,2,1)模型，R^2在0.95以上，Q检验p值为0.19 > 0.05，认为残差为白噪声，且预测值与实际值动态趋势基本一致，预测值在真实值0.33%误差内波动，ARIMA (0,2,1)模型对美国COVID-19累计确诊人数预测精度很高，对疫情防控具有很强的指导意义。

关键词

ARIMA模型，ACF/PACF，COVID-19，白噪声，累计确诊人数

Based on the Application of the ARIMA Model in the Cumulative Number of Confirmed Cases of COVID-19 in the United States

Wang Xie[1], Yin Mou[2], Jiashuo Song[3]

[1]School of Automation, Shenyang University of Aeronautics and Astronautics, Shenyang Liaoning
[2]School of Nursing, College of Medicine and Technology, Zunyi Medical University, Zunyi Guizhou

文章引用: 谢旺, 牟银, 宋佳硕. 基于 ARIMA 模型在美国 COVID-19 累计确诊人数中的应用[J]. 统计学与应用, 2020, 9(6): 979-987. DOI: 10.12677/sa.2020.96103

[3]School of Automation, Shenyang University of Aeronautics and Astronautics, Shenyang Liaoning
Email: 3379822855@qq.com

Received: Nov. 16th, 2020; accepted: Dec. 14th, 2020; published: Dec. 21st, 2020

Abstract

To address the adverse impact of COVID-19 on people around the world, this article collected the cumulative number of daily COVID-19 diagnosed in countries and territories, mainly the United States, from January 20 to June 1, 2020. Auto Regressive Integrated Moving Average Model (ARIMA) was established to analyze and predict the cumulative number of diagnosed cases in the United States. With SPSS25.0 and MATLAB2019a as fitting methods, combined with the R^2 and Q test to evaluate the fitting effect, the predicted and real values for the cumulative number of confirmed cases for the last 5 days were compared to evaluate the prediction accuracy of the model and the cumulative number of confirmed cases for the next 10 days. The results show that the original sequence can fit the ARIMA (0,2,1) model well after two differences. The R^2 is above 0.95, and the p value of the Q test is 0.19 > 0.05 that is regarded as white noise, and the predicted value is basically consistent with the actual value dynamic trend. The forecast value fluctuated within 0.33% of the true value. The ARIMA (0,2,1) model is under a high accuracy in predicting the number of COVID-19 diagnosed in the United States, which has a strong guiding significance for epidemic prevention and control.

Keywords

ARIMA Model, ACF/PACF, COVID-19, White Noise, Cumulative Number of Confirmed Cases

1. 引言

2000 年以来，新发传染病不断涌现，给国家发展和人民安全造成严重威胁。例如：2003 年急性呼吸综合征，2009 年 H1N1 流感，2012 年中东呼吸综合征，2013 年 B7N9 流感，2014 年黄热病，2015 年埃博拉病毒，2016 年寨卡病毒，2019 年新型冠状病毒肺炎(coronavirus disease 2019, COVID-19)等这些传染病给国家和人民生命安全造成极大的影响。COVID-19 主要传播方式为呼吸道传播和接触传播，其病发现象与感冒相似，严重者可导致死亡[1]。截止于 2020 年 6 月 1 日，全球 COVID-19 累计确诊人数持续上升，中国累计确诊案例为 84,588 人，美国累计确诊人数为 1,734,040 人，俄罗斯累计确诊人数为 414,178 人，在欧洲感染严重的国家中，英国累计确诊人数为 274,766 人，德国累计确诊人数为 181,815 人，全球各个国家和地区累计确诊人数热力图，见图 1 所示，据目前疫情形式来看，美国疫情十分严峻，累计确诊人数超过 100 万人，未来仍有大幅上升趋势，当地政府应加大疫情管控力度，提高医疗水平，保障人民生命安全。

本文研究基于自回归求和滑动平均(autoregressive integrated moving average, ARIMA)模型在 COVID-19 中的运用，预测美国未来累计确诊人数，以及分析疫情波动趋势的方法。

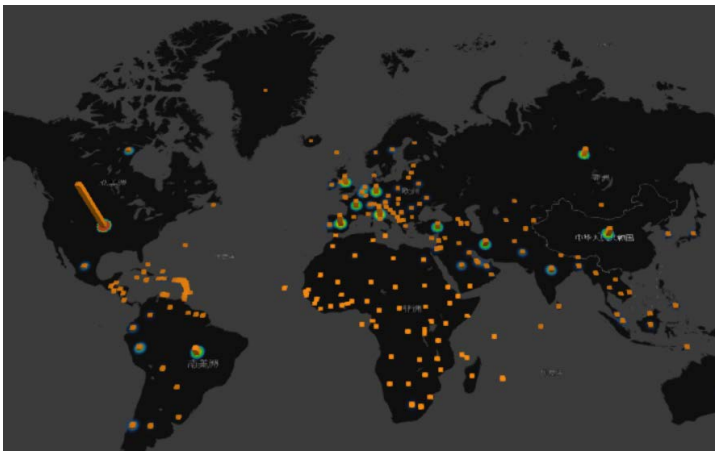

Figure 1. Thermal chart of global cumulative number of confirming cases
图 1. 全球累计确诊人数热力图

2. 资料与方法

2.1. 数据资料

COVID-19 数据来自世界卫生组织(WTO)提供的以美国为主的各个国家和地区每日累计确诊人数变化数据，收集日期为 2020 年 1 月 20 日~2020 年 6 月 1 日。

2.2. ARIMA 模型的建立与分析

2.2.1. ARIMA 模型与时间序列预测的基本思想

时间序列预测是将预测目标按时间顺序排列起来，构建成一个所谓的时间序列，从所构成的这一组时间序列分析过去的变化规律，推理未来变化的可能性及变化趋势和规律(见图 2)，其基本理论是：一方面承认事物在时间尺度上的延续性，运用过去时间序列数据变化规律就能推算出事物发展趋势[2] [3]；另一方面又要舍弃客观因素影响所产生的随机影响，为消除客观因素带来的偏差，利用历史数据进行统计分析，并对数据进行离群值处理，其次利用该数据进行建模。ARIMA 模型是基于时间序列的一种模型，在建立 ARIMA 模型时，要求时间序列数据是平稳的，即时间序列要求是零均值的平稳的随机序列。应用 ARIMA 模型进行时间序列数据预测时，主要有时序数列平稳性的识别、拟合模型的估计和诊断、最优模型的选取与预测三个阶段，利用这三个阶段选取最合适的 ARIMA 模型进行预测分析。

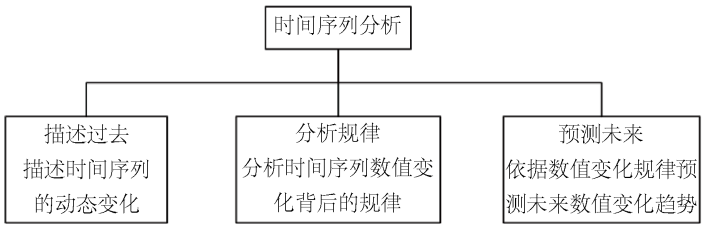

Figure 2. Time series analysis
图 2. 时间序列分析

2.2.2. 研究方法

疫情期间，美国 COVID-19 累计确诊人数是按时间序列发展的，发展趋势各不相同，因此，本文采用时间序列分析方法建模，通过对数据的分析与处理，选用求和自回归移动平均模型(ARIMA)模型进行

建模，并对未来做 10 期预测分析。

ARIMA 模型建模流程如下：

(1) 数据预处理

本文收集了 2020 年 2 月 20 日~2020 年 6 月 1 日内的数据，由于受到医疗设备限制，美国在检验病人是否被 COVID-19 感染时，除核酸检验外，还增加了临床检验等手段，根据病人发病症状进行判断，为了防止疫情的扩散，医生会将疑似感染者列入感染者隔离观察，导致累计确诊人数异常增加，为了还原原始的数值变化规律，将该数据作离群值处理，完成加性、移动水平、革新、瞬态、局部趋势、可加修补操作流程，见图 3 所示。

- 加性(Additive)：影响单个观测值的离群值。
- 移位水平(Level shift)：从某个序列点开始将所有观测值移动到一个常数的离群值，移位水平可能由策略的更改而造成的。
- 革新(Innovational)：在某个特定的序列点附加到噪声项的离群值。对于平稳的序列，革新离群值将影响多个观测值；对于不平稳的序列，它可能影响在某个特定的序列点开始的每个观测值。
- 瞬态(Transient)：其影响按指数衰减到 0 的离群值。
- 局部趋势(Local trend)：从某个特定的序列点开始局部趋势的离群值。
- 可加修补(Addtitive path)：由两个或更多连续可加离群值构成的组，选择此离群值类型将导致除了检测加性离群值的组以外，还检验各个加性离群值。

输出的时间序列数据绘制时间序列图和自相关 ACF 图，如果是平稳时间序列，其走势绕某个固定值上下波动，ACF 图有迅速衰减的趋势，并进行白噪声检验。

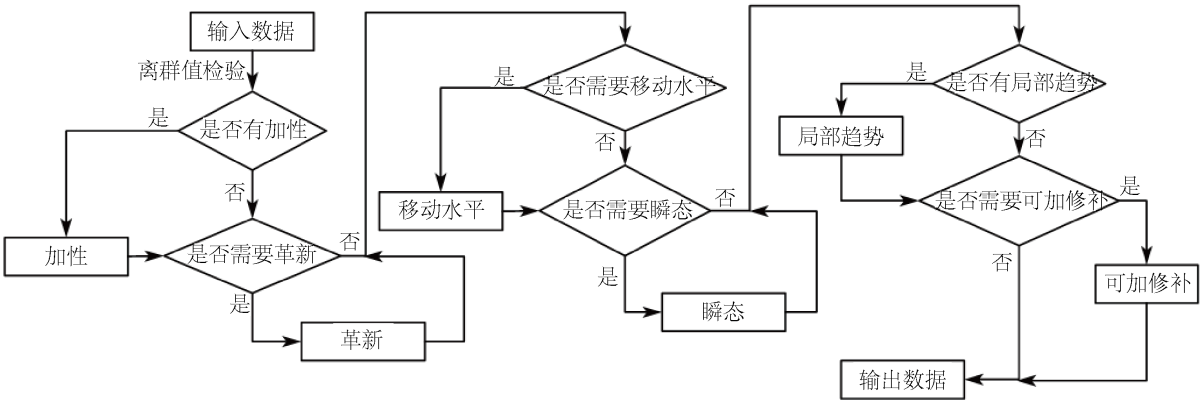

Figure 3. Outlier test flow charts
图 3. 离群值检验流程图

(2) 数据平稳性识别

通过自相关图 ACF 和偏自相关图 PACF 以及时间序列图综合识别该序列是否平稳，若该序列不是平稳的时间序列，需将其做差分处理(1 阶差分或 2 阶差分)，使数据平稳化。

(3) ARIMA 模型 p，d，q 参数估计

P，d，q 是时间序列 ARIMA 模型的三个参数，p 是指时间序列数据本身的滞后数；d 是指时间序列数据稳定时差分次数；预测模型中预测误差的滞后数用 q 表示，分别通过 ACF 和 PACF 的趋势确定该模型的 p，d 值，并结合 AIC 准则(Akaike 信息准则)综合考虑，防止过拟合现象。ARIMA(p,d,q) 序列 AIC 定阶准则为：

$$\min \mathrm{AIC} = n\ln\hat{\sigma}_\varepsilon^2 + 2\left(p+q+1\right) \tag{1}$$

若 $p=\hat{p}, q=\hat{q}$ 时，上式达到最小值，则认为序列是 ARIMA$\left(\hat{p},\hat{d},\hat{q}\right)$；若 ARIMA$\left(p,d,q\right)$ 模型含有未知均值参数 μ 时，模型为：

$$\vartheta\left(L\right)\left(y_t-\mu\right)=\theta\left(L\right)\varepsilon_t \tag{2}$$

见表 1 所示，d 是根据时序数列平稳的差分次数确定的，假设三个参数确定后，时间序列 ARIMA 预测模型也就确定，ARIMA 模型的数学表达式为：

Table 1. Recognition of ARIMA (p, d, q) model
表 1. ARIMA (p, d, q)模型的识别

模型	自相关函数(ACF)	偏自相关函数(PACF)
AR (p)	拖尾	p 阶截尾
MA (q)	q 阶截尾	拖尾
ARMA (p,q)	拖尾	拖尾

$$y_t' = \alpha_0 + \sum_{i=1}\alpha_i y_{t-i}' + \varepsilon_t + \sum_{i=1}^{q}\beta_i\varepsilon_{t-i} \tag{3}$$

且 $y_t' = \Delta^d y_t = \left(1-L\right)^d y_t$

$$\left(1-\sum_{i=1}^{p}\alpha_i L^i\right)\left(1-L\right)^d y_t = \alpha_0 + \left(1-\sum_{i=1}^{q}\beta_i L^i\right)\varepsilon_t$$

$\Uparrow$ $\Uparrow$ $\Uparrow$

AR$\left(p\right)$ d阶差分 MA$\left(q\right)$

其中，L 是滞后算子，α_0 是常数。

(4) 模型评价

在步骤(3)的基础上，进行白噪声检验和 Q 检验该序列的自相关系数 ACF 和偏自相关系数 PACF 均不应超过置信区间，即其数值与 0 比较不应有显著性差异，认为该序列为白噪声序列，若该序列不是白噪声序列，需重复步骤一、步骤二、步骤三操作，重新定阶和 p，q 的确定；若该序列是白噪声序列，计算 R^2，若 R^2 越接近于 1，拟合效果越好。

2.3. ARIMA 模型的运用

本文收集了以美国为主 2020 年 2 月 20 日~2020 年 6 月 1 日内累计确诊人数变化数据，由于美国疫情如今相当严峻，对疫情的管控尤为重要，本文主要对美国疫情累计确诊人数进行建模分析，并对未来 10 日内累计确诊人数进行预测分析，并对美国政府提供具有参考性的建议。

3. 结果

3.1. 数据预处理

将美国 2020 年 2 月 20 日~2020 年 6 月 1 日期间每日累计确诊人数按时间序列排列，完成加性、移动水平、革新、瞬态、局部趋势、可加修补操作流程，绘制时间序列图，见图 4 所示，该序列具有急剧上升的趋势，不满足数据平稳性条件，需对该数据进行差分操作，使其数据平稳化。

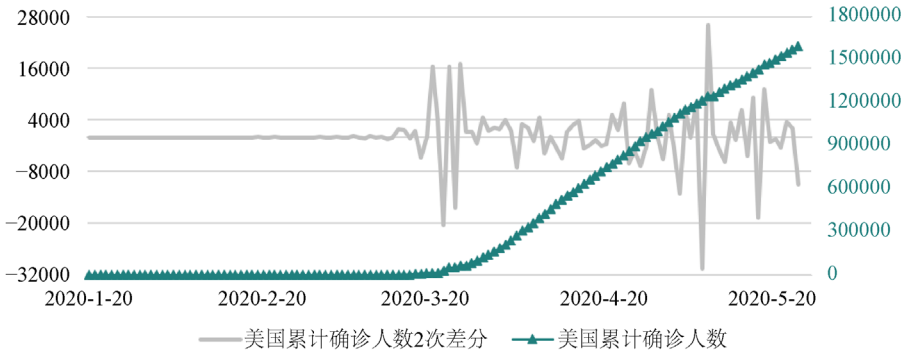

Figure 4. Time series and differential time series of the cumulative number of confirmed cases in the United States
图 4. 美国累计确诊人数时序图及差分时序图

3.2. 数据平稳化

从美国疫情发展趋势来看，累计确诊人数呈“J”型曲线增长，非平稳的时间序列，需对该序列做差分处理，见图 5 所示，1 阶差分后，时序数据仍然有上升的趋势，ACF 和 PACF 均已超出置信区间，判断该序列非平稳；2 阶差分后，时序数据几乎在 0 附近上下波动(见图 4)，ACF 和 PACF 较为均匀分布在置信区间内，可判断该时序数据经 2 次差分后数据平稳，可以对差分后的时序数据做参数估计。

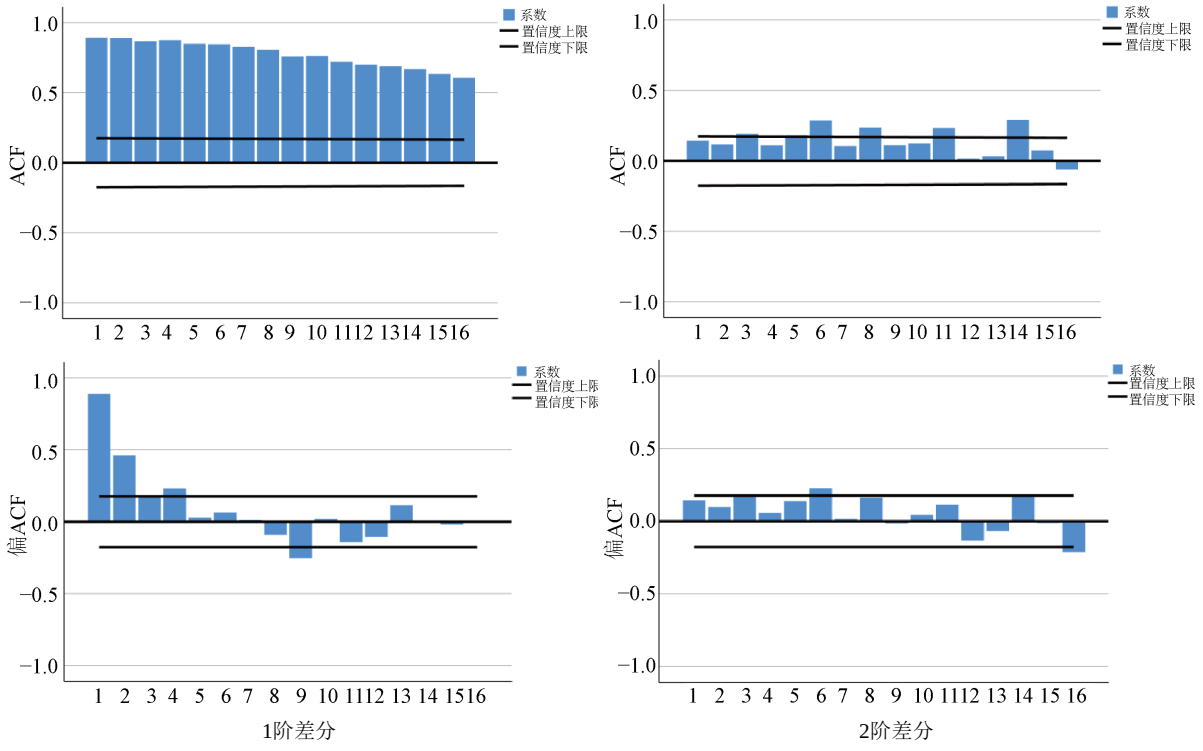

Figure 5. ACF and PACF of COVID-19 differential sequences
图5. COVID-19差分序列ACF图和PACF图

3.3. 参数估计

通过数据平稳化后，确定差分次数为2，然后选取不同的p，q值，进行不同组合的拟合优度检验，提

取不同模型的AIC值，以二者值相对最小模型为最优模型，所选取的最优模型的参数要求都要有统计学意义。结合选取规则确定美国累计确诊人数时间序列预测模型为ARIMA (0,2,1)模型，并对该模型的残差进行白噪声检验，Q值为21.867，其显著性p值为0.19 > 0.05，Q检验接受原假设，该残差是白噪声，残差白噪声检验结果如表2所示，残差的ACF和PACF如图6所示，两者大致在95%CI内，且R^2为0.959，说明该模型拟合效果良好，利用粒子群算法求解参数(见表3)，并对ARIMA (0,2,1)进行口径拟合[4] [5]，模型为：

$$y_t = 2y_{t-1} - y_{t-2} + \varepsilon_t - 0.499\varepsilon_{t-1} \tag{4}$$

运用该模型可完成美国累计确诊人数预测。

Table 2. Residual white noise test results
表2. 残差白噪声检验结果

延迟	ACF	标准误差	PACF	标准误差	延迟	ACF	标准误差	PACF	标准误差
1	−0.002	0.087	−0.002	0.087	13	0.116	0.097	0.111	0.087
2	−0.134	0.087	−0.134	0.087	14	0.027	0.098	−0.013	0.087
3	−0.164	0.089	−0.168	0.087	15	0.127	0.098	0.1	0.087
4	0.19	0.091	0.175	0.087	16	−0.043	0.099	−0.015	0.087
5	0.077	0.094	0.039	0.087	17	−0.016	0.099	−0.052	0.087
6	0.002	0.094	0.02	0.087	18	−0.073	0.099	−0.1	0.087
7	0.027	0.094	0.108	0.087	19	0.002	0.1	−0.092	0.087
8	0.135	0.094	0.135	0.087	20	−0.089	0.1	−0.172	0.087
9	0.078	0.096	0.087	0.087	21	−0.007	0.1	−0.107	0.087
10	0.04	0.096	0.099	0.087	22	0.076	0.1	0.02	0.087
11	0.017	0.096	0.067	0.087	23	0.077	0.101	0.003	0.087
12	−0.053	0.096	−0.071	0.087	24	−0.029	0.101	0.005	0.087

Table 3. Parameter estimation and significant results of particle swarm optimization
表3. 粒子群算法参数估计及显著结果

MA参数估计				
延迟	估算	标准误差	t值	显著性
1	−0.449	0.09	−4.992	0.001

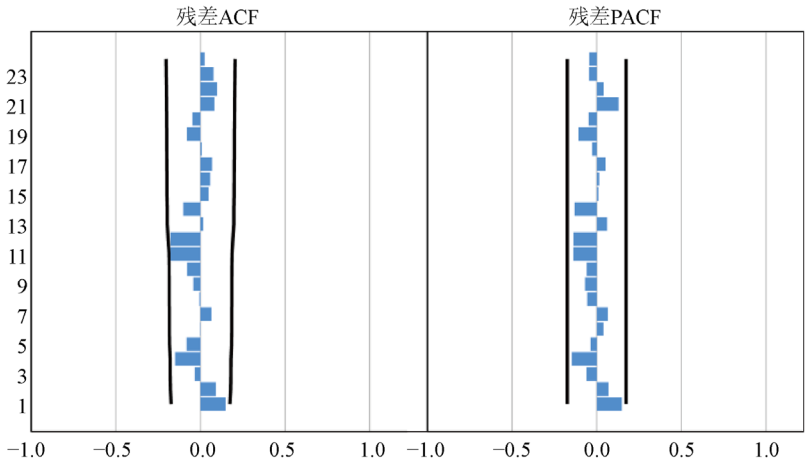

Figure 6. ACF/PACF residual
图6. 残差ACF/PACF图

3.4. 预测运用

运用ARIMA (0,2,1)模型对美国累计确诊人数后5日数据进行回带验证，见表4所示，结果显示，在美国回带验证结果的相对误差大小不超过0.33%，并对未来10内的累计确诊人数进行预测估计(见表5)，在2020年6月11日美国累计确诊人数达到1,928,458人，按照现发展趋势，美国感染人数仍有大幅上升的趋势，见图7所示，几乎每天增加20,000人，美国政府应加大疫情管控力度和有效治疗手段，尽量控制感染人数的增加。

Table 4. Verification results of the daily number of confirmed cases in the US in the last 5 days
表4. 美国后5日每日确诊人数回带验证结果

时间	真实值	预测值	95%置信限		相对误差(%)
5 月 28 日	1658896	1655127	1652409	1657845	0.227199
5 月 29 日	1675258	1676244	1671235	1681253	0.058857
5 月 30 日	1694864	1697361	1689775	1704946	0.147327
5 月 31 日	1716078	1718477	1708028	1728926	0.139796
6 月 1 日	1734040	1739594	1726013	1753175	0.320292

Table 5. ARIMA (0,2,1) models predict the cumulative number of confirmed cases in the United States in the next 10 days
表5. ARIMA (0,2,1)模型预测未来10日美国累计确诊人数

时间	预测值	95%置信限	
6月2日	1753755	1744146	1763363
6月3日	1773166	1758093	1788239
6月4日	1792578	1771225	1813930
6月5日	1811989	1783649	1840329
6月6日	1831401	1795435	1867366
6月7日	1850812	1806635	1894989
6月8日	1870224	1817288	1923159
6月9日	1889635	1827428	1951842
6月10日	1909047	1837082	1981011
6月11日	1928458	1846272	2010644

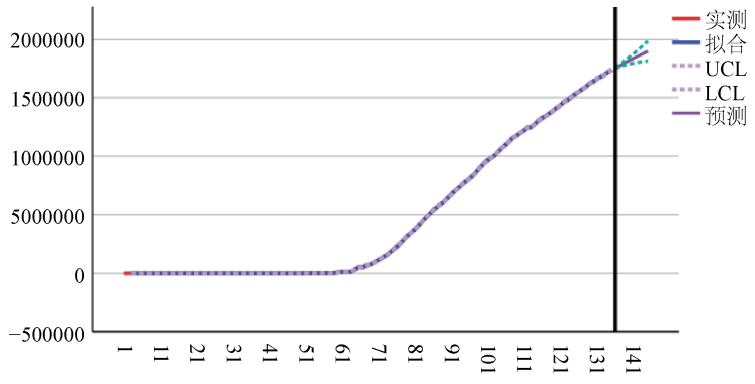

Figure 7. Projected curve of cumulative number of confirming cases in the United States
图7. 美国累计确诊人数预测曲线

4. 结论

新型冠状病毒严重危害人类生命财产安全，给国家及世界造成严重的伤害，为了研究 COVID-19 的传播规律，运用合理的模型对疾病进行预测警告，便于采取相应的防控措施。本文基于该问题建立了ARIMA (0,2,1)模型对美国累计确诊人数进行预测，该模型对于短期预测效果突出，具有很强的预测价值。根据预测结果显示，美国累计确诊人数仍有大幅上升的趋势，说明美国采取的防控措施效果不明显，美国政府应及时调整疫情管控措施，缓解感染人数的上升。

参考文献

[1] 包娅薇, 邵明, 陈雨婷, 刘旭祥, 丁晓芹, 潘贵霞, 潘发明, 李小静. 自回归求和滑动平均(ARIMA)模型在全球新型冠状病毒肺炎发病人数预测中的应用[J]. 中华疾病控制杂志, 2020, 24(5): 543-548.

[2] 白璐, 郭佩汶, 范晋蓉. 湖北省新冠肺炎确诊人数的建模与预测分析[J]. 检验检疫学刊, 2020, 30(2): 10-12.

[3] 王家亮, 钟霞, 沈诗华, 郭德莹, 胡洁, 刘文佳, 孟凡祥. 基于ARIMA乘积季节模型预测医院感染患病率趋势和季节性[J]. 安徽预防医学杂志, 2020, 26(5): 338-334.

[4] 孔德川, 潘浩, 郑雅旭, 姜晨彦, 吴寰宇, 陈健. ARIMA 模型在上海市猩红热发病率预测中的应用[J]. 实用预防医学, 2020, 27(8): 1011-1013.

[5] 赵景平, 谢晴, 曹玉婷. ARIMA 模型在军队呼吸道传染病发病预测中的应用研究[J]. 解放军预防医学杂志, 2020, 38(8): 1-4.