Advances in Applied Mathematics 应用数学进展, 2021, 10(4), 1175-1183
Published Online April 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.104127

Hans汉斯

新型冠状病毒疫情分析及预测研究

郑梦迪，孙菊贺，刘　昊

沈阳航空航天大学理学院，辽宁 沈阳
Email: ipurple1997@163.com, juhesun@163.com, 1722993580@qq.com

收稿日期：2021年3月22日；录用日期：2021年4月11日；发布日期：2021年4月28日

摘　要

本文围绕着新型冠状病毒疫情进行了探究。首先收集了133个国家的时间序列数据，我们发现有不同国家表现出同一特点，为使国家更具代表性，我们从中删除了一些国家，最终确定了31个国家，根据其疫情发展特点，按照系统聚类的方法分类。其次，确定了疫情发展及管控的影响因素，采用灰色关联度分析和模糊综合评价的方法，对主要国家疫情管控效果进行了综合评价。最后，考虑到新型冠状病毒的潜伏期较长，建立了SEIR传染病模型，运用动力学方法模拟疫情发展过程，该过程的关键是寻找最优参数，我们采用最小二乘法。接着预测疫情发展情况并对其进行了F检验，得到F值大于理论值，表明模型拟合效果较好。

关键词

聚类分析，灰色关联度，模糊综合评价，SEIR，F检验

Analysis and Prediction of Novel Coronavirus Epidemic

Mengdi Zheng, Juhe Sun, Hao Liu

School of Science, Shenyang Aerospace University, Shenyang Liaoning
Email: ipurple1997@163.com, juhesun@163.com, 1722993580@qq.com

Received: Mar. 22nd, 2021; accepted: Apr. 11th, 2021; published: Apr. 28th, 2021

Abstract

This article explores the novel coronavirus epidemic. Firstly, we collected time series data from 133 countries and found that there are different countries showing the same characteristics. In order to make the country more representative, we deleted some countries and finally identified 31 countries. According to the epidemic development characteristics, they were classified accord-

文章引用: 郑梦迪, 孙菊贺, 刘昊. 新型冠状病毒疫情分析及预测研究[J]. 应用数学进展, 2021, 10(4): 1175-1183.
DOI: 10.12677/aam.2021.104127

ing to the method of systematic clustering. Secondly, the factors affecting the development and control of the epidemic were determined, and the methods of grey correlation analysis and fuzzy comprehensive evaluation were used to evaluate the effectiveness of epidemic control in major countries. Finally, considering the long incubation period of the new coronavirus, we established the SEIR model and the dynamic method was used to simulate the development process of the epidemic. The key to this process is to find the optimal parameters; we use the least squares method. Then the development of the epidemic was predicted and F test was performed on the prediction model; the F value was greater than the theoretical value, which indicated that the model fitting effect was good.

Keywords

Cluster Analysis, Grey Correlation, Fuzzy Comprehensive Evaluation, SEIR, F Test

1. 引言

新型冠状病毒(COVID-2019)目前已经成为世界流行性传染疾病，对各国的经济、人民的生活、生命财产造成了巨大影响。由于世界各个国家的经济状况、社会体制的不同，再加上人们对疫情重视程度的差异，抗击疫情的方法也不甚相同。有的国家已经基本摸清新冠病毒传播的特点，积极采取有效措施，新冠病毒得到了有效控制，而有的国家还处在疫情初期，还需要继续加强疫情防控。由此可见，防控新型冠状病毒疫情已然成为全球人民共同需要攻克的难题，因此了解世界主要国家疫情特点和防控状况是十分必要的。

对此，我们需要收集相关的时间序列数据[1]，分析世界主要国家的疫情发展特点并对其进行分类。其次，需要确定疫情发展及管控的影响因素，建立合理的数学模型，对主要国家疫情管控效果进行综合评价。最后，根据疫情发展特点，建立合理的数学模型，预测世界主要国家的疫情发展趋势，并检验模型效果。

2. 主要国家疫情特点分析及分类

2.1. 主要国家疫情特点分析

反映疫情发展特点最直接的指标就是每个国家确诊人数、新增确诊人数、治愈人数、死亡人数的增长幅度。由于折线图能更直观地反映这一特点，故可用折线图分析各个国家的疫情特点。下面是对比较有代表性的 6 个国家(美国、中国、西班牙、印度、巴西、澳大利亚)进行疫情特点分析，分析结果如下：

图 1(左)为 2020 年 1 月 21 日~2020 年 7 月 4 日世界主要 6 个国家的新冠确诊人数。从图中可以看出，3 月 21 日之前，中国确诊人数明显高于其它国家，其它国家确诊人数几乎为 0。3 月 21 日之后，中国确诊人数趋于平缓；美国、巴西、印度这三个国家确诊人数明显上升，其中，美国上升速度最快，巴西次之，印度最慢；西班牙 4 月 20 日之后，上升幅度趋于平稳；澳大利亚确诊人数较稳定。

图 1(右)为 2020 年 4 月 7 日~2020 年 7 月 4 日世界主要 6 个国家的新冠确诊人数。从图中可以看出，美国新增确诊人数波动幅度一直较大；5 月 7 日之后，巴西新增确诊数波动幅度变大且整体呈现增长趋势；5 月 17 日之后，印度新增确诊人数，波动幅度较大；其他国家新增确诊人数虽然也在增长，但幅度相对美国、巴西，波动较小。

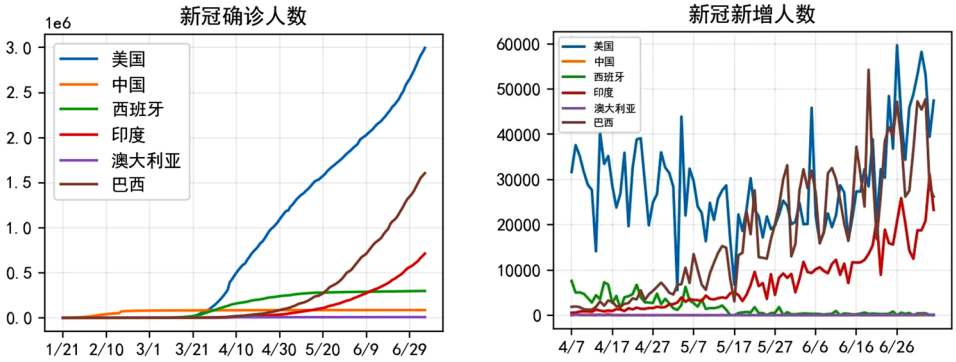


Figure 1. Cumulative number of confirmed cases and new number of confirmed cases
图 1. 累计确诊及新增确诊人数

图 2(左)为 2020 年 1 月 21 日~2020 年 6 月 19 日世界主要 6 个国家新冠病毒的治愈人数。从图中可以看出，2 月 10 日之前，这 6 个国家治愈人数基本为 0；2 月 20 日~3 月 21 日之间，中国新冠治愈人数明显高于其它国家，通过对图 1 的分析，了解到这期间其他国家疫情还未开始，因此其治愈人数几乎为 0。3 月 21 日之后，中国治愈人数趋于平缓，美国、巴西、印度这几个国家曲线明显上升，西班牙刚开始有小幅度上升，5 月 15 日左右趋于平缓；澳大利亚治愈人数有轻微增加，最后趋于平稳。

图 2(右)为 2020 年 1 月 21 日~2020 年 6 月 19 日世界主要 7 个国家的新冠病毒死亡人数。从图中可以看出，2 月 10 日之前，这 6 个国家死亡人数基本为 0；2 月 10 日~3 月 11 日之间，中国新冠病毒开始出现死亡人数，累计死亡人数比较稳定，其它国家死亡人数几乎为 0。3 月 21 日之后，中国死亡人数趋于平缓，美国、巴西增幅较快；西班牙刚开始有小幅度上升，5 月 10 日左右趋于平缓；印度从 4 月 30 日开始，死亡人数开始上升；澳大利亚死亡人数基本趋于平稳。

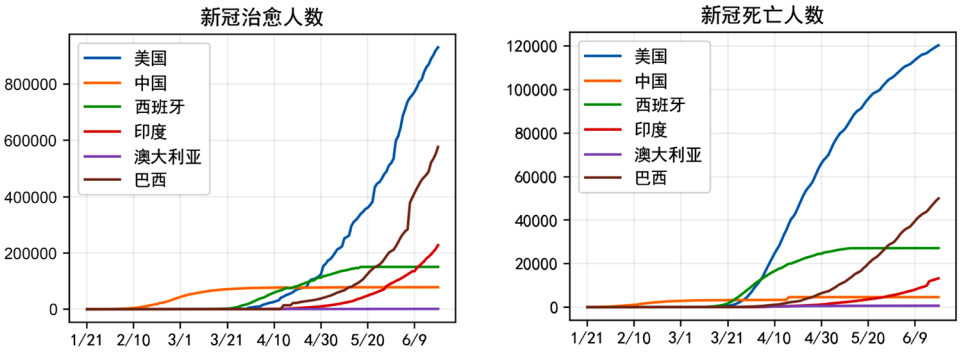


Figure 2. Number of cured and deaths
图 2. 治愈人数及死亡人数

综合上述分析，3 月 1 日中国确诊人数、治愈人数和死亡人数基本趋于平稳，这说明中国在疫情初期认识到了新冠病毒传播特点，并采取积极有效措施，使疫情得到有效控制，西班牙的确诊人数、治愈人数和死亡人数，于 5 月 20 日趋于平稳，由此可知，西班牙的管控效果在 5 月 17 日有明显提高。3 月 21 日开始，美国、巴西、印度的确诊人数、治愈人数和死亡人数明显升高，到 6 月 19 日，仍然还处在上升期，由此可见，美国、巴西、印度疫情防控效果较差，需及时采取有效的防控措施。

2.2. 聚类分析

聚类分析是统计中最常用的一种基本方法。由于我们需要根据时间序列数据，将世界主要国家进行

分类，因此，我们可以进行聚类。基于欧氏距离在坐标轴正交旋转时的稳定性，可采用欧氏距离计算变量间的相似性度量[2]。通过以上分析发现，有不同国家表现出同一特点，为使国家更具代表性，我们从中删除了一些国家，最终确定了 31 个国家，对其 1 月 21 日~6 月 19 日的确诊人数使用 SPSS 软件进行聚类，聚类结果如下图 3 所示。

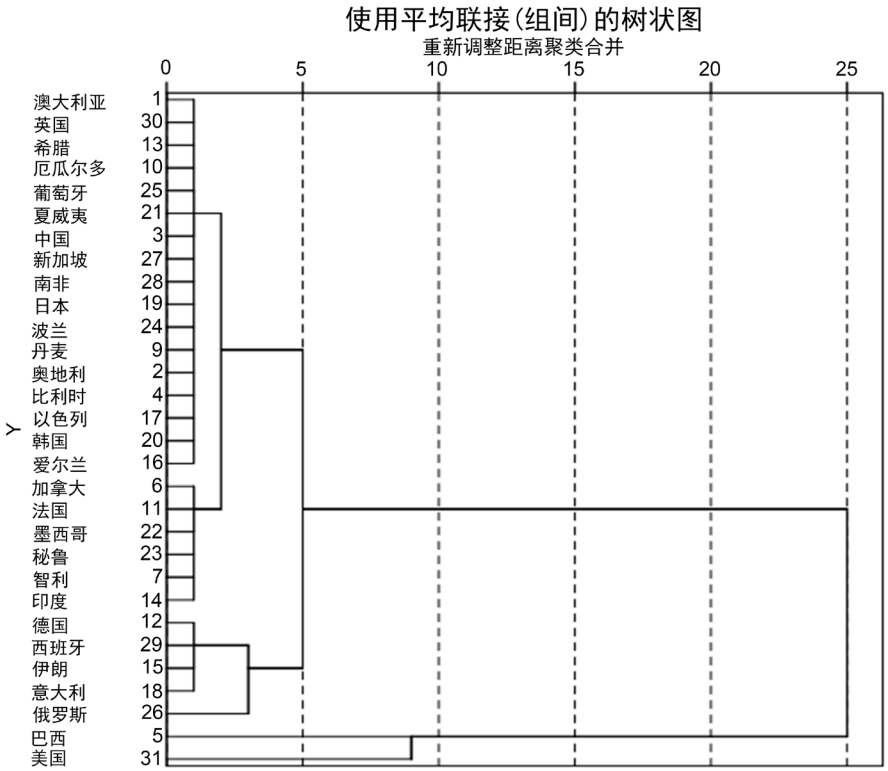


Figure 3. Class average pedigree
图 3. 类平均法谱系图

从图 3 可以看出，世界各主要国家被分为了 4 类。第一类为美国、巴西；第二类德国、西班牙、伊朗、意大利、俄罗斯；第三类为加拿大、法国、墨西哥、秘鲁、智利、印度；第四类中国、新加坡等。

3. 模糊综合评价模型

根据上述对疫情发展特点和管控效果的分析，我们选取了 6 个具有代表性的国家：中国、美国、西班牙、巴西、印度、澳大利亚。选取了 6 个影响疫情管控效果的指标：医院数量、感染率、人口密度、核酸检测(万)、呼吸机数量(每百万人)、封城时间。

3.1. 灰色关联度分析

数据归一化处理：

为了使各项指标具有可比性，需要对搜集到的各指标数据进行无量纲化处理。这里可以采取离差标准化，它是对原始数据的线性变换，将结果值映射到[0, 1]之间，转换函数如下[3]：

$$X^* = \frac{X - \min}{\max - \min}$$

得到标准化后的数据如下：

$$data=\begin{pmatrix}1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00\\ 1.00 & 1.00 & 0.14 & 1.00 & 1.00 & 0.02\\ 0.19 & 0.31 & 1.00 & 0.00 & 0.01 & 0.08\\ 0.02 & 0.08 & 0.20 & 0.50 & 0.01 & 0.03\\ 0.70 & 0.02 & 0.70 & 0.60 & 0.22 & 0.11\\ 0.91 & 0.08 & 0.03 & 0.70 & 0.21 & 0.01\\ 0.04 & 0.02 & 0.17 & 0.8 & 0.01 & 1.00\end{pmatrix}$$

具体计算方法：

1) 选取参考数列

$$x_0=\left\{x_0(k)\mid k=1,2,\cdots,n\right\}=\left(x_0(1),x_0(2),\cdots,x_0(n)\right)$$

设有 6 个比较数列

$$x_i=\left\{x_i(k)\mid k=1,2,\cdots,n\right\}=\left(x_i(1),x_i(2),\cdots,x_i(n)\right),i=1,2,3,4,5,6$$

这里 $n=6$ 表示主要国家的数量。我们选取各指标所期望的最优值作为参考序列。医院人数、核酸检测数量、呼吸机数量、封城时间相对来说较长，效果较好；感染率、人口密度则相反，则参考序列可取为：

$$x_0=\left\{33009,0.608,2.8,10000,450,10\right\}$$

2) 数据归一化处理：

按照离差标准化方法进行处理。

3) 计算灰色关联度：

通过灰色关联系数公式

$$\xi_i(k)=\frac{\min\limits_s\min\limits_t\left|x_0(t)-x_s(t)\right|+\rho\max\limits_s\max\limits_t\left|x_0(t)-x_s(t)\right|}{\left|x_0(k)-x_i(k)\right|+\rho\max\limits_s\max\limits_t\left|x_0(t)-x_s(t)\right|}$$

计算出灰色关联度 r。

$$r=\left(0.87,0.78,0.65,0.63,0.58,0.44\right)$$

3.2. 模糊综合评价

1) 选取隶属度函数：根据灰色关联度，将相关因素按最大值、最小值划分区间，建立隶属度函数[4]。

$$\mu_A=\begin{cases}1, & 0.44\le x\le 0.58\\ \left(\dfrac{b-a}{b-a}\right)^x, & 0.58<x\le 0.78\\ 0, & 0.78<x\le 0.87\end{cases}$$

2) 根据隶属函数计算出 6 个国家对应的不同隶属度，见表 1。

Table 1. Membership table
表 1. 隶属度关系表

国家	中国	澳大利亚	巴西	西班牙	印度	美国
医院数量	0.802	0.789	0.891	0.678	0.781	0.783
呼吸机数量	0.761	0.766	0.858	0.663	0.774	0.766
人口密度	0.557	0.629	0.579	0.568	0.565	0.562

Continued

封城时间	0.810	0.8	0.577	0.78	0.804	0.607
感染率	0.936	0.885	0.675	0.731	0.92	0.632
核酸检测数	0.833	0.787	0.663	0.579	0.621	0.732

3) 确定模糊关系矩阵 R

$$R=\begin{pmatrix}0.802 & 0.789 & 0.891 & 0.678 & 0.781 & 0.743\\ 0.761 & 0.766 & 0.858 & 0.663 & 0.774 & 0.766\\ 0.557 & 0.629 & 0.579 & 0.568 & 0.565 & 0.562\\ 0.810 & 0.8 & 0.577 & 0.78 & 0.804 & 0.607\\ 0.936 & 0.885 & 0.675 & 0.731 & 0.92 & 0.632\\ 0.833 & 0.787 & 0.663 & 0.579 & 0.621 & 0.732\end{pmatrix}$$

4) 根据相关度，各影响因素在决策中占的权重为

$$A=(0.22,0.20,0.17,0.16,0.13,0.12)$$

5) 计算各个国家综合得分

$$B=AL=(0.7746,0.7712,0.7257,0.6676,0.7454,0.6881)$$

由此可知：中国疫情管控效果最好，澳大利亚次之，美国及西班牙管控效果较差。

4. SEIR 模型

SEIR 是常见的一种描述传染病传播的模型，SEIR 模型在 SIR 的基础上增加了潜伏者，这更加符合新型冠状病毒的实际情况，易感人群在发病前期会经历潜伏期，一段时间之后才表现出症状[4]。因此，我们选用 SEIR 模型更为准确。其基本假设是将环境中的所有人群分为了四类：

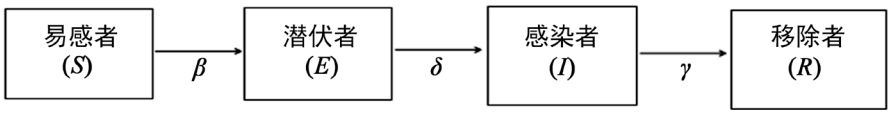
1) 易感者(健康者)：指未得病，但与感染者接触后有被感染概率的人；

2) 潜伏者：指已被感染，但还没表现症状的人；

3) 感染者(病人)：指感染上病毒的人，他可以传播给易感人群；

4) 移除者(病愈者)：被病毒感染之后，病愈或死亡的人，这部分人不再参与感染和被感染的过程。

SEIR 模型在以上三类人群中存在三个转换关系[5]：



β：易感者与感染者接触时被传染的概率，反映了疾病的传播强度，β 越大，易感人群和感染人员接触后被传染的可能性越大。δ 反映了潜伏者转换为感染者的概率。γ：感染人群以固定的平均速率恢复或死亡，这个概率称为恢复系数，它取决于感染的平均时间。

4.1. 动力学模拟疫情过程

基于上述介绍，四类人群数量随时间的动态变化规则可用以下常微分方程组来表示[5] [6]：

$$\frac{\mathrm{d}S}{\mathrm{d}t}=-r\delta\frac{SI}{N},\ \frac{\mathrm{d}E}{\mathrm{d}t}=r\delta SI/N-\beta E,\ \frac{\mathrm{d}I}{\mathrm{d}t}=\beta E-\gamma I,\ \frac{\mathrm{d}R}{\mathrm{d}t}=\gamma I.$$

4.2. 优化算法进行参数辨识

通过上述介绍，我们知道 SEIR 模型实际就是运用动力学模型，对疫情发展过程进行模拟，采用感染系数、恢复系数来刻画疾病传染和治愈的过程。此过程最重要、最关键的一点就是获取精确的模型参数，建立精确的模型，进而达到较好的预测效果。我们的主要任务就是确定以下最优参数[7]：

1) 感染系数 β、转化系数 δ、恢复系数 γ；

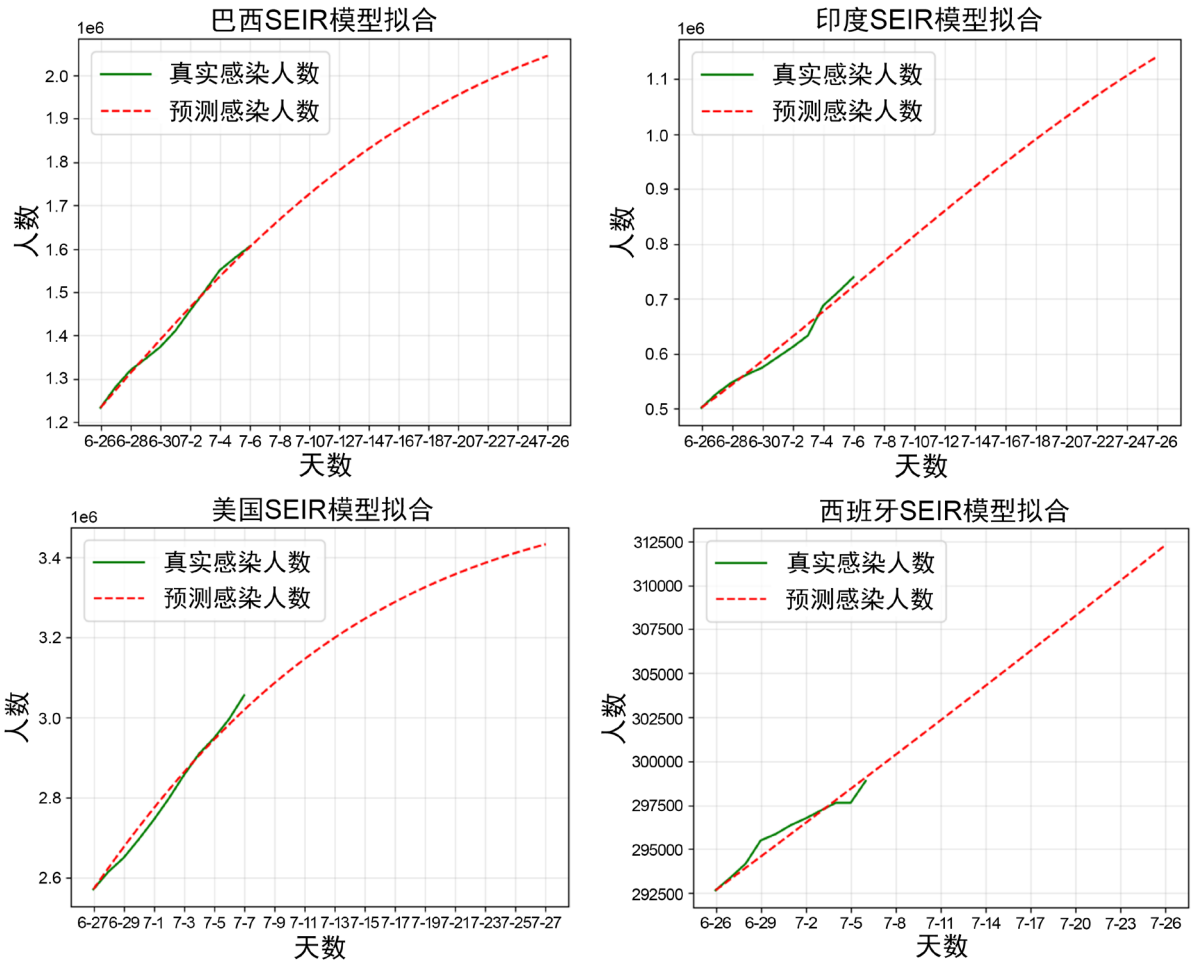
2) 易感人群初值 $S(0)$、潜伏人群初值 $E(0)$、感染人群初值 $I(0)$、移除人群初值 $R(0)$。

其中、感染人群初值 $I(0)$、移除人群初值 $R(0)$，可由当日的确诊人数、治愈人数、死亡人数得到；对于恢复系数 γ，常取恢复天数的倒数，由于随着疫情发展，各国采取更为有效的防控手段，所以我们使用 6 月 17 日到 6 月 27 日的数据计算 γ_1，再使用 6 月 28 日到 7 月 7 日算出 γ_2，最后取其平均值作为恢复系数 γ，因此，我们的主要任务变为了确定 $S(0)$、$E(0)$、β、δ。

最优参数值的确定可以采用最小二乘法：先给定一个大致区间，使用穷举法，使得真实值与所得值之差的平方和最小的参数值即为最优参数。

4.3. SEIR 模型预测

通过最小二乘法确定最优参数后，对 6 个国家 6 月 27 日~7 月 7 日的确诊人数进行拟合和预测，预测效果如下图 4 所示：



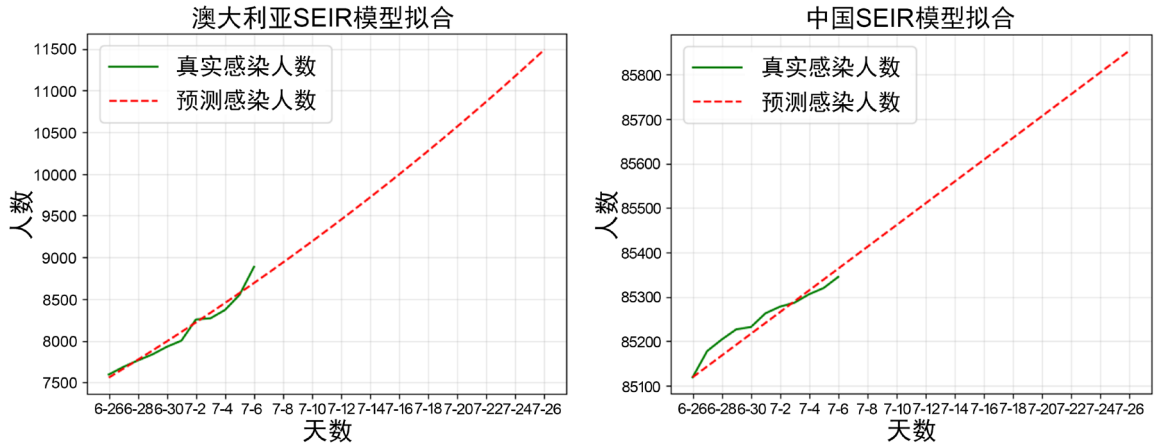


Figure 4. Predictions of confirmed patients in major countries
图 4. 主要国家确诊患者预测

由图 4 可以得到各国的预计感染患者数量，可以根据各国人口数，确定确诊患者数量在本国的比例，以此评估各国新冠肺炎的严重程度，见表 2。

Table 2. Estimated proportion of people diagnosed on July 24
表 2. 预估确诊患者比例

国家	预估确诊患者数量	本国人口(百万)	预估确诊患者比例(%)
美国	3,200,000	3.27×10^4	0.891
巴西	1,800,000	2.09×10^4	0.858
印度	1,100,000	13.53×10^4	0.579
澳大利亚	11,280	2499.24	0.577
西班牙	310,375	4672.37	0.675
中国	85,800	13.95×10^4	0.663

从预估确诊患者数量在本国占比来看，美国、巴西的严重程度为第一梯队，西班牙为第二梯队，印度、澳大利亚为第三梯队，中国最低。

4.4. 模型检验

检验回归模型的效果可以采用 F 检验。它验证的是偏回归系数是否不全为 0，直接从回归效果检验回归方程的显著性。

1) 构造 F 统计量

我们知道平方和分解式： $SST = SSE + SSR$

其中，真实值和均值差的平方和为 $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$，预测值和均值差的平方和为 $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$，真实值和预测值差的平方和为 $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$，不难看出，$SSE$、$SSR$ 都会随着模型的变化而变化，可构造统计量：

$$F = \frac{SSR/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$$

2) 计算 F 值

该模型得到的 $F=174.637$，理论的 $F=2.502$，由于计算的 F 值远远大于理论 F 值，所以拒绝原假设，说明回归方程是显著的。

5. 总结

抗击新型冠状病毒疫情已经成为了世界各个国家的共同责任，了解新型冠状病毒的传播特点有助于抗击疫情。为更好地了解世界各个国家的疫情情况，本文建立了聚类分析、模糊综合评价和 SEIR 预测模型。利用统计理论、灰色关联度分析和最小二乘法分别对模型进行数据处理、求解、预测和检验。在一般的 SIR 模型基础上，我们考虑到了潜伏者，建立了 SEIR 模型，预测效果较为准确。然而新型冠状病毒的发展情况，不仅仅与已选取的指标有关，还受到一些不确定性因素的影响，例如境外输入病例的出现。因此模型还有待优化和改进。

参考文献

[1] 约翰霍普金斯大学. 全球新冠疫情统计数据[Z/OL]. https://coronavirus.jhu.edu/map.html, 2020-07-05.

[2] 何晓群. 多元统计分析[M]. 北京: 中国人民大学出版社, 2015.

[3] 司守奎, 孙玺菁. 数学建模算法与应用[M]. 北京: 国防工业出版社, 2011.

[4] 徐宝春. 基于 SIR 模型的 SARS 传染病研究[D]: [硕士学位论文]. 青岛: 山东大学, 2019.

[5] 马知恩, 周义仓, 王稳地. 传染病动力学的数学建模与研究[M]. 北京: 科学出版社, 2004.

[6] 韩中庚, 马晓军, 胡宗云. 数学建模竞赛: 获奖论文精选与点评(第二卷) [M]. 北京: 科学出版社, 2013.

[7] 王志心, 刘治, 刘兆军. 基于机器学习的新型冠状病毒(COVID-19)疫情分析及预测[J]. 生物医学工程研究, 2020, 39(1): 1-5.