# 基于新冠疫情数据的分析与建模

（杨浩然 10195501441）

**摘要：**2019 年 12 月爆发的 COVID-19 病毒对世界各国造成了巨大的影响，生命财产损失不计其数。从掌握疫情基本规律的现实意义出发，本文基于 Kaggle 公开的疫情数据，从可视化的角度对疫情状况进行了分析，通过 Exponential Smoothing 方法和 ARIMA 预测模型对未来疫情的发展做出了预测。

**关键词：**新冠病毒　数据分析　预测

# Analysis and Modeling Based on COVID-19 Data

**Abstract:** The COVID-19 outbroke in December 2019 has had a huge impact on countries around the world, causing untold loss of life and property. Based on the practical significance of learning the basic rules of the plague, this paper analyzed the plague situation from the perspective of data visualization based on the data published by Kaggle, and predicted the future development of the plague through Exponential Smoothing method and ARIMA prediction model.

**Key words:** COVID-19　data analysis　prediction

## 一、数据集与数据处理

### 1. 数据集介绍

新冠肺炎是指 2019 新型冠状病毒感染导致的肺炎。2019 年 12 月以来，湖北省武汉市部分医院陆续发现了多例有华南海鲜市场暴露史的不明原因肺炎病例，证实为 2019 新型冠状病毒感染引起的急性呼吸道传染病。3 月 11 日，世卫组织认为当前新冠肺炎疫情可被称为全球大流行。

疫情爆发后，各国及时更新每日数据。本文使用的数据集来自 Kaggle。

*COVID19 Daily Updates* 包含每日更新的共 983419 个数据，每个数据包含 8 个属性：国家/地区（Country/Region）、省/州（Province/State）、纬度（Latitude）、经度（Longitude）、确诊人数（Confirmed）、治愈人数（Recovered）、死亡人数（Deaths）、日期（Date）。如图 1-1-1。

| | Country/Region | Province/State | Latitude | Longitude | Confirmed | Recovered | Deaths | Date |
|---|---|---|---|---|---|---|---|---|
| **0** | China | Anhui | 31.8257 | 117.2264 | 1.0 | NaN | NaN | 2020-01-22 |
| **1** | China | Beijing | 40.1824 | 116.4142 | 14.0 | NaN | NaN | 2020-01-22 |
| **2** | China | Chongqing | 30.0572 | 107.8740 | 6.0 | NaN | NaN | 2020-01-22 |
| **3** | China | Fujian | 26.0789 | 117.9874 | 1.0 | NaN | NaN | 2020-01-22 |
| **4** | China | Gansu | 35.7518 | 104.2861 | NaN | NaN | NaN | 2020-01-22 |

图 1-1-1

### 2. 数据初处理

本文主要进行了数据格式和补齐缺省数据的处理。如图 1-2-1。

将缺省值用 0 填充，对数据格式进行了合适的转换。在不同任务中进行了数据分类、聚合等操作。

总的来说，*COVID19 Daily Updates* 数据集较为规范，可用性高。

| | Country/Region | Date | Confirmed | Recovered | Deaths |
|---|---|---|---|---|---|
| **0** | Afghanistan | 2020-02-24 | 1.0 | 0.0 | 0.0 |
| **1** | Afghanistan | 2020-02-25 | 1.0 | 0.0 | 0.0 |
| **2** | Afghanistan | 2020-02-26 | 1.0 | 0.0 | 0.0 |
| **3** | Afghanistan | 2020-02-27 | 1.0 | 0.0 | 0.0 |
| **4** | Afghanistan | 2020-02-28 | 1.0 | 0.0 | 0.0 |

图 1-2-1

## 二、数据可视化

主要选取中、美的数据，对每日疫情变化趋势进行可视化，并且展示中国大陆部分省市数据（图 2-1-1）。对世界疫情状况直观表现（图 2-1-2）。

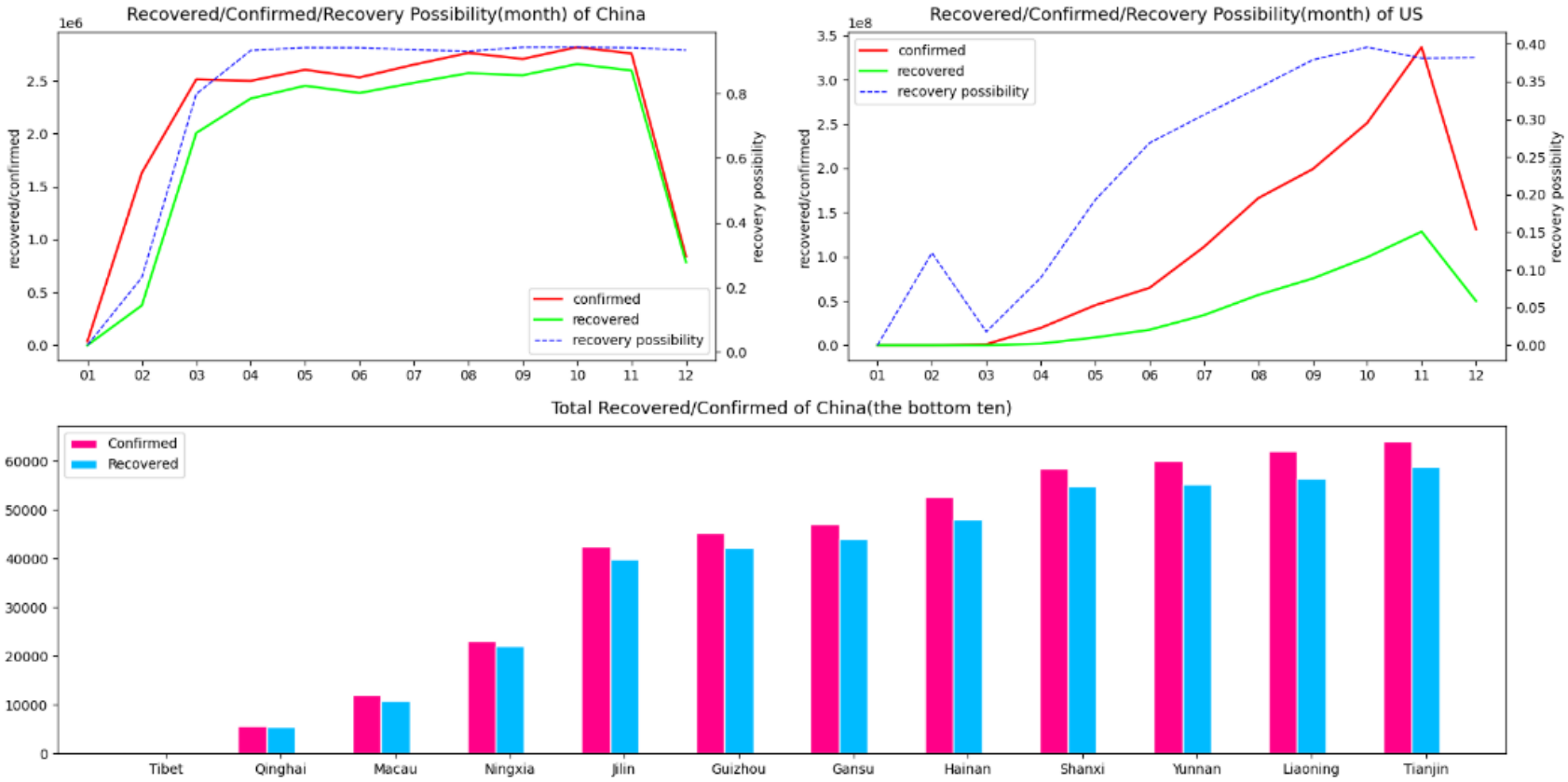

Recovered/Confirmed/Recovery Possibility(month) of China
1e6
recovered/confirmed
recovery possibility
confirmed
recovered
recovery possibility
Recovered/Confirmed/Recovery Possibility(month) of US
1e8
recovered/confirmed
recovery possibility
confirmed
recovered
recovery possibility
Total Recovered/Confirmed of China(the bottom ten)
Confirmed
Recovered
Tibet
Qinghai
Macau
Ningxia
Jilin
Guizhou
Gansu
Hainan
Shanxi
Yunnan
Liaoning
Tianjin

图 2-1-1

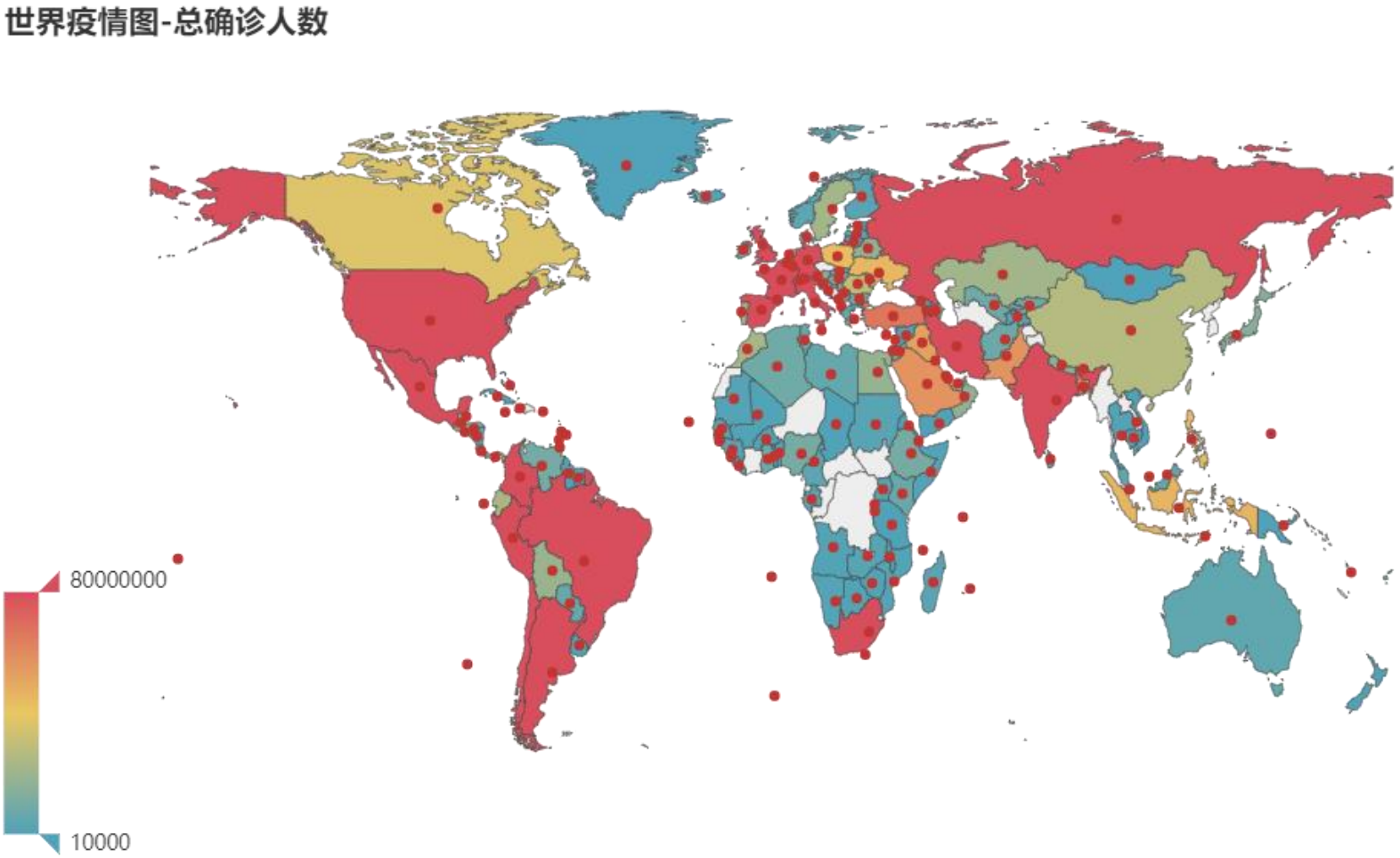



图 2-1-2

# 三、数据建模及预测

## 1. Exponential Smoothing Method

通过对 statsmodels.tsa.api 中的 ExponentialSmoothing 模块的直接调用，使用默认参数，得到以下结果。

从可视化观察角度来看，模型性能并不好。

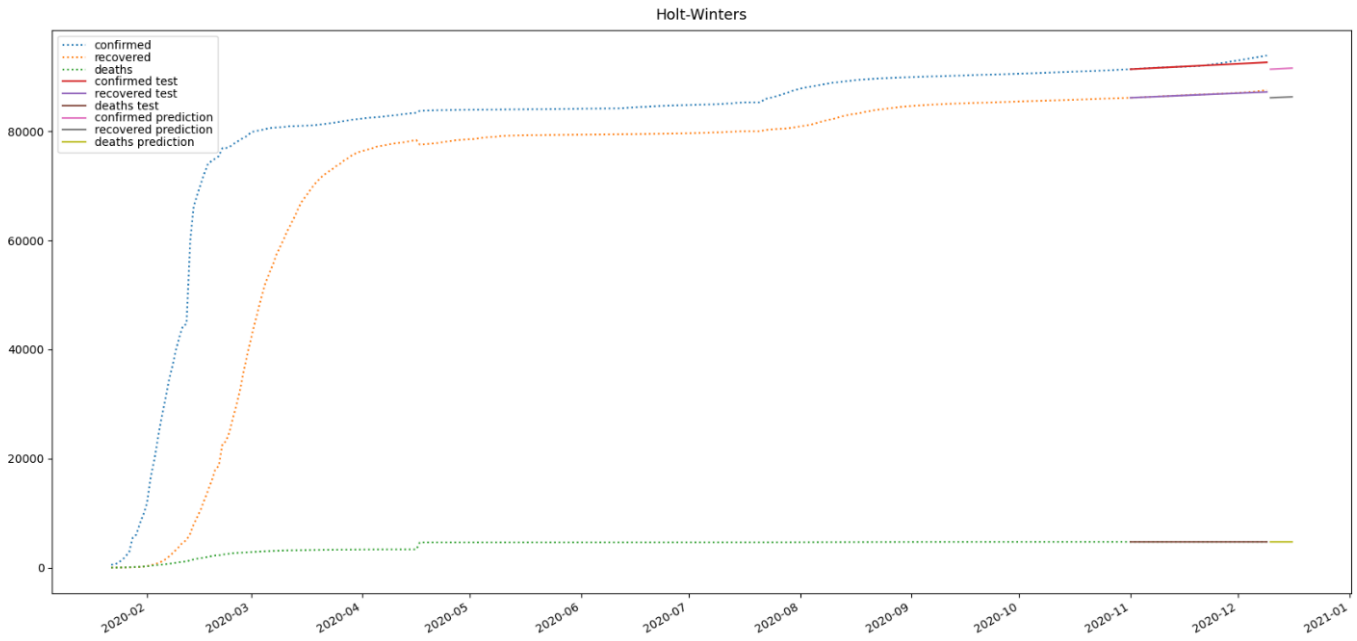



图 3-1-1

| Date | confirmedPred | recoveredPred | deathsPred |
|---|---|---|---|
| 2020-12-10 | 91399.055656 | 86163.382766 | 4739.298521 |
| 2020-12-11 | 91432.111312 | 86191.422547 | 4739.597041 |
| 2020-12-12 | 91465.166968 | 86219.462328 | 4739.895562 |
| 2020-12-13 | 91498.222624 | 86247.502109 | 4740.194082 |
| 2020-12-14 | 91531.278280 | 86275.541890 | 4740.492603 |
| 2020-12-15 | 91564.333935 | 86303.581671 | 4740.791123 |
| 2020-12-16 | 91597.389591 | 86331.621452 | 4741.089644 |

图 3-1-2

| | confirmed | recovered | deaths |
|---|---|---|---|
| RMSE | 481.574553 | 98.000588 | 2.121754 |

图 3-1-3

## 2. ARIMA Prediction Model

### 2.1 模型简介

ARIMA 模型（Autoregressive Integrated Moving Average Model），差分整合移动平均自回归模型，又称整合移动平均自回归模型（移动也可称作滑动），是时间序列预测分析方法之一。

ARIMA(p, d ,q)模型是 ARMA(p, q)模型的扩展。ARIMA(p, d, q)模型可以表示为：

$$(1-\sum_{i=1}^{p}\emptyset_i L_i)(1-L)^d X_t = (1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t$$

其中$L$是滞后算子（Lap operator），$d \in Z, d > 0.$

### 2.2 模型定义

非平稳时间序列，在消去其局部水平或者趋势之后，其显示出一定的同质性，也就是说，此时序列的某些部分 与其它部分很相似。这种非平稳时间序列经过差分处理后可以转换为平稳时间序列，那么称这样的时间序列为齐次非平稳时间序列，其中差分的次数就是齐次的阶。

将∇记为差分算子，那么有

$$\nabla^2 y_t = \nabla(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2}$$

对于延迟算子$B$，有

$$y_{t-p} = B^p y_t, \forall p \geq 1$$

因此可以得出

$$\nabla^k = (1-B)^k$$

设有$d$阶其次非平稳时间序列，那么有$\nabla^d y_t$是平稳时间序列，则可以设其为ARMA(p, q)模型，即

$$\lambda(B)(\nabla^k y_t) = \theta(B)\varepsilon_t$$

其中，

$$\lambda(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \cdots - \lambda_p B^p \,, \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$$

当差分阶数$d$为 0 时，ARIMA 模型就等同于 ARMA 模型，即这两种模型的差别就是$d$是否等于零，也就是序列是否平稳，ARIMA 模型对应着非平稳时间序列，ARMA 模型对应着平稳时间序列。

### 2.3 实验步骤

1）数据预处理

| | Country/Region | Date | Confirmed | Recovered | Deaths |
|---|---|---|---|---|---|
| **2020-01-22** | China | 2020-01-22 | 547.0 | 28.0 | 17.0 |
| **2020-01-23** | China | 2020-01-23 | 639.0 | 30.0 | 18.0 |
| **2020-01-24** | China | 2020-01-24 | 916.0 | 36.0 | 26.0 |
| **2020-01-25** | China | 2020-01-25 | 1399.0 | 39.0 | 42.0 |
| **2020-01-26** | China | 2020-01-26 | 2062.0 | 49.0 | 56.0 |

| | Confirmed |
|---|---|
| **2020-01-22** | 547.0 |
| **2020-01-23** | 639.0 |
| **2020-01-24** | 916.0 |
| **2020-01-25** | 1399.0 |
| **2020-01-26** | 2062.0 |

| | Recovered |
|---|---|
| **2020-01-22** | 28.0 |
| **2020-01-23** | 30.0 |
| **2020-01-24** | 36.0 |
| **2020-01-25** | 39.0 |
| **2020-01-26** | 49.0 |

| | Deaths |
|---|---|
| **2020-01-22** | 17.0 |
| **2020-01-23** | 18.0 |
| **2020-01-24** | 26.0 |
| **2020-01-25** | 42.0 |
| **2020-01-26** | 56.0 |

2）数据的平稳性处理及检验

由图 3-2-1 可以看出三个时间序列波动性比较大，不是稳定的时间序列，可以做差分转化为线性趋势（图 3-2-2）。

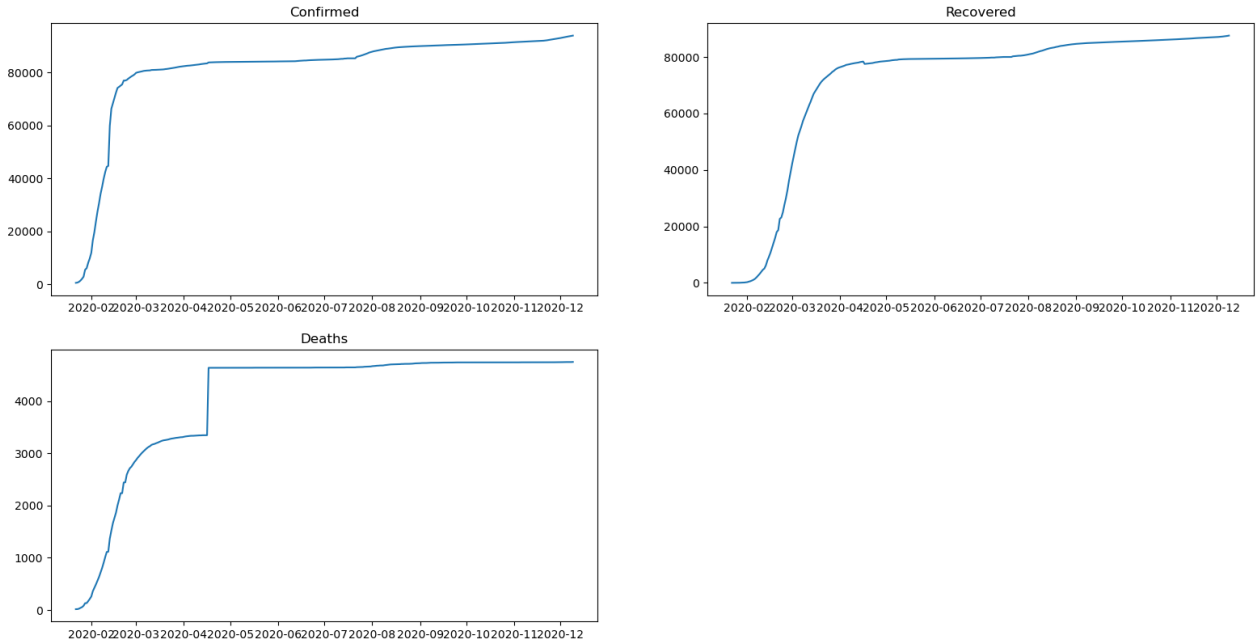



图 3-2-1

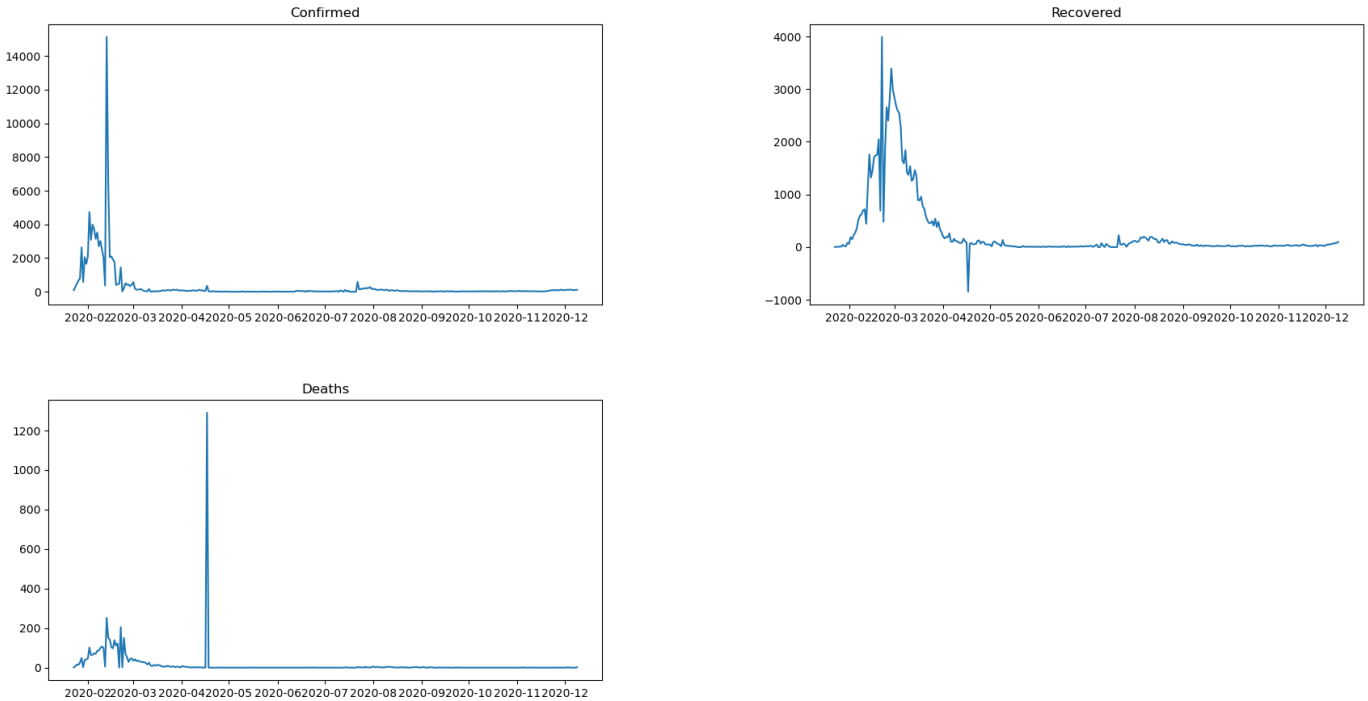



图 3-2-3

做差分处理之后，对时间序列进行 ADF 检验。

| | |
|---|---|
| **Test Statistic Value** | -4.39524 |
| **p-value** | 0.000302638 |
| **Lags Used** | 16 |
| **Number of Observations Used** | 305 |
| **Critical Value(1%)** | -3.45197 |
| **Critical Value(5%)** | -2.87106 |
| **Critical Value(10%)** | -2.57184 |

图 3-2-4 （confirmed ADF）

| | |
|---|---|
| **Test Statistic Value** | -3.59239 |
| **p-value** | 0.00590826 |
| **Lags Used** | 16 |
| **Number of Observations Used** | 305 |
| **Critical Value(1%)** | -3.45197 |
| **Critical Value(5%)** | -2.87106 |
| **Critical Value(10%)** | -2.57184 |

图 3-2-5 （recovered ADF）

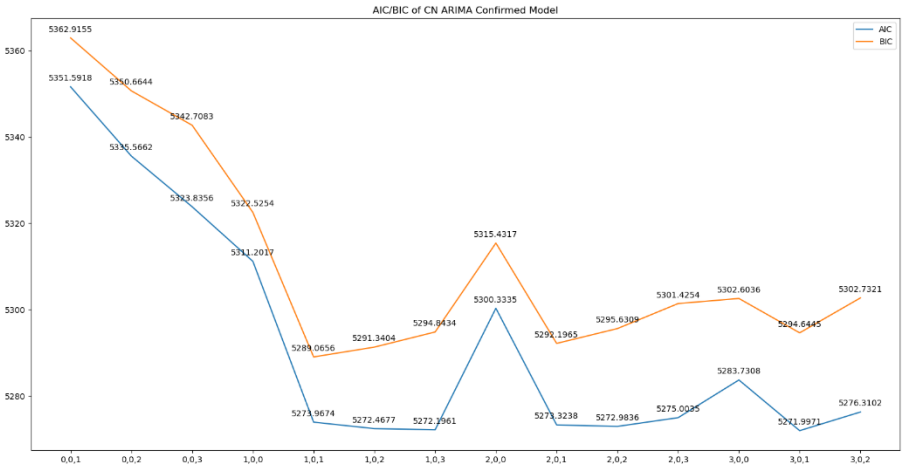
| | |
|---|---|
| **Test Statistic Value** | -10.8349 |
| **p-value** | 1.67147e-19 |
| **Lags Used** | 1 |
| **Number of Observations Used** | 320 |
| **Critical Value(1%)** | -3.45095 |
| **Critical Value(5%)** | -2.87061 |
| **Critical Value(10%)** | -2.5716 |

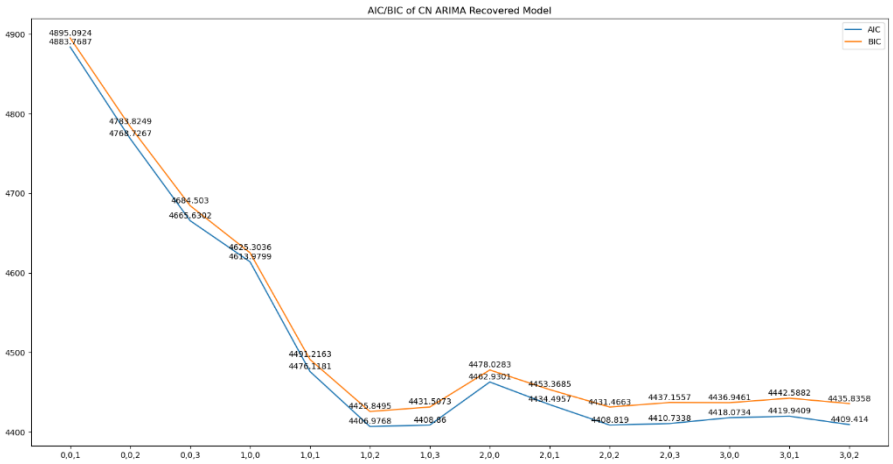
图 3-2-6 （deaths ADF）

差分之后的序列基本达到稳定，并且通过了 ADF 检验。

3）确定模型参数

这里主要根据 AIC 准则和 BIC 准则相结合的方式来确定 ARMA 模型的阶数，应当选取 AIC 和 BIC 值达到最小的那一组为理想阶数。



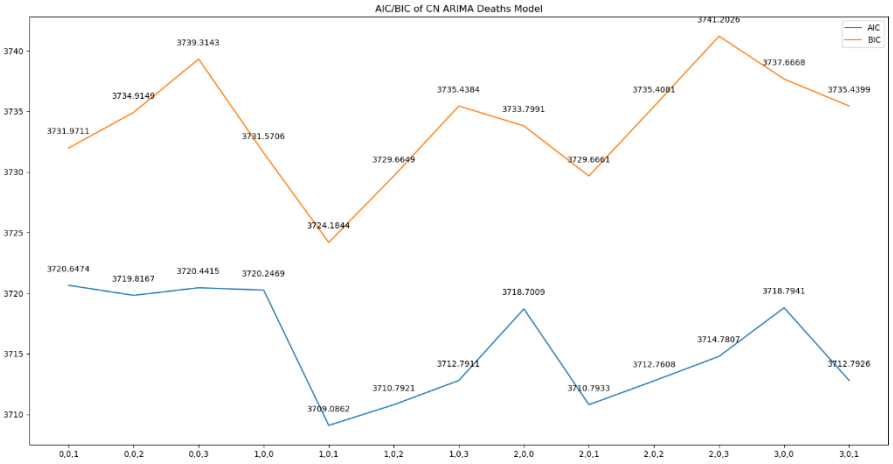



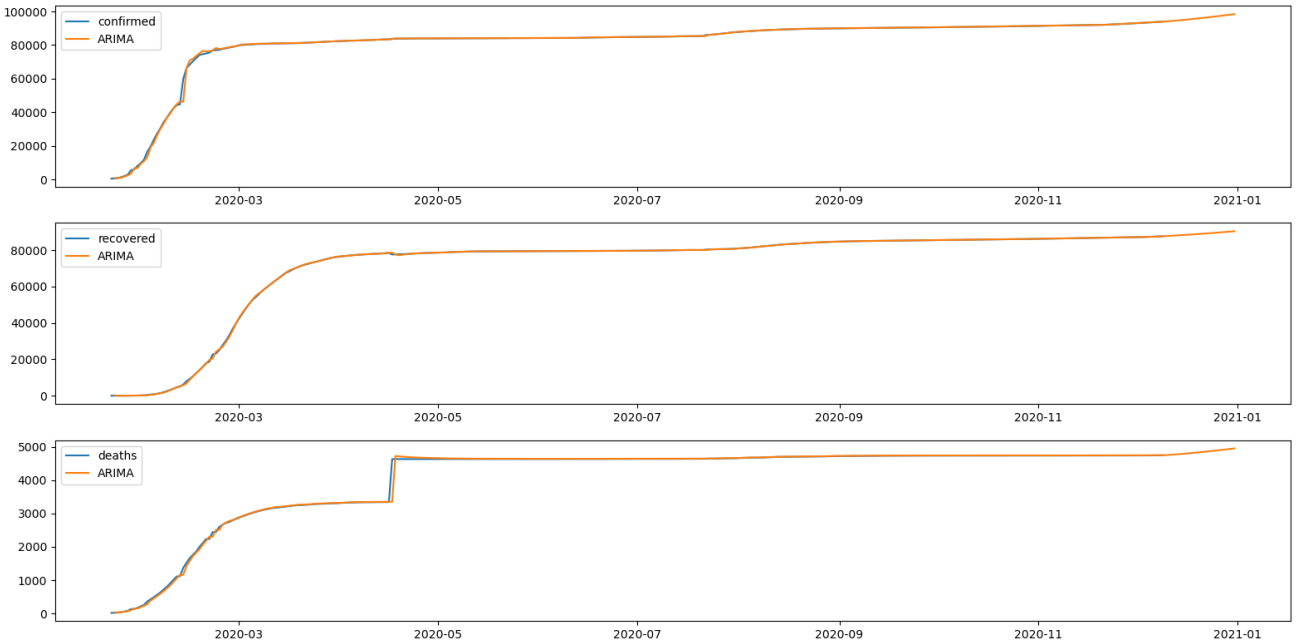

图 3-3-1

三组的参数(p, d, q)的选择如下：

confirmed: (1,0,2)

recovered: (1,0,2)

deaths: (1,0,1)




4）模型训练与数据预测

图 3-4-1

可以看到拟合程度很好。对未来七天的预测如下：

| | confirmed_pre | | recovered_pre | | deaths_pre |
|---|---|---|---|---|---|
| **2020-12-10** | 94041.291007 | **2020-12-10** | 87670.212866 | **2020-12-10** | 4754.410351 |
| **2020-12-11** | 94195.167702 | **2020-12-11** | 87770.109679 | **2020-12-11** | 4761.209900 |
| **2020-12-12** | 94356.234096 | **2020-12-12** | 87873.174916 | **2020-12-12** | 4768.381351 |
| **2020-12-13** | 94524.146896 | **2020-12-13** | 87979.330610 | **2020-12-13** | 4775.908176 |
| **2020-12-14** | 94698.579196 | **2020-12-14** | 88088.500710 | **2020-12-14** | 4783.774584 |
| **2020-12-15** | 94879.219703 | **2020-12-15** | 88200.611039 | **2020-12-15** | 4791.965483 |
| **2020-12-16** | 95065.771986 | **2020-12-16** | 88315.589246 | **2020-12-16** | 4800.466454 |

图 3-4-2