桂林电子科技大学
第十八届大学生数学建模竞赛

承　诺　书

我们仔细阅读了桂林电子科技大学第十八届大学生数学建模竞赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

参赛队员信息：

序号	姓名	学号	专业	电话号码
队长	蔡响	2000301708	信息安全	13328372882
组员	赵浚帆	2000300236	网络空间安全	18007710093
组员	陈乐	2000300309	计算机科学与技术	13687898384

基于 SEIR、ARIMA 的多模型预测新冠肺炎疫情

摘要

新型冠状病毒，从 2019 年年底至今已在全世界范围内传播，到目前为止已有 5 种“关切的变异株”。其导致全球超过 600 万人死亡，是自 1918 年以来最大的全球健康危机。要对国家或地区进行疫情情况分析和预测，需要综合考虑其防疫政策和整体医疗水平及其人口基数大小。本文旨在建立模型满足对疫情的情况分析和未来发展的预测。

针对问题一：分析提供的数据，选取部分有代表性的国家和中国的数据并选择性剔除部分无用多余的数据，结合当国应对政策进行比较，进行曲线模拟和占比图求解。通过对比不同的国家应对新冠疫情的政策及其疫情的发展情况，建立了多个数据指标，抽取处理过的数据进行拟合求解其数值，用于后两问中的模型建立和求解，根据不同政策的国家进行适合其国情的数学模型搭建。

针对问题二：按照不同的防控疫情策略，在考虑对疫情的防控方面许多国家依据自身情况采取了不同的应对方案，以此作为分类标准， 选取三个具有代表性的国家进行分类模型的搭建。根据传统的传染病模型 SIR，综合考虑到存在潜伏者情况的 SEIR 模型，存在变异情况使得康复者能重新转化为易感人群的 SEIRS 模型。这三个模型合理的对应了英中美三国的实际情况。在第一问确定的数据指标的基础上，结合马尔科夫链，在传染病动力学基础上进行设立微分方程进行模拟。据现有数据进行拟合并对未来预测。

针对问题三：对于第二问设计的传统传染病模型，其对感染率是估算的，在后续的其他的因素影响下，是无法根据事实进行动态调整的，无法满足中国不采取动态请零政策情况下的情况模拟和预测。在不采取动态清零的政策下，中国的感染率和日均接触率以及治愈率等指标都会有所改变。使用具有时间序列预测性的 ARIMA 模型更合理。结合美国的特殊性，抽取美国的疫情数据后进行 ARIMA 模型后进行训练，获得的数据和现有的数据进行拟合。求解特征值自回归项和移动平均项数，达标后预测未来一段时间的疫情情况。另抽取英国的疫情数据后放入 SIR 模型求解日均接触率和日均治愈率，拟合达标后预测未来一段时间的疫情情况。完成两种不同模型的训练后。依据两国之间的人口系数和医疗水平进行换算，将中国数据放入 ARIMA、SIR 模型中进行预测，得到不采取“动态清零”政策而效仿西方国家政策后的中国疫情发展情况。

关键词：SIR 系列传染病预测模型； ARIMA 预测模型；马尔科夫链；时间序列预测

一、问题重述

1.1 问题的背景

新型冠状病毒（SARS-CoV-2），从 2019 年年底至今已在全世界范围内传播开来，到目前为止已有 5 种“关切的变异株”分别为阿尔法（Alpha）、贝塔（Beta）、伽玛（Gamma）、德尔塔（Delta）和和奥密克戎（Omicron）。病毒引发的新冠肺炎（COVID-19）已导致全球超过 600 万人死亡，是自 1918 年以来最大的全球健康危机。

从当前的情况来看，新冠病毒是一种具有易变异特性，高潜伏风险，高传播风险，高危害风险的新型冠状病毒。根据近两年的相关情况，新冠病毒借助其感染前期的隐蔽性和不易察觉的症状迷惑人们，对其产生错误的判断，借助其传播方式的快速、易感的特点大规模的在人群中传播,随后借助其感染后期的搞致命性使部分感染者迅速丧失生命，尽管最近的时间已经研发了对于新冠的疫苗和特效药，但距离临床实验和真正上市还有相当一段时间，所以我们对新冠的重视程度依旧要保持一个很高的程度。对疫情采取正确有效的措施的前提是对新冠病毒得到传染模式有--定的了解，而借助数据分析是一种有科学依据的方式，而庞杂的数据往往让人无法直观抽取信息要素，无法客观给出定性分析，借助数学工具对数据进行可视化操作则成为分析数据的首选方式。在对病毒传染模式有一定了解的基础上,如何根据已有数据对病毒感染规模进行准确有依据的预测应成为考虑的重点,因为借助从数学模型的角度出发所做出的预测是科学高效的。在建立的模型的基础上如何根据实际情况来对模型进行定性的修正，使模型的预测更准确更贴合实际也是一个需要考虑的问题。

1.2 题目所给信息及参数

本比赛在 2020 年就已经关注了该传染病在我国的传播情况并根据相关数据进行了分析和探讨。截至北京时间 2022 年 6 月 6 日，根据 Worldometer 实时统计数据，全球累计确诊新冠肺炎病例约 5.3 亿例，累计死亡病例约 632 万例。现我们从美国约翰斯·霍普金斯大学（Johns Hopkins University）实时统计数据网站得到三个数据文件（包含全球新冠肺炎确诊病例、死亡病例和治愈病例）。

附件 1：全球新冠肺炎确诊病例数据（截止 2022 年 6 月 4 日）；附件 2：全球新冠肺炎死亡病例数据（截止 2022 年 6 月 4 日）；附件 3：全球新冠肺炎治愈病例数据（截止 2022 年 6 月 4 日）；附件 4：数据指标说明。

1.3 所需解决的问题

1. 根据附件所提供的全球的疫情数据对全球新冠疫情情况进行描述统计，提取整合有效的数据信息，包括但不限于某国家的病情情况、感染率、死亡率、

康复率、死亡率、易感人群等信息。

2. 依据各国或地区病情的不同变化特点和各国的相关政策进行分析，利用适合各国国情的数学建模方法建立基于问题一的数据分析的分类模型，阐释各类别数学模型特征。

3. 通过比较我国与其他国家疫情情况拜年话的不同，解读不同国家的政策后，利用合理的数学模型分析如果我国不采取“动态清零”的总方针，那么我国的疫情会有何变化，并阐述问题三的模型建立过程和适当的求解分析过程。

二、问题分析

针对不同的国家，地区和相对应的医疗水平进行对应的数据指标分析。主要分析感染率，病人接触率，治愈率，以及传染期接触数。模型构建还需要考虑到新冠肺炎的无症状感染者这一特殊的情况，根据这些指标进行相轨线分析，合理的进行疫情的分析和未来疫情走向以及各地区、国家的防疫政策研究。

2.1 问题一的分析

附件 1 中给出了全球新冠疫情情况的数据，包含各地区每日患病人数。需要解决的就是进行数据的筛选，统计整合出有用的数据来进行模型拟合分析，可视化呈现国家和地区的病情。获取感染率，治愈率，日均接触率等指标，并且求解出对应的数值，利用曲线图和占比图将数据呈现。

2.2 问题二的分析

在问题一数据分析的基础上，选取部分的国家进行分析。选取合适的数学模型进行分析，已有传染病的 SIR 模型，将合适国家的数据进行模拟计算，分析得其特点。新冠传染的过程中密切跟随各人群的变化走向，根据对应的比率动态变化。引用马可夫链思想进行状态关联，给出对应的状态率微分方程，分析该模型的结果阈值，结对应国家的政策进行合理分析。考虑到新冠肺炎的特殊，存在无症状感染者的情况，后续改进为 SEIR(存在无症状感染者)和 SEIRS(治愈者可重新患病)两个更为符合实际的模型，新冠经历多次变异，须考虑到合理的治愈后可重新感染的情况，合理分析选取特征数据的各国的疫情特征。

2.3 问题三的分析

我国与其他国家不同的变化情况最主要的原因来自于我们疫情期间较为严厉的政策，导致在一段时间内，传染率、治愈复发和传染期接触情况基本为 0。治愈率取决于国家的医疗水准，不采取“动态清零”政策的情况会在现有的数据上会存在偏差。合理取一水平较近的国家进行拟合。分析不采取“动态清零的国

家”模拟出对应的感染率和传染期接触情况，根据中国人口各基数的不同数据进行整理模拟，进而求解我国不采取该政策的情况下疫情发展的趋势。

三、模型假设

1.新冠肺炎的传播途径仅为现已知的飞沫传播和接触传播；
2.新冠肺炎的传播只存在于人传人之间；
3.不考虑年龄因素、当地的环境气候因素以及其他存在抗体的人群对该冠状病毒的影响；
4.假定传播的主题在城镇，人口总量大体在 9 亿人保持不变；
5.不考虑境外输入风险；
6.接触即可判定为有感染风险，属于密集人群。

四、符号说明

符号	意义	单位
N	总人口数	人
E	潜伏者、无症状感染者	人
D	死亡人数	人
I	感染人数	人
S	易感人数	人
R	康复人数	人
r_1	感染者接触的人数	人
b_1	传染新冠病毒的概率	
a	潜伏者变为感染者的概率	
r_2	潜伏者接触的人数	人
b_2	潜伏者、无症状感染者感染给正常人的概率	

g_1	感染者康复概率
d	感染者日致死率
g_2	潜伏者、无症状感染者康复后转易感者概率

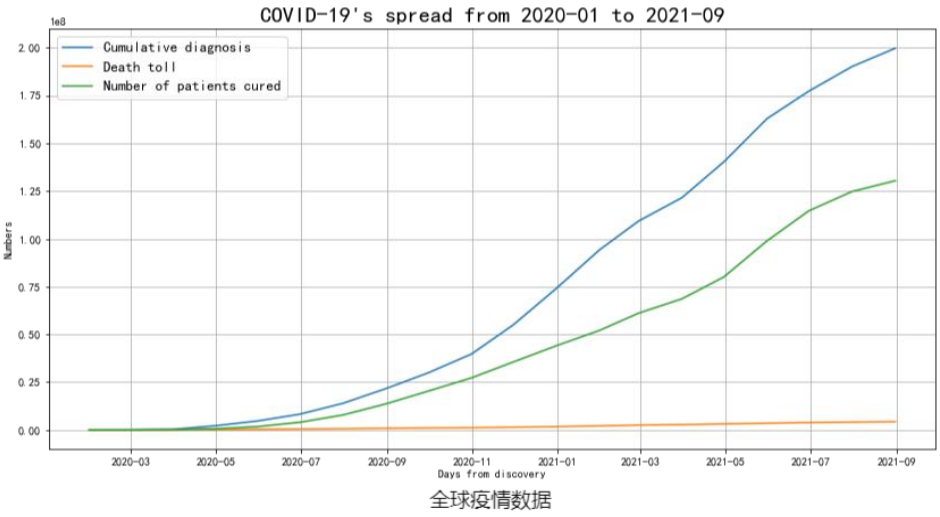
五、模型建立与求解

5.1 问题一模型的建立与求解

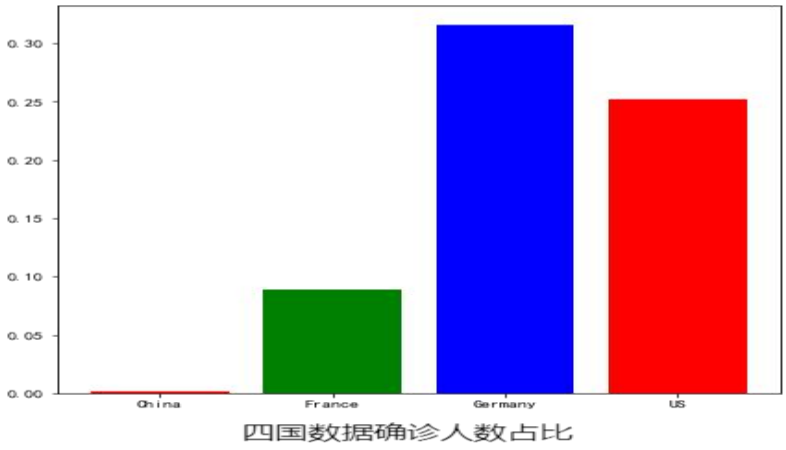
问题一聚焦对于附件提供的数据进行描述统计。附件一提供了对新冠肺炎感染人数从2020年1月22号至2022年6月4日的每日感染人数的统计。因为附件一包含的国家过多，但是大部分国家采取的政策和方针是相似的.现分成实行两种不同政策的国家。第一种：美国大部分的州采取的是放任疫情发展，允许大规模人群聚集，尽管人群自新冠肺炎爆发后有戴口罩的习惯，但美国政府的放任仍旧对疫情的传播起到推动作用；相似的还有英国的群体免疫政策，在疫情爆发初期各个国家没有疫苗的产生，所以采用群体免疫的政策来确保整个国家的安全。虽然2022年有疫苗的普及，但对数据的分析预测主要集中在疫情爆发的初始阶段，所以英国的传播模型也可以类似与美国，同理较多西方国家都是采取此类政策，故采取相似的政策的国家可以一并采取一种模型进行数据分析。第二种不同的政策是中国的动态清零政策。中国的疫情传播因为人人有安危意识，戴口罩，居家隔离，定期做核酸检测的原因，疫情在中国的发展只能限于还没有被发现的时间，遏制疫情的时间也会比正常的情况短上很多，在被发现后动态清零政策会让疫情的爆发较快来到拐点，所以对于类似中国的数据分析和模型建立是另一种情况。综上所述，在此表中读取文件中几个国家的数据建立散点图，进行分析描述。

下图是全球疫情和抽取的四个国家的新冠肺炎传播感染人群数量的对比情况。从对比我们可以看出全球的新冠肺炎疫情仍旧在肆虐全球，虽然中国的占比是最少的但是由于经贸等各种原因，中国的国门是不可能关上的，所以我们对新冠的防控

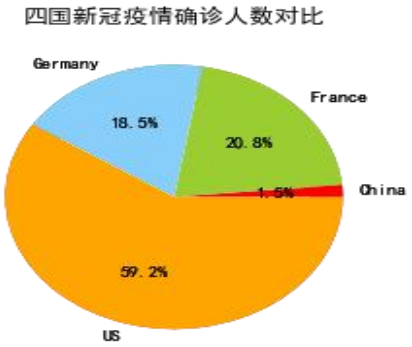
不能松懈，对于新冠的预测和提前准备是非常有必要的。



图一：全球疫情曲线图



图二：四国确诊人数占比



图三：四国新冠人数对比

下图是中美的确诊、死亡和恢复的曲线对比图。根据数据分析统计，我们可以直观的发现美国的确诊人数远远大过中国的确诊人数，究其原因是因为中国和美国的政策不同，人们对生命财产安全的看法有本质上的区别才导致了对于现在的情况。这对于接下来的节目和分析提供了非常明确的不同方向。

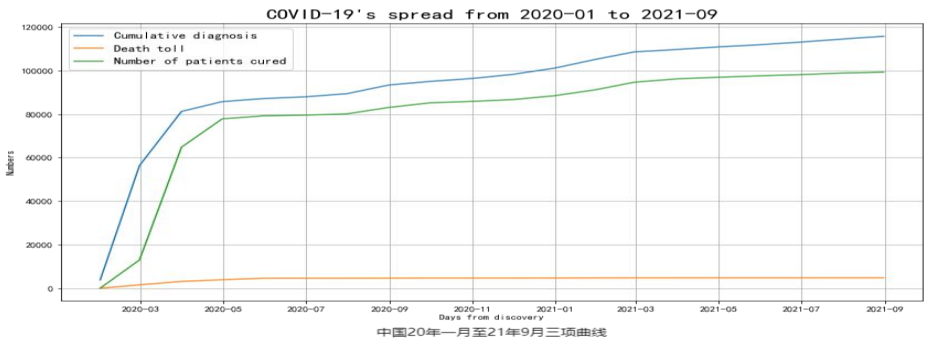
图四：中美疫情对比

下图是中国2020年1月至2021年8月的数据，蓝色曲线代表确诊感染的累计数，绿色代表累计康复人数，橙色代表累计死亡人数。从下面的图表中我们可以分析得到绝大部分的感染者都能被治愈，少部分的会病情严重甚至危机生命。对于该曲线前4个月，因为对于新冠病毒的陌生和缺少前期的防范意识导致感染人数大幅超越恢复人数，所以依据曲线我们推算得出指标：

① g（感染者康复率）为 0.1；

② d（感染者日致死率）为 0.05% ；

在后期随着各种医疗卫生情况的好转后期确诊和康复的比率维持相对不变，暂不做假设。



图五：中国20年1月至21年9月的确诊人数、治愈人数、死亡人数曲线图

下图所示是新冠肺炎在中国传播时的折线图，比较的是疑似病例的折线和确诊病例的折线。对于如下的折现折线部分进行分析得到疑似病例远高于感染者，原因在于新冠肺炎具有极强的传播性，在感染者所到的场所都是高危地区，根据现有的经验符合我们的分析；同时在疫情传播的初期人们的防范意识不足，人群接触的情况与平时无差别。根据资料和折线图我们得到如下指标

③ r1（感染者接触人数）= 23 ；

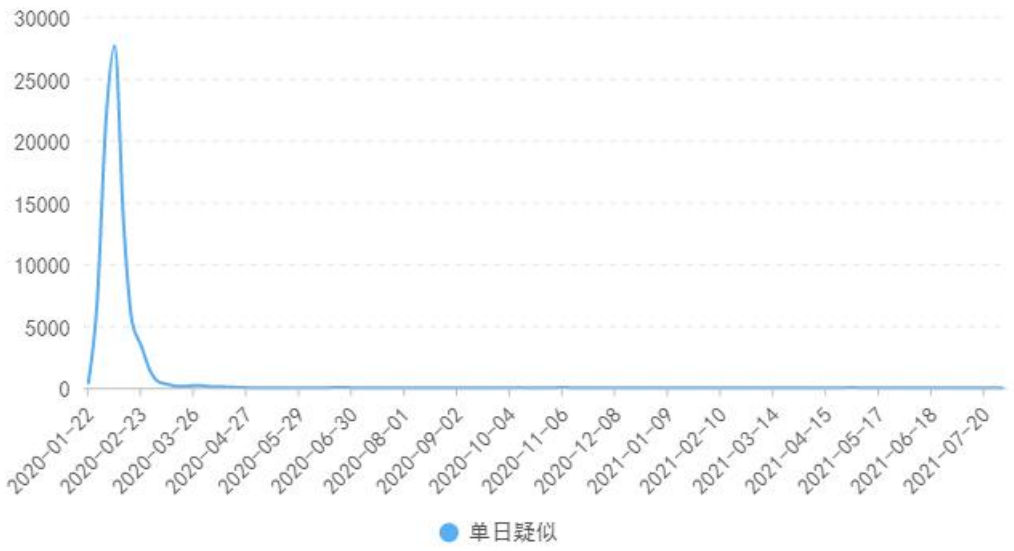
④ b1（传染新冠病毒的概率）= 0.037；

⑤a（潜伏者变为感染者的概率）= 0.127；

对于潜伏者来说，因为症状表现的不明显所以同样对少去不去公共场合缺少意识，接触人数依旧可以看成与日常生活无异。

⑤ r2（潜伏者接触的人数）= 52；

⑥ b2（潜伏者、无症状感染者感染给正常人的概率）= 0.0263 ；



图六：中国20年1月22日至21年7月20日的单日新增疑似曲线

5.2 问题二的模型建立和分类

本节针对问题一分析得出的数据指标，选取部分特征不同的国家进行疫情情况分析，进行模型的分类和搭建，同时对其模型进行数据模拟。进行数据预测，结合真实情况进行模型的分析评价。分析各个国家针对疫情的措施以及真实的情况，现在的疫情防范主流有中国这种严厉的“遇见一起，扑灭一起”的措施，还有诸如西方各国的“群体免疫”策略，以及医疗卫生情况良好，防范措施适中，但是不够严谨的部分国

家。我们从中抽取合适的国家数据进行模型建立。

5.2.1 模型的分类

中国针对疫情的政策采取的是“动态清零”的政策，遵循“外防输入、内防反弹”防控策略的前提下，当出现本土新冠肺炎病例时，采取有效的综合性防控措施，“发现一起、扑灭一起”，快速切断疫情传播链，使每一起疫情及时终止，感染者“清零”，实现以最小成本取得最大成效的一种防控目标。中国防范疫情的措施严厉程度高，在人口密集处的防范相对严格，故该情况下中国的日均接触率和感染率会相对低，但同时严厉的措施仍会存在潜伏者，检测不出来，则误认为健康的状态，故此，我们将采用传染病 SEIR 模型对中国数据进行分析。

西方部分国家支持“群体免疫”的政策，是指人群对传染的抵抗力，群体免疫水平高，表示群体中对传染具有抵抗力的动物百分比高。因为，疾病发生流行的可能性不仅取决于动物群体中有抵抗力的个体数，而且与动物群体中个体间接触的频率有关。如果群体中有 70%—80%的动物有抵抗力，就不会发生大规模的爆发流行。其中英国首席科学顾问帕特里克·瓦兰斯称，英国大约需要感染 60%的人来获得群体免疫力。从中合理分析，此情况下，英国人群对疫情防范措施疏忽，可以假设无防护能力，即由确诊者接触，则必定感染的情况，可以认为其无潜伏者存在，在这一问中先不考虑新冠肺炎的变异，后续会给出更加合理的考虑变异的模型。则据英国情况可以采用 SIR 传染病模型来进行数据分析。

对于美国这种，医疗卫生条件较佳，但防范措施不够严厉的情况，日均接触率不如放肆政策的国家高，感染率适中，但是仍会存在潜伏者的情况，即无症状感染者。同时需要考虑到痊愈后仍可能重新变回乙肝人群的情况，采取传染病 SEIRS 模型进行数据分析。

5.2.2 SIR 模型建立

SIR 模型是一种国际通用的传统传染病模型。 该模型是研究传染病的传播速度、感染人数以及发展 趋势等问题的一种非线性动力学模型，对传染病有效控制和预防具有指导的作用 。人群分为 S（易感人群）、I（已确诊者）、R（治愈者）。假设该国家的总人口为 N，则可得：

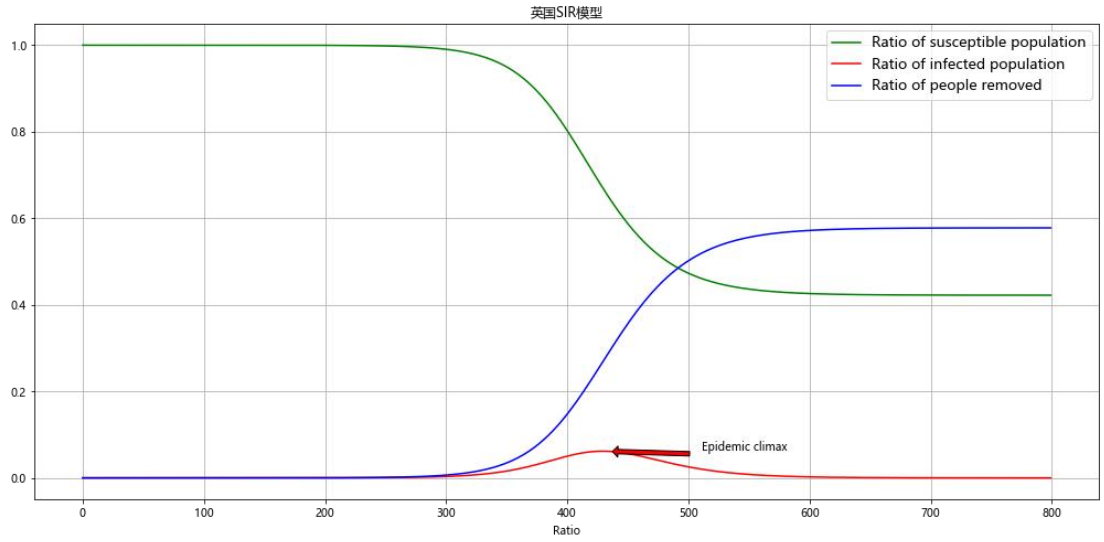
$$N = S + I + R \tag{1}$$

由下图描述疫情情况转移图：

建立如下的微分方程组

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{r_1 b_1 SI}{N} & (2) \\ \dfrac{dI}{dt} = \dfrac{r_1 b_1 SI}{N} - g_1 I & (3) \\ \dfrac{dR}{dt} = g_1 I & (4) \end{cases}$$

模拟后得出如下的状态曲线图：



图七：英国 SIR 状态曲线图

分析方程当 $\frac{dS}{dt} \leqslant 0$ 时，S＞0，则 S 的无穷值存在，$\frac{dR}{dt} \geqslant 0$ 时，R≤1，R 的无穷值也存在，同理 I 的无穷值也是存在的。分析阈值 $\sigma = r_1 / g_1$，其中当 S 小于 $1/\sigma$，I 是单调减小至 0 。S 会单调减小到 S 的无穷值。我们认为，当确诊这 I 在一段时间增长，才认为传染病在蔓延，则可得出 $1/\sigma$ 就是我们要找的边界分析值。S＞$1/\sigma$ 则肺炎开始传播，增大阈值 $1/\sigma$，使得 S≤1，肺炎疫情就可以控制住。

我们判断，该国家的卫生医疗水平越高，日接触率越低，日治愈率越大，就可以得到 $1/\sigma$ 越大，所以上述指标的改动可以有利的影响新冠疫情在该国家的蔓

延。

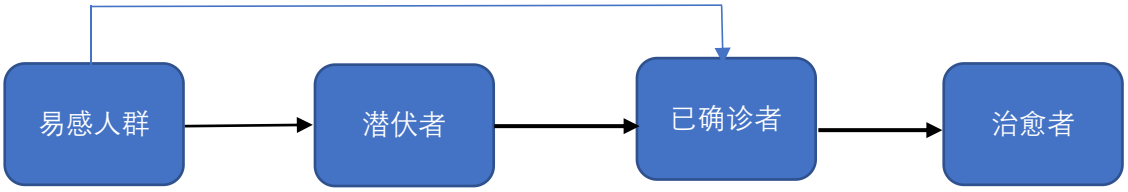
阻止疫情可以通过增大$1/\sigma$，另外也可以减少 S 的值，即降低易感人群的在总人口中的比例。忽略初始的确诊者，假设移出的治愈者为r_0，则有$S=1-r_0$，为了满足$S\le 1/\sigma$这一条件，是可以通过群体免疫的策略来调整治愈者的移出比例，满足$S\le 1/\sigma$这个条件的条件下，是可以做到制止传染病的蔓延的。

但这种群体免疫方法有一定的局限性，需要病后免疫者合理的分布在全体人口中，实际上这是很难做到的。

5.2.2 SEIR 模型建立

SEIR 模型和 SIR 模型的最大区别是，考虑到了存在潜伏者，即我们在新冠期间称为无症状感染者的人群，他们在被确诊者接触后 14 天内会病发，同时他们的活动轨迹中所接触到的易感者也会有一定几率被感染。结合中国疫情防范的措施，选取 SEIR 模型较为合理。

SEIR 模型中人群分为 S（易感人群）、E（无症状感染者）、I（已确诊者）、R（治愈者）。假设该国家的总人口为 N。则可以表示为$N=S+Z+I+R$。传染流程如下图所示：



由于是连续时间马尔可夫过程，可以构建关于时间 t 的动力方程：

$$\begin{cases} S_{t+1}=S_t-(r_1b_1S_tI_t)/N-(r_2b_2S_tE_t)/N & (5)\\ E_{t+1}=E_t+(r_1b_1S_tI_t)/N+(r_2b_2S_tE_t)/N-aE_t & (6)\\ I_{t+1}=I_t+aE_t-g_1I_t & (7)\\ R_{t+1}=R_t+g_1I_t & (8) \end{cases}$$

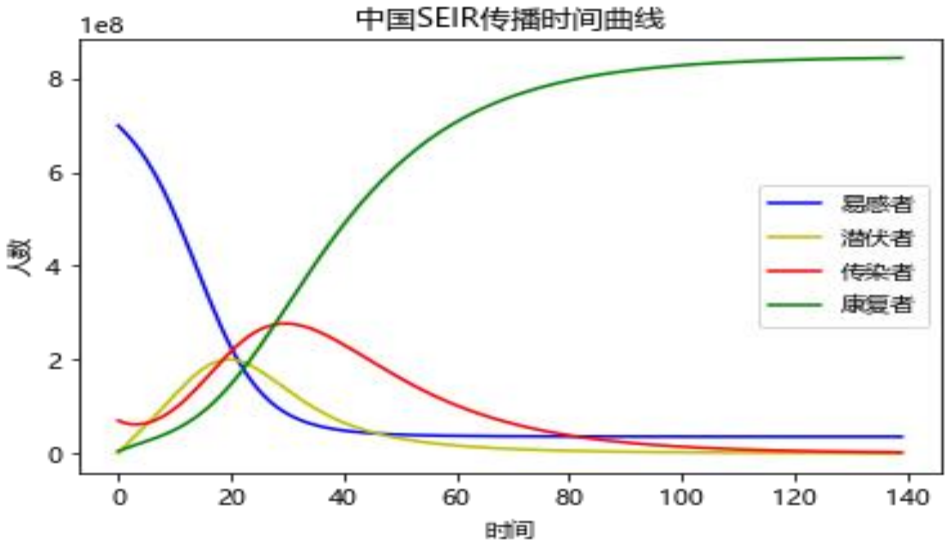
构建微分方程组：

$$\begin{cases} \dfrac{dS}{dt}=-[(r_1b_1S_tI_t/N)+(\mathrm{r_2b_2E_tS_t}/N)] & (9)\\ \dfrac{dR}{dt}=aE_t-g_1I_t & (10) \end{cases}$$

$$\frac{dE}{dt} = (r_1 b_1 S_t I_t) / N + (r_2 b_2 S_t E_t) / N - a E_t \quad (11)$$

$$\frac{dR}{dt} = g_1 I_t \quad (12)$$

进行模型建立和数据代入分析后得出如下结果图表：



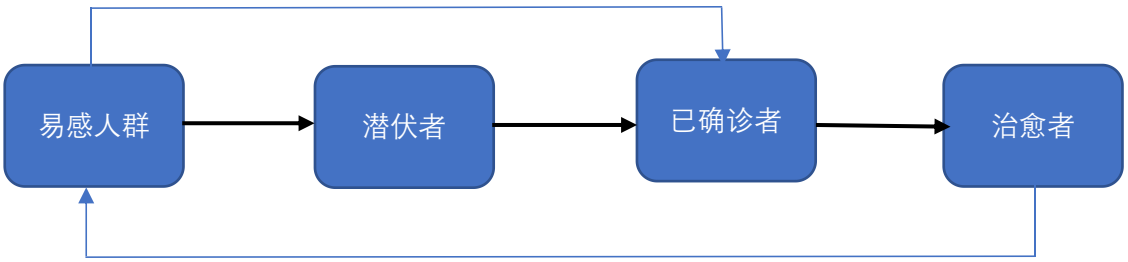
图八：中国 SEIR 传播曲线

可以发现大致和我们国家的疫情情况吻合，评估疫情情况依赖于其 I 和 E 之间的接触人群数，以及相应的感染率。依然采用 $\sigma = r_1 / g_1$ 作为指标，来进行比对。它由整体的医疗水平、治愈率接触率来决定。假定当该值在一定时间内达到稳定，疫情会趋向于稳定。设初始的潜伏者为 0，分析 $1/\sigma = \frac{g_1}{(b_1 r_1 + b_2 r_2)\,d}$ 我们可以得出，$1/\sigma$ 是跟 S 相关的，控制 $1/\sigma$，使其增大，是可以控制疫情的。采取的措施一即为减少 E 和 I 人群对易感人群的接触率，通过减少人们的出行，以及对传播途径进行阻断传播的方式，来达到降低接触率。措施二，也可以提高治愈率和降低死亡率，这涉及整个国家的医疗水平和人民身体健康程度，在短时间内是达到效果的，所以措施一是较为有效的，但我们通过数据模拟对比发现，措施二对整体阈值的变化率影响更为明显，所以措施一实用，措施二高效。但二者都有着相互的联系，我们要做的就是尽量延缓 E 和 I 的高峰期到来，使其尽可能往后移，为我们的医疗水平等硬实力的提升争取时间。同时就是实行措施一，严厉管控疫情，减少接触率，减少感染率，为前者争取时间。

总结该类模型的特点，符合中国的疫情特征，即我们采取了严厉的管控措施，但同时考虑到了新冠的潜伏者存在，采用 SEIR 模型，在 S 和 I 之间，还有 S 到 E 的这一条传播途径，适用于管控策略严厉，人口基数大，存在人与人接触率较高，但通过切断传播途径和降低接触率，等待医疗水平逐渐提升（新冠疫苗的开发）后能够出现疫情好转的情况。

5.2.3 SEIRS 模型建立

SEIRS 模型是一种国际通用的传统传染病模型。该模型是研究传染病的传播速度、感染人数以及发展 趋势等问题的一种非线性动力学模型，对传染病有效控制和预防具有指导的作用。建立在 SEIR 的基础上进行的修改，新冠病毒历经多次变异，会使得原本已经获得了抗体的康复者，重新变为新的易感人群，成为潜伏者和确诊人群的传播对象。结合美国的疫情情况和相应的医疗水平和人民身体素质。得出结论：医疗水平高，人民身体素质良好，采取了一定的疫情防控措施（阻断传播，减少接触，自我隔离等），但由于其文化经济的影响，对外文化经济交流对其影响不可忽略，考虑该情况，则美国会存在外来感染的新型变异的新冠病毒，使得已经获得了抗体的康复者重新被感染，采用 SEIRS 建模拟合分析。传染过程如下：



总体人群分类和 SEIR 类似，分为 S（易感人群）、E（无症状感染者）、I（已确诊者）、R（治愈者）。总人口 $N=S+E+R+I$，且总人口不变（不考虑境外输入，仅仅考虑对外交流过程存在的感染变异情况）。引入新的变量 g_2（潜伏者、无症状感染者康复后转易感人群的概率）。仍然是采用马可尔夫链的时间连续思想建立对应的动力学方程：

$$\begin{cases} S_{t+1}=S_t-(r_1b_1S_tI_t)/N-(r_2b_2S_tE_t)/N+g_2R_t & (12)\\ E_{t+1}=E_t+(r_1b_1S_tI_t)/N+(r_2b_2S_tE_t)/N-aE_t & (13)\\ I_{t+1}=I_t+aE_t-g_1I_t & (14)\\ R_{i+1}=R_t+g_1I_t-g_2R_t & (15) \end{cases}$$

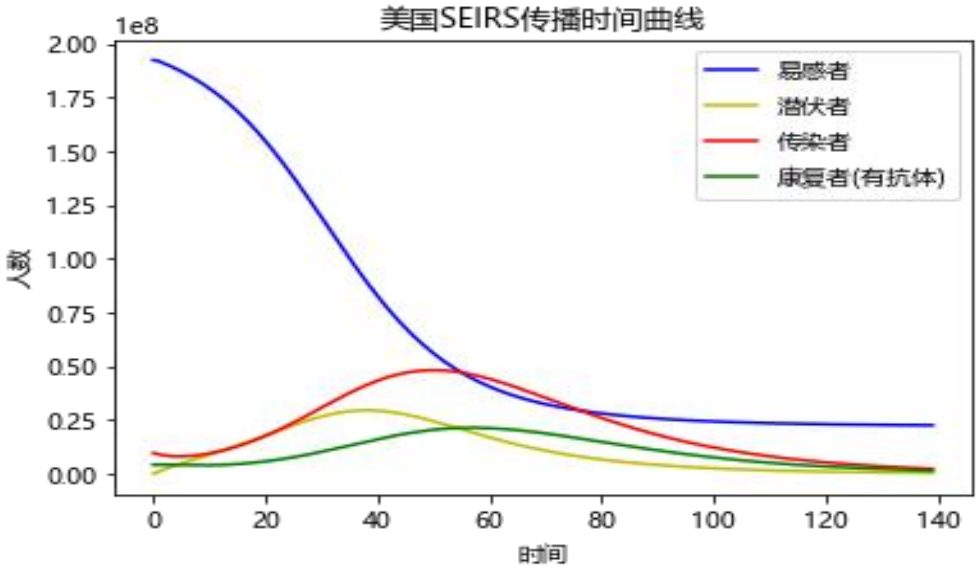
建立构建微分方程：

$$\begin{cases} \dfrac{dS}{dt}=-[(r_1b_1S_tI_t/N)+(r_2b_2E_tS_t/N)]+g_2R_t & (16)\\ \dfrac{dR}{dt}=aE_t-g_1I_t & (17) \end{cases}$$

$$\frac{dE}{dt} = (r_1 b_1 S_t I_t) / N + (r_2 b_2 S_t E_t) / N - aE_t \quad (18)$$

$$\frac{dR}{dt} = g_1 I_t - g_2 R_t \quad (19)$$

通过方程进行模型拟合后得出如下的变化情况图表：



图九：美国 SEIR 传播曲线

分析图表中可得出该模型的具体特征。假定确诊人群和潜伏者前期有一个初始值 I_0 和 E_0，这个值由当时的医疗条件和人民身体素质联合确定，由此我们通过分析得出其传播情况的判定标准值为：$1/\sigma = \frac{(g_2 - d)S_t}{(b_1 r_1 + b_2 r_2) g_1 R_t}$，直观的看出 $1/\sigma$ 和易感人群 S 存在着正比关系，所以在美国疫情情况走向图中也可以发现易感者降低到一定程度后减缓速率会降低，这是因为受到康复者重新转化为易感人群的影响。可以发现的是，传染者和潜伏者所达到的高峰并不高，高峰后转低后仍会维持在一定的程度，但不会趋于 0，这是由美国的具体情况而定，人口基数适中，地广，远距离传播存在困难。分析标准值，想要减少疫情的扩散则需要增大该阈值，可以发现增大易感人群是可以实现的，因为当易感人群占大多数时，传染者、康复者、潜伏者的占比低，则康复者仍会持续转化为易感人群，从短暂的时间来看，增加易感人群可取，但长远来看是不可取的，因为变异的情况未知，在传染性极强的情况下，短时间内庞大的易感人群和极强的传染性能瞬间摧毁医疗防线，会使得疫情不可收拾。所以正确的做法仍然是，减少接触率、传染率和医疗水平

以及人民的身体素质，提高治愈率。

该模型适用于有着一定的医疗卫生基础和经济基础，能够承受住强变异性病毒来袭的国家，其中要考虑该国的具体对外政策和对内的管控政策，从该模型拟合出来的结果来看，美国雄厚的实力对付新冠也是有点不足，仍需在防疫政策上采取强硬的措施。

5.3 问题三模型的建立与求解

本问题从多方数据和官方发布的通告研究得出中国始终坚持“动态清零”政策，而以美国英国为首的多个西方国家采取的是“群体免疫”、消极对待的政策。基于两个的政策的显著不同及问题一、问题二的分析，我们对中国也使用 SEIRS 的数学模型，但由于有误差存在，为更加精确的估计发展情况，我们采用了 ARIMA 的模型进行拟合后预测。

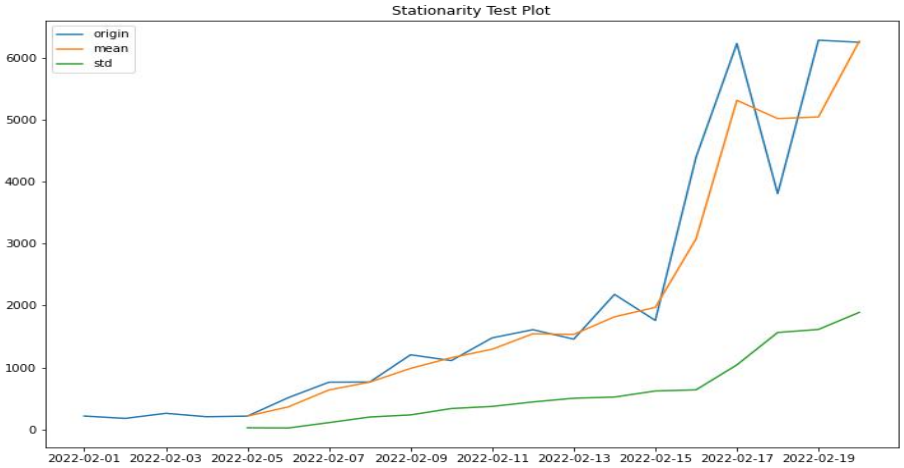
5.3.1 数据处理

（1）判定数据是否平稳化。

我们通过平均值和标准差的公式计算判定进行平稳化判定。平稳的数据应该不随时间的变化有较大的改动。计算公式如下：

$$\begin{cases} \text{平均值公式：}\ \bar{X} = \dfrac{X_1,\ldots,X_n}{n} & (20) \\ \text{标准差公式：}\ S = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} & (21) \end{cases}$$

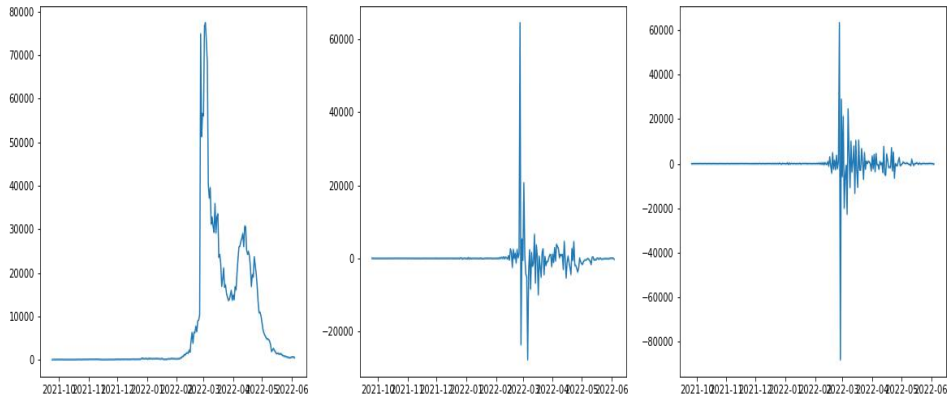
如下图所示，其平均值随时间呈现递增的趋势。因此数据不具有平稳性。



图十：中国 2 月 1 至 2 月 19 日的新增确诊数、平均值数、标准差数

（2）平稳化处理。

通过 pandas 中的 diff 函数进行二阶差分处理后，数据基本平稳。如下图是对差分方程模拟曲线的过程：



图十一：原始数据、一阶差分、二阶差分

5.3.2 不同 ARIMA 模型以及选择

（1）自回归模型 AR

自回归模型描述当前值与历史值之间的关系，用变量自身的历史时间数据对自身进行预测。自回归模型必须满足平稳性的要求。

自回归模型首先需要确定阶数 p，表示用几期的历史值来预测当前值。p 阶自回归模型的公式定义为：

$$y_t = \mu + \sum_{i=1}^{n}{}^{p} \gamma_i y_{t-i} + \varepsilon_t \qquad (22)$$

上式中 yt 是当前值,u 是常数项,p 是阶数 ri 是自相关系数,et 是误差。自回归模型有如下的限制：

1、自回归模型是用自身的数据进行预测

2、时间序列数据必须具有平稳性

3、自回归只适用于预测与自身前期相关的现象

（2）移动平均模型 MA

移动平均模型关注的是自回归模型中的误差项的累加，q 阶自回归过程的公式定义如下：

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{n} {}^{q}\theta_i \varepsilon_{t-i} \quad (23)$$

移动平均法能有效地消除预测中的随机波动。

（3）自回归移动平均模型 ARMA

自回归模型 AR 和移动平均模型 MA 模型相结合，我们就得到了自回归移动平均模型 ARMA(p,q)，计算公式如下：

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (24)$$

（4）差分自回归移动平均模型 ARIMA

将自回归模型、移动平均模型和差分法结合，我们就得到了差分自回归移动平均模型 ARIMA(p,d,q)，其中 d 是需要对数据进行差分的阶数。

（5）模型选择

通过各种比较得知，ARIMA 模型将各种模型的优点进行集合，是一个最为完备的模型。因此我们选择了 ARIMA 模型进行预测分析，同时混合了 SEIR 模型，以求数据更加贴合实际。

5.3.3 ARIMA 模型搭建

5.3.3.1 模型识别和定阶

模型的识别问题和定阶问题，主要是确定 p，d，q 三个参数，差分的阶数 d 一般通过观察图示，1 阶或 2 阶即可。这里我们主要介绍 p 和 q 的确定。我们首先介绍两个函数。

1、自相关函数 ACF(autocorrelation function)描述的是时间序列观测值与其过去的观测值之间的线性相关

性。计算公式如下：

$$ACF(\mathrm{k}) = \rho_k = \frac{Cov(\mathrm{y}_t, \mathrm{y}_{t-k})}{Var(\mathrm{y}_t)} \quad (25)$$

其中 k 代表滞后期数，如果 k=2，则代表 yt 和 yt-2

2、偏自相关函数 PACF(partial autocorrelation function)描述的是在给定中间观测值的条件下，时间序列观测值预期过去的观测值之间的线性相关性。

3、拖尾和截尾

拖尾指序列以指数率单调递减或震荡衰减，而截尾指序列从某个时点变得非常小；

出现以下情况，通常视为(偏)自相关系数 d 阶截尾：

1）在最初的 d 阶明显大于 2 倍标准差范围；

2）之后几乎 95%的(偏)自相关系数都落在 2 倍标准差范围以内；

3）且由非零自相关系数衰减为在零附近小值波动的过程非常突然。

出现以下情况，通常视为(偏)自相关系数拖尾：

1）如果有超过 5%的样本(偏)自相关系数都落入 2 倍标准差范围之外；

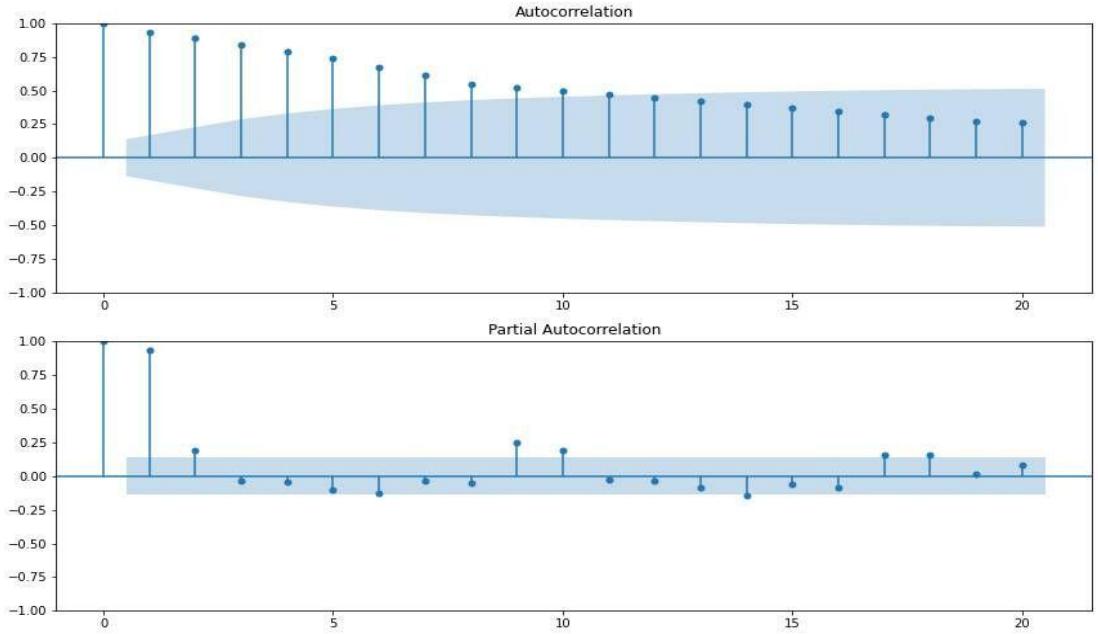
2）或者是由显著非 0 的(偏)自相关系数衰减为小值波动的过程比较缓慢或非常连续。

5.3.3.2 确定阶数和参数估计

p，q 的确定基于如下的规则：

模型（序列）	AR(p)	MA(q)	ARMA(p, q)
自相关函数	拖尾	第q个后截尾	拖尾
偏自相关函数	第p个后截尾	拖尾	拖尾

给出如下的拖尾和截尾情况：



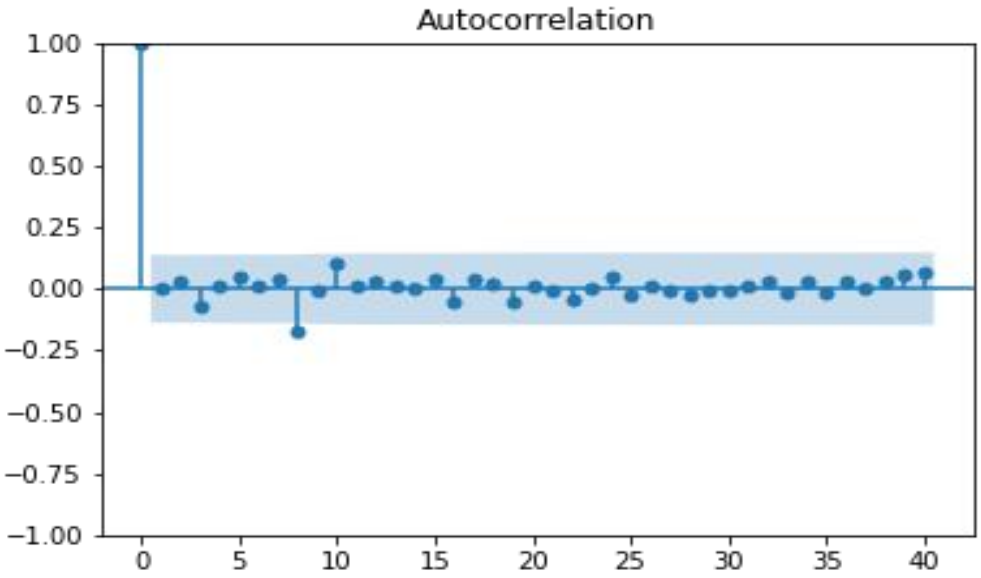
图十二：ACF 和 PACF 的置信区间

分析可以得知数据自相关系数 13 阶拖尾，偏自相关系数 4 阶截尾。通过拖尾和截尾对模型进行定阶的方法，具有很强的主观性。结合最终的预测误差来确定 p，q 的阶数，在相同的预测误差情况下，根据奥斯卡姆剃刀准则，模型越小越好。所以，平衡预测误差和参数个数，即根据信息准则函数法来确定模型的阶数。预测误差通常用平方误差即残差平方和来表示。

通过网络搜索的方式来寻找 AR 模型最佳的 p,q 组合，用 AIC 和 BIC 两个准则进行试验，得出 p,q 最优值 AIC(6,4) BIC(3,3)。

5.3.3.3 模型检验

检验如下两个目标：参数估计的显著性（t 检验），检验残差序列的随机性给，即残差之间是独立的。如下给残差序列的随机性通过相关函数法来检验，得出如下的自相关函数图，依次来看，检测合格。

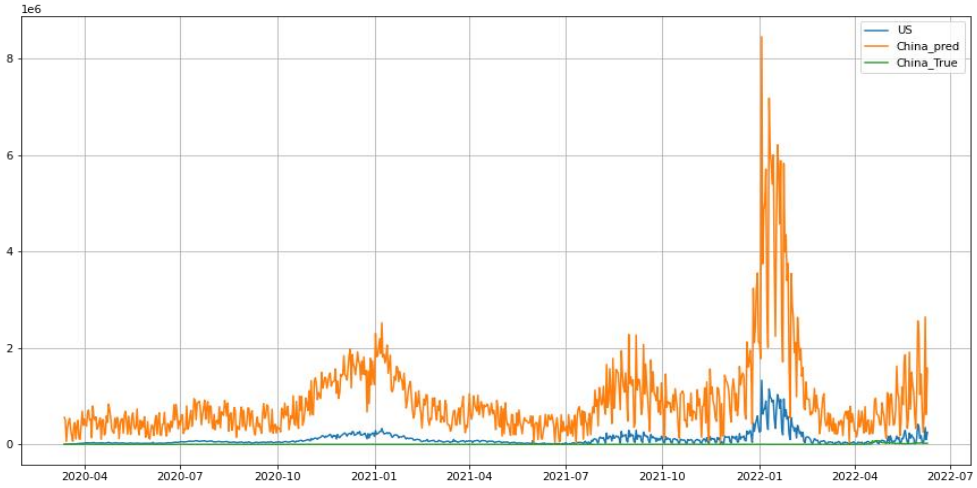


图十三：自相关函数图

5.3.4 模型预测

预测主要有两个函数，一个是 predict 函数，分别是 forecast 函数，和 predict。predict 中进行预测的时间段必须在训练 ARIMA 模型的数据中，forecast 则对训练数据集末尾下一个时间段的值进行预估。

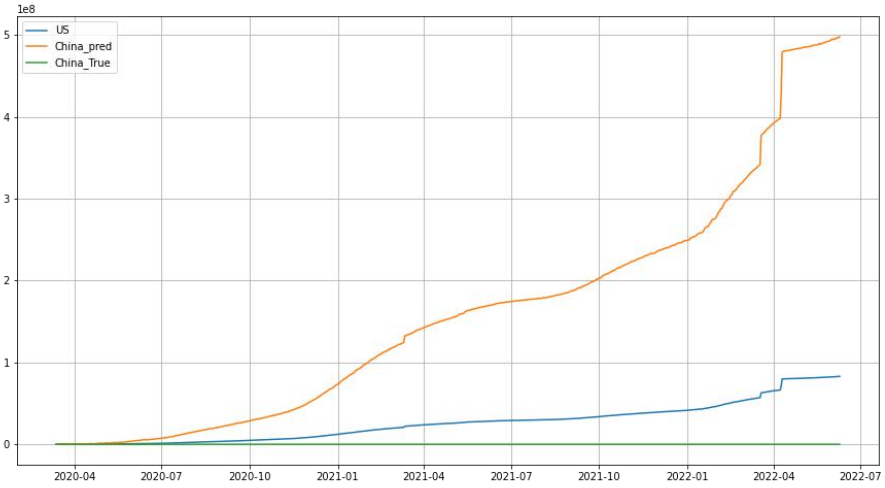
（1）考虑到美国不采取动态清零的情况，且基础条件和中国相似，基于美国的疫情情况进行 ARIMA 模型的训练，得出结果后对中国进行合理的预测，预测结果如下：



图十四：蓝色：美国真实数据 橙色：不采取：动态清零“中国预测数据

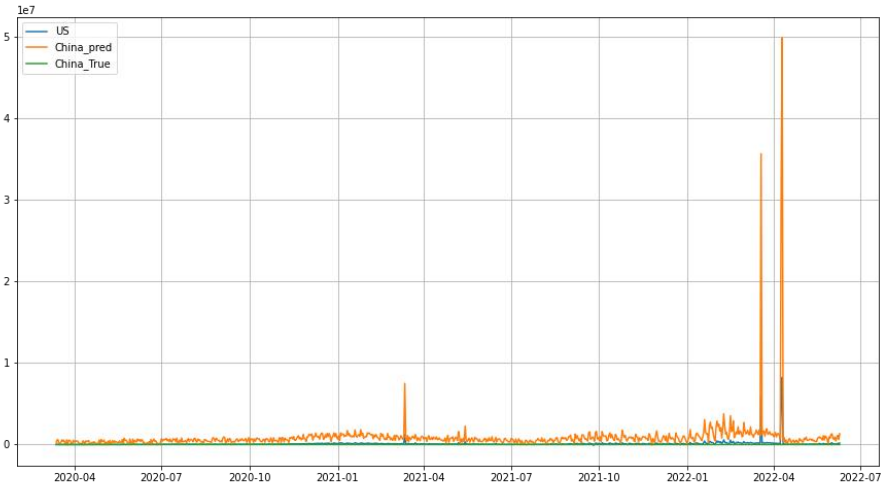
如上结果可以得出，中国如若不采取动态清零情况，前期稳定增长，在经历几次增长波峰后会在 2022 年 1 月左右迎来最高的爆发期。和美国的曲线对比，由于中国人口基数的区别，中国经历的小波峰会多于美国，但会在大致相同的时期达到爆发的巅峰，峰值过后会下降，但在一段时间内均有反弹的可能。

如下给出日均治愈率的曲线图：



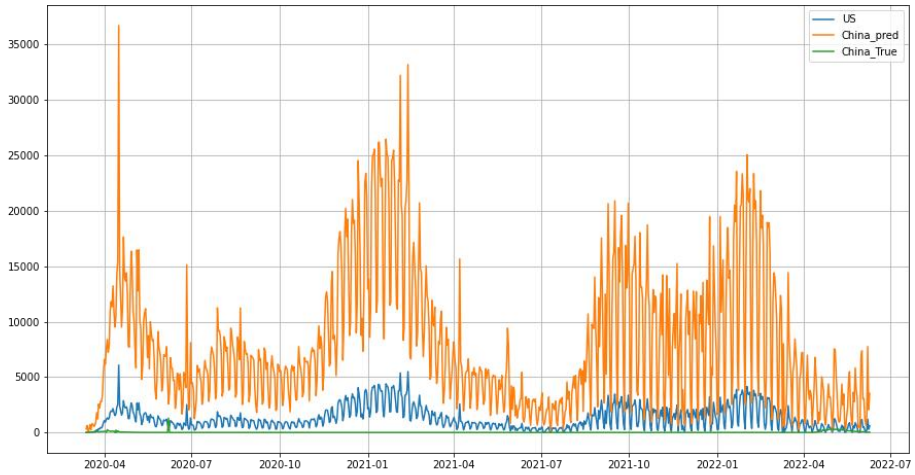
图十五：日均治愈率

分析得出在疫情到来之后日均治愈率会随着感染者成正比的关系，增长速率愈来愈大。日均治愈率聚合后得出如下的月治愈率：



图十六：月治愈率

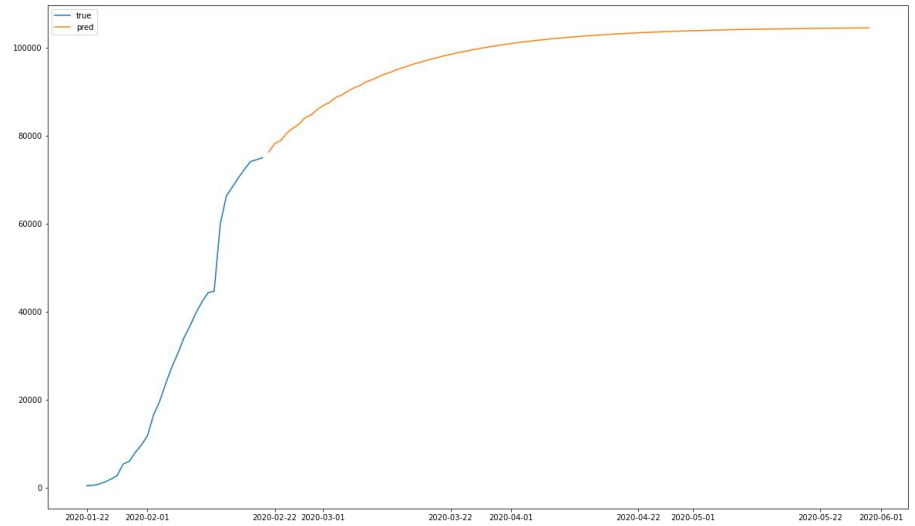
同理根据模型预测，给出如下的死亡情况：



图十七：橙色：不采取“动态清零”中国死亡曲线 蓝色：美国死亡曲线

分析得出在疫情初期死亡情况达到巅峰，随着治愈率增加会逐渐降低，但前后会受感染情况的影响，经历数次波峰，结合实际情况，预测情况是合理的。

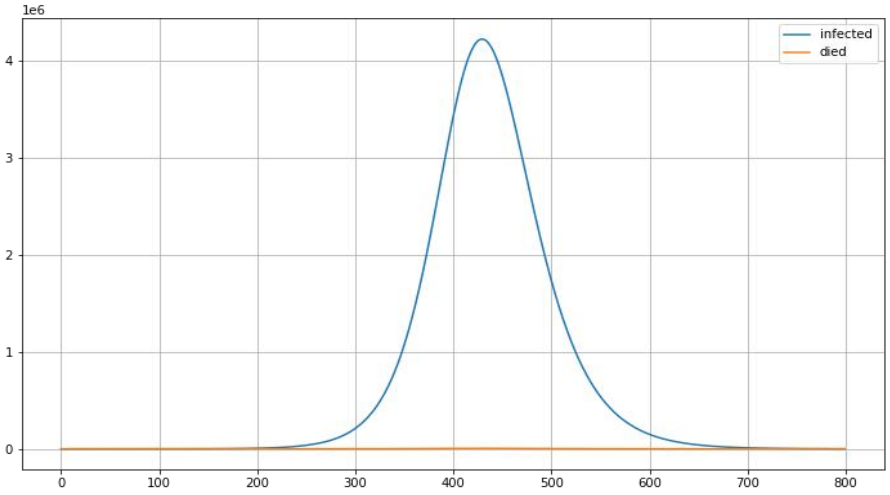
（2）作为对比给出中国自身的参照情况。分析得出结论，采取疫情初期严重时未采取严厉管控措施，可视为是未动态清零下的一个小样本。取 2020 年 1 月 24 日开始后的一个月的时间内时间作为训练基础数据，得出结果后进行后续的预测，即可作为对照的一直不采取动态清零情况下中国的疫情情况走向。如下所示：



图十八：不采取“动态清零”的预测曲线

据折线图可以发现，在前期的增长速率大，在预测阶段，依然保持增长趋势，但会有着相应的减缓，在 2020 年 6 月开始进入平缓的阶段，随后会进入递减阶段，大致与上述中在美国数据的基础上训练的 ARIMA 得出的结果基本吻合。

（3）此处再给出另一个解体思路，英国采取的防疫政策为“群体免疫”，可认为是松懈的低防备政策，在基于第二问中 SIR 模型情况下进行模拟计算，算出 SIR 模型中的日接触率和日治愈率。后续对中国进行 SIR 模型拟合，估测中国在采取不动态清零情况下的疫情发展情况。得出的死亡情况和治愈情况结果如下：



图十九：SIR 情况下 蓝色：中国感染人数 橙色：中国死亡人数

（4）总结来看，在本题中采取的 ARIMA 模型预测结果具有一定的理论基础性和数据真实性，对预测有着一定参考性，可以发现中国在不采取动态清零的情况下疫情情况会严重许多，故动态清零是合理的政策。

六、模型评价与推广

6.1 模型评价

6.1.1 模型优点

（1）分析所选取的三个分类模型，都相对符合其选择的国家疫情情况。SIR 模型符合传统传染病问题，求解过程严谨，数据确切可信。SEIR 在 SIR 基础上结合了潜伏者存在的情况，优化后模拟的数据与实际贴合，有一定的合理预测性。SEIRS 是对 SIR 的另一个改进，增加了实际存在变异的情况，考虑更为全面，一样严格的符合马克尔夫链，时间连续性保证数据求解过程严谨且真实可信，通过拟合，也有一定合理的预测能力。

（2）第三问中采用的 ARIMA 模型，结合了二者的优点。SEIR 模型可以较好的

预测疫情的发展趋势，但对感染率是估算的，在后续的其他的因素影响下，是无法根据事实进行动态调整的。ARIMA 模型基于时间序列预测，具有由内生变量所决定不受外界因素干预的优点很好地弥补了 SEIR 模型的弊端。该模型学习到一般传染病的发展趋势，又避免后期因管控措施、病毒变异等影响，造成外生变量的改变，而进行的分阶段划分分析，根据历史数据的发展趋势，合理统筹未来接下来疫情发展的优点。

（3）通过与 SEIR、ARIMA、SIR 等模型在不同时间、不同地 点预测分析对比，得出本文提出的 ARIMA 模型有着较好的数据真实性、逻辑严谨性、更贴合事实等优点，在对国家未来预防其 他类型传染病具有良好的应用价值。

6.1.2 模型缺点

（1）预测的时候部分数据规律不明确，导致部分预测结果得不够准确。

（2）ARIMA 模型过程较为复杂，计算时需要同时处理的数据过多，对计算机的算力有一定的要求。

（3）分类模型中，给出的模型不够贴合实际，在某些特殊的情况影响下，所预测的结果不够稳定，大量数据的处理，对模型的严谨性高。

6.2 模型的改进

（1）筛选数据的时候分类处理，分类导入到模型中进行估测，使数据更加准确。同时给模型设定评价指标，进行合理性校验。

（2）利用循环迭代的方法对模型进行优化，减少计算要求，更新算法，采用时间复杂度和空间复杂度较低的算法，来减少对计算机的要求。

（3）后续根据 ARIMA 模型的特性增加新的评估预测指标，来评价数据的精确度，减少平均、绝对、平均绝对百分比误差。

6.3 误差分析

因为各类模型中的预测是基于现存数据进行特征值提取和数学推算得出，对病毒的变异情况、人群的防范意识的改变、可能出现的特效药和疫苗等原因考虑不周会导致数据的偏差。我们利用 ARIMA 模型、SEIR 模型、SIR 模型推算出的疫情数据和现实数据进行比对，得到平均相对误差值。随机选取 5 个月的数据，得到相对误差，最后平均相对误差为 7.21%。

第一组 ARIMA	第二组 SEIR	第三组 SIR
4.58	7.61	9.42

参考文献

[1]姜启源，谢金星，数学模型(第四版)[M]，北京：高等教育出版社，2018,85-94

[2]董章功,宋波,孟友新.基于 SEIR-ARIMA 混合模型的新冠肺炎预测[J].计算机与现代化,2022(02):1-6.

[3]陈田木,赵泽宇,芮佳,余珊珊,祝媛钊,徐静文,刘星纯,王瑶,杨蒙,李佳,刘若云,谢昉,雷照,赵本华,王明斋,苏艳华.厦门市新型冠状病毒肺炎人群传播能力计算与防控措施效果的模拟评估[J].厦门大学学报(自然科学版),2020,59(03):298-303.

[4]肖佳.西非三国埃博拉疫情控制模型的建立与分析［Ｄ］.重庆：重庆大学,２０１８.

[5] 潘典雅，基于 ARIMA 模型的吉林省 GDP 分析及预测，中国集体经济，2021.8

[6]傅惠民，刘成瑞，马小兵．时间序列均值和方差函数的确定方法[J].机械强度，2004，26(2)：164-169.

[7] 桂思思，基于 ARIMA 与线性回归组合模型的汽车销量预测分析，计算机与数字工程， 2021.08

[8] 司宛玲，司守奎，数学建模简明教程[M]北京：国防工业出版社，2019,185-189

附录

一. ARIMA 模型的 python 代码(因代码过多,具体代码见附录压缩包)

```
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
confirmed_china=pd.read_csv('./data/American.csv',encoding='gbk',index_col    =
'date',parse_dates=['date'])
confirmed=confirmed_china[50:]
# confirmed['confirmed']=confirmed['confirmed'].diff(1).fillna(0)
confirmed['Close_diff_1'] = confirmed['confirmed'].diff(1)
confirmed['Close_diff_2'] = confirmed['Close_diff_1'].diff(1)
fig = plt.figure(figsize=(20,6))
ax1 = fig.add_subplot(131)
ax1.plot(confirmed['confirmed'])
ax2 = fig.add_subplot(132)
ax2.plot(confirmed['Close_diff_1'])
ax3 = fig.add_subplot(133)
ax3.plot(confirmed['Close_diff_2'])
plt.savefig('./img/Q3/diff')
plt.show()
```

二. SIR、SEIR、SEIRS 模型的 python 代码（因代码过多，具体代码见附录压缩包）

```
gammaguess = (df1["治愈人数"] + df1["死亡人数"]) / df1["累计确诊"]
gamma = gammaguess.mean()
fig, ax = plt.subplots(figsize=(14, 7))
ax.plot(df1.index, gammaguess)
ax.set_title("SIR model of γ changes")
ax.set_xlabel('Days from discovery')
ax.set_ylabel('γ')
ax.grid(True)
plt.show()

'''
SIR 模型的参数确定
'''
class estimationInfectionProb():
    def __init__(self, estUsedTimeIndexBox, nContact, gamma):
        self.timeRange = np.array([i for i in range(estUsedTimeIndexBox[0],
estUsedTimeIndexBox[1] + 1)])
```

```
        self.nContact, self.gamma = nContact, gamma
        self.dataStartTimeStep = 0
    def setInitSolution(self, x0):
        self.x0 = 0.04
    def costFunction(self, infectionProb):
        res = np.array(np.exp((infectionProb * self.nContact - self.gamma) *
self.timeRange) - \
    df1.loc[self.timeRange - self.dataStartTimeStep, '累计确诊'])
        return (res ** 2).sum() / self.timeRange.size
    def optimize(self):
        self.solution = minimize(self.costFunction, self.x0,
method='nelder-mead', options={'xtol': 1e-8, 'disp': False})
        return self.getSolution()
    def getSolution(self):
        return self.solution.x
    def getBasicReproductionNumber(self):
        self.basicReproductionNumber = self.nContact * self.solution.x[0] /
(self.gamma)
        print("basic reproduction number:", self.basicReproductionNumber)
        return self.basicReproductionNumber
```

三. 文件列表

1.代码包列表

1）ARIMA 模型代码

2）绘制 csv 文件代码

3）绘制曲线比率代码

4）SIR、SEIR、SEIRS 模型代码

名称	修改日期	类型	大小
ARIMA.ipynb	2022/6/14 4:19	IPYNB 文件	455 KB
draw_csv.ipynb	2022/6/14 4:19	IPYNB 文件	102 KB
draw_ratio.ipynb	2022/6/14 4:19	IPYNB 文件	30 KB
SIRandSEIRandSEIRS.ipynb	2022/6/14 4:16	IPYNB 文件	331 KB

2. 美国前 10 日数据（具体 csv 文件见压缩包）

国家	日期	累计确诊	新增确诊	累计治愈	新增治愈	累计死亡	新增死亡
美国	2022/6/9	87238151	249480	83009072	183073	1035620	589
美国	2022/6/8	86988671	96848	82825999	71713	1035031	283
美国	2022/6/7	86988671	351184	82825999	154475	1035031	1201
美国	2022/6/6	86637487	114926	82671524	86993	1033830	239
美国	2022/6/5	86522561	19504	82584531	56742	1033591	20
美国	2022/6/4	86503057	52454	82527789	64283	1033571	202
美国	2022/6/3	86450603	181223	82463506	69428	1033369	506
美国	2022/6/2	86269380	122425	82394078	90988	1032863	453
美国	2022/6/1	86503057	356102	82527789	224699	1033571	1161

3. 世界疫情前 10 天总数据（具体 csv 文件见压缩包）

日期	累计确诊	新增确诊	累计治愈	新增治愈	累计死亡	新增死亡
2022/6/10	539714877	530663	512329763	491136	6329728	1295
2022/6/9	539184214	628419	511838627	517585	6328433	1640
2022/6/8	538555795	637148	511321042	562258	6326793	1838
2022/6/7	537918647	604265	510758784	721866	6324955	1608
2022/6/6	537314382	389237	510036918	532890	6323347	917
2022/6/5	536925145	414991	509504028	521429	6322430	737
2022/6/4	536510154	528935	508982599	440681	6321693	1064
2022/6/3	535981219	597440	508541918	562611	6320629	1348
2022/6/2	535383779	646924	507979307	583743	6319281	1743

4. 法国前 20 天总数据（具体 csv 文件见压缩包）

法国	累计确诊	累计治愈	累计死亡
2020-01-22	0	0	0
2020-01-23	0	0	0
2020-01-24	2	0	0
2020-01-25	3	0	0
2020-01-26	3	0	0
2020-01-27	3	0	0
2020-01-28	4	0	0
2020-01-29	5	0	0
2020-01-30	5	0	0
2020-01-31	5	0	0
2020-02-01	6	0	0
2020-02-02	6	0	0

2020-02-03	6	0	0
2020-02-04	6	0	0
2020-02-05	6	0	0
2020-02-06	6	0	0
2020-02-07	6	0	0
2020-02-08	11	0	0
2020-02-09	11	0	0

5. 中美前十天数据对比（具体 csv 文件见压缩包）

日期	中国累计确诊	中国治愈人数	中国死亡人数	美国累计确诊	美国治愈人数	美国死亡人数
2020-01-22	643	28	17	1	0	0
2020-01-23	920	30	18	2	0	0
2020-01-24	1406	36	26	2	0	0
2020-01-25	2075	39	42	5	0	0
2020-01-26	2877	49	56	5	0	0
2020-01-27	5509	58	82	5	0	0
2020-01-28	6087	101	131	5	0	0
2020-01-29	8141	120	133	6	0	0
2020-01-30	9802	135	171	7	0	0

6. 中国前 10 天疑似新增数据（具体 csv 文件见压缩包）

日期	累计疑似
2020 年 1 月 16 日	0
2020 年 1 月 17 日	0
2020 年 1 月 18 日	0
2020 年 1 月 19 日	0
2020 年 1 月 20 日	54
2020 年 1 月 21 日	37
2020 年 1 月 22 日	393
2020 年 1 月 23 日	1072
2020 年 1 月 24 日	1965