新冠疫情可视化分析及时间序列预测

主要工作：

1.数据整理
2.可视化
3.时间序列预测

关于数据:
数据来源于 github：
https://github.com/BlankerL/DXY-COVID-19-Data/releases/tag/2020.12.22
数据并不完整，且每日多次爬取，有大量重复数据。

第一部分：数据整理
首先载入数据，并转 tibble 表
对数据执行一系列操作
从全球数据筛选中国数据，筛选确诊，治愈，死亡数据
因为每天更新数次，每天有好几个数据，没有每天单独的数据
于是通过切片得到日期和时分秒，再以以日期省市为分组，筛选最大时分秒，以此得到每日最新数据
再剔除国内没有数据的日期
再以日期为组，将各省市数据加起来得到全国数据

```
nation <- DXYArea %>%
  #从全球数据筛选中国数据，筛选确诊，治愈，死亡数据
  filter(countryName=="中国") %>%
  select(updateTime, provinceName, cityName,city_confirmedCount,
         city_curedCount,city_deadCount) %>%
  #每天有好几个数据，通过切片得到日期和时分秒，
  mutate(Date=substr(updateTime,1,10)) %>%
  mutate(time= as.numeric(paste(substr(updateTime,12,13),
         substr(updateTime,15,16),substr(updateTime,18,19),sep = ""))) %>%
  #以日期省市为分组，筛选最大时分秒，以此得到每日最新数据
  group_by(Date, provinceName, cityName) %>%
  filter(time == max(time)) %>%
  #剔除国内没有数据的日期
  filter(Date!="2020-01-22" ) %>%
  filter(Date!="2020-01-23" ) %>%
  #再以日期为组，将各省市数据加起来得到全国数据
  group_by(Date) %>%
  summarise(confirmedCount = sum(city_confirmedCount,na.rm=TRUE),
            curedCount = sum(city_curedCount,na.rm=TRUE),
            deadCount = sum(city_deadCount,na.rm=TRUE)) %>%
  #数据不完整，只有前39行较准确
  head(39)
```

不过数据不完整
只有前 39 行较准确，故取前 39 行

	Date	confirmedCount	curedCount	deadCount
28	2020-02-20	74539	16689	2119
29	2020-02-21	75980	18700	2236
30	2020-02-22	76110	20960	2343
31	2020-02-23	76936	23183	2443
32	2020-02-24	77238	24880	2591
33	2020-02-25	77399	27487	2659
34	2020-02-26	77899	29944	2713
35	2020-02-27	78387	32808	2742
36	2020-02-28	78466	36012	2785
37	2020-02-29	78995	39057	2831
38	2020-03-01	79588	41909	2868
39	2020-03-02	79703	44539	2910
40	2020-03-03	79574	46875	2941
41	2020-03-04	79670	49438	2973

再挑选数据完整的 2020-2-23。以省份为组，将各市数据加起来得到各省当天数据

```
province <- DXYArea %>%
  #从全球数据筛选中国数据，筛选确诊，治愈，死亡数据
  filter(countryName=="中国") %>%
  select(updateTime, provinceName, cityName,city_confirmedCount,city_curedCount
         ,city_deadCount) %>%
  #每天有好几个数据，通过切片得到日期和时分秒，
  mutate(Date=substr(updateTime,1,10)) %>%
  mutate(time= as.numeric(paste(substr(updateTime,12,13),substr(updateTime,15,16),
                                substr(updateTime,18,19),sep = ""))) %>%
  #以日期省市为分组，筛选最大时分秒，以此得到每日最新数据
  group_by(Date, provinceName, cityName) %>%
  filter(time == max(time)) %>%
  #选择数据较完整的最新日期2020年2月23日
  filter(Date=="2020-02-23" ) %>%
  #再以省份为组，将各市数据加起来得到各省数据
  group_by(provinceName) %>%
  summarise(confirmedCount = sum(city_confirmedCount,na.rm=TRUE),
            curedCount = sum(city_curedCount,na.rm=TRUE),
            deadCount = sum(city_deadCount,na.rm=TRUE)) %>%
  mutate(level=1)
```

省份数据中缺失台湾，香港，故手动添加台湾，香港数据
再为各省确诊人数设置级别，各级别转因子

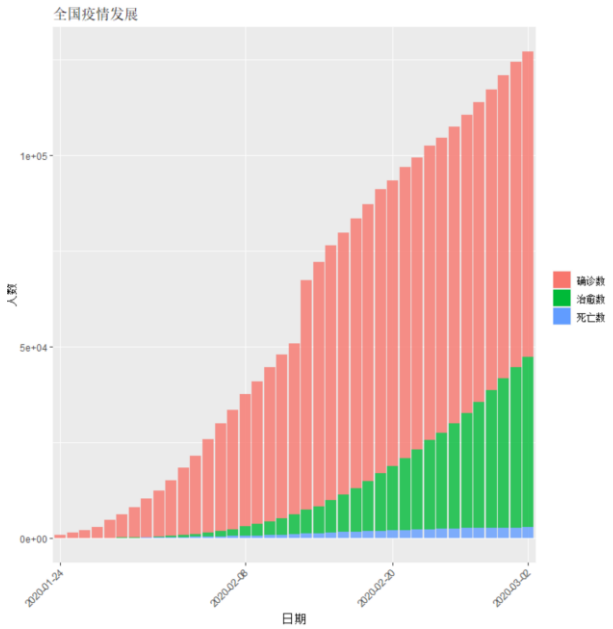
```
#手动添加台湾，香港数据
province$provinceName[27] <- "台湾省"
province$provinceName[30] <- "香港特别行政区"
province$confirmedCount[27] <- 25
province$confirmedCount[30] <- 52

#为各省确诊人数设置级别
attach(province)
province[confirmedCount <= 9, ]$level <- 1
province[confirmedCount > 9 & confirmedCount <= 99, ]$level <- 2
province[confirmedCount > 99 & confirmedCount <= 499, ]$level <- 3
province[confirmedCount > 499 & confirmedCount <= 999, ]$level <- 4
province[confirmedCount > 999 & confirmedCount <= 10000, ]$level <- 5
province[confirmedCount > 10000, ]$level <- 6
province<-province %>%
  mutate(level = factor(level, ordered = TRUE))
```

筛选 2 月 23 日湖北省各市的疫情数据

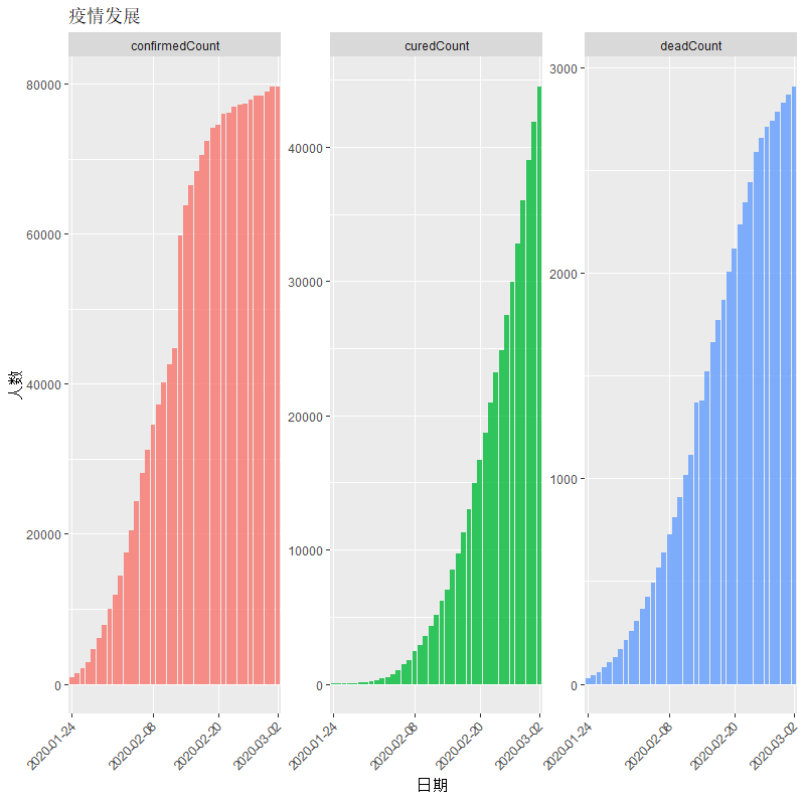
```
city <- DXYArea %>%
  #从全球数据筛选中国数据，筛选确诊，治愈，死亡数据
  filter(countryName=="中国") %>%
  select(updateTime, provinceName, cityName,city_confirmedCount,city_curedCount
         ,city_deadCount) %>%
  #每天有好几个数据，通过切片得到日期和时分秒，
  mutate(Date=substr(updateTime,1,10)) %>%
  mutate(time= as.numeric(paste(substr(updateTime,12,13),substr(updateTime,15,16),
                                substr(updateTime,18,19),sep = ""))) %>%
  #以日期省市为分组，筛选最大时分秒，以此得到每日最新数据
  group_by(Date, provinceName, cityName) %>%
  filter(time == max(time)) %>%
  #选择数据较完整的最新日期2020年2月23日
  filter(Date=="2020-02-23" ) %>%
  filter(provinceName=="湖北省")
```

第二部分：可视化分析

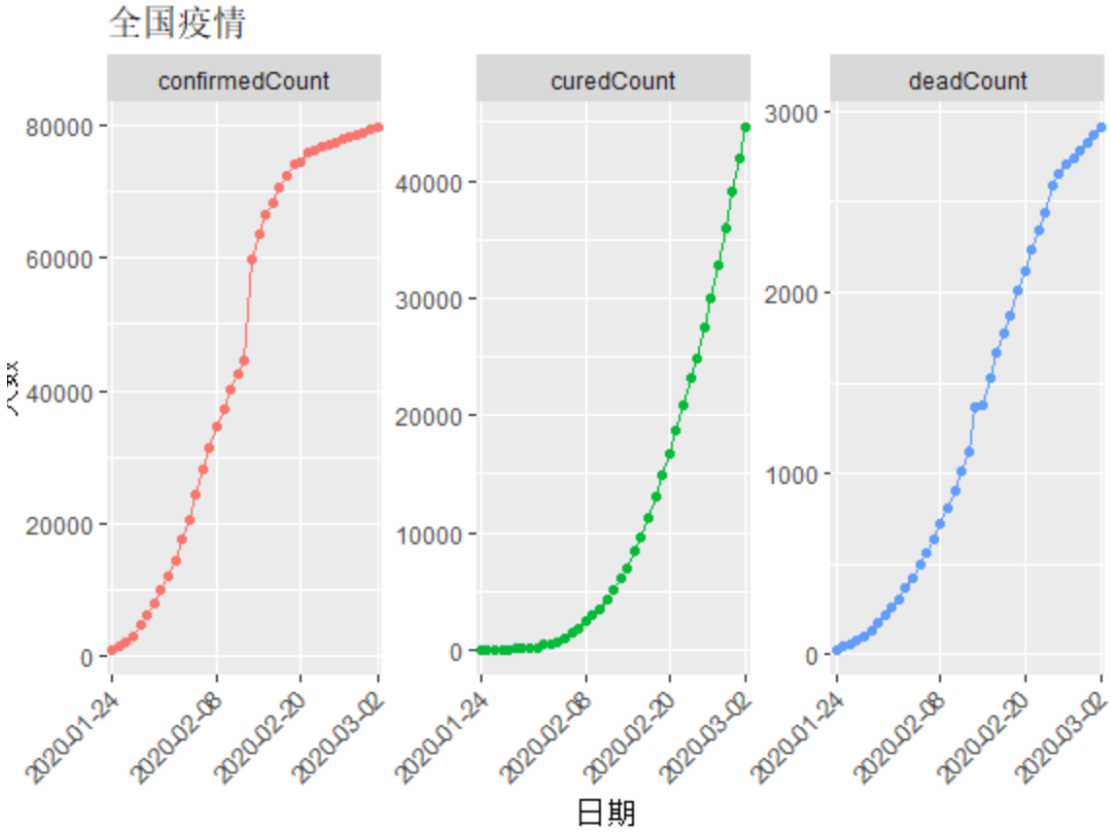
可视化部分全程使用 ggplot2 画图，下图为 2020 年 2 月 23 日各省份确诊条状图：



facet_wrap 分开的图：

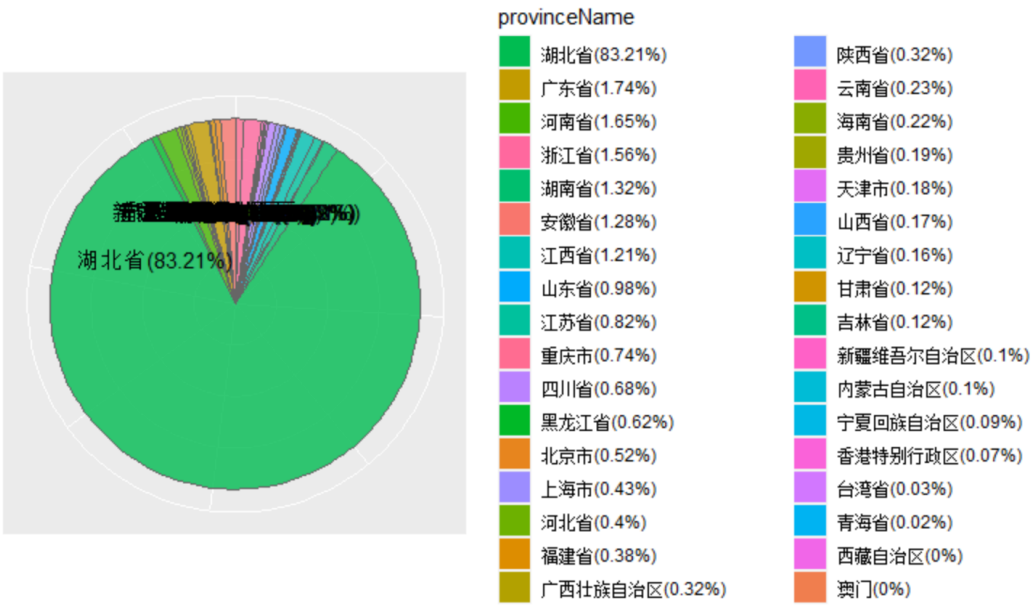


全国疫情数据折线图：



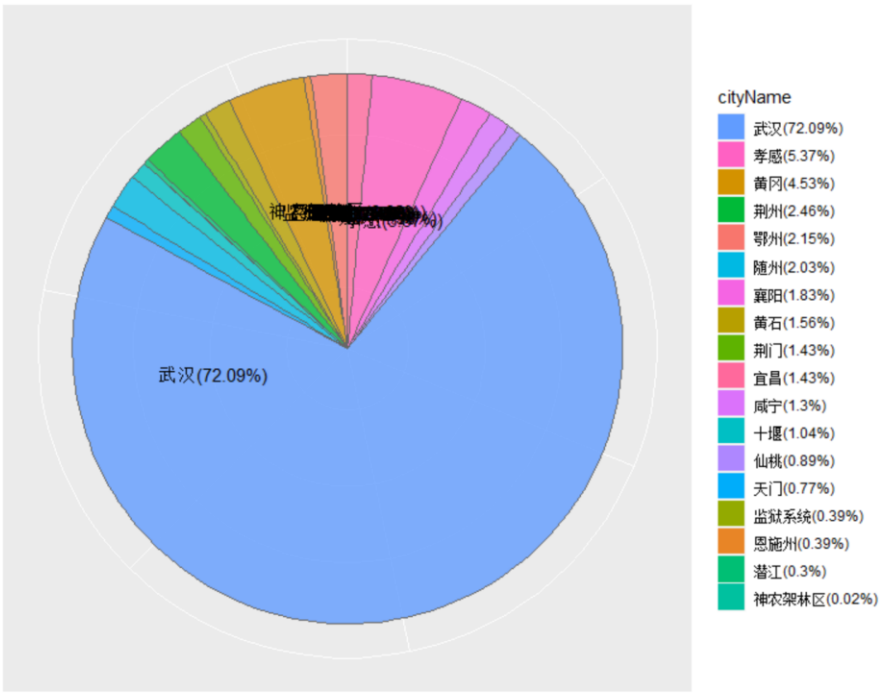
从上边的图中可以看出确诊数被控制了下来，死亡数也有所趋缓，治愈数极速增加。

2020-2-23 日各省市确诊人数占比饼状图:



从 2.23 日各省市确诊数占比中，可以看出 83%的确诊案例在湖北省，湖北的疫情严重。

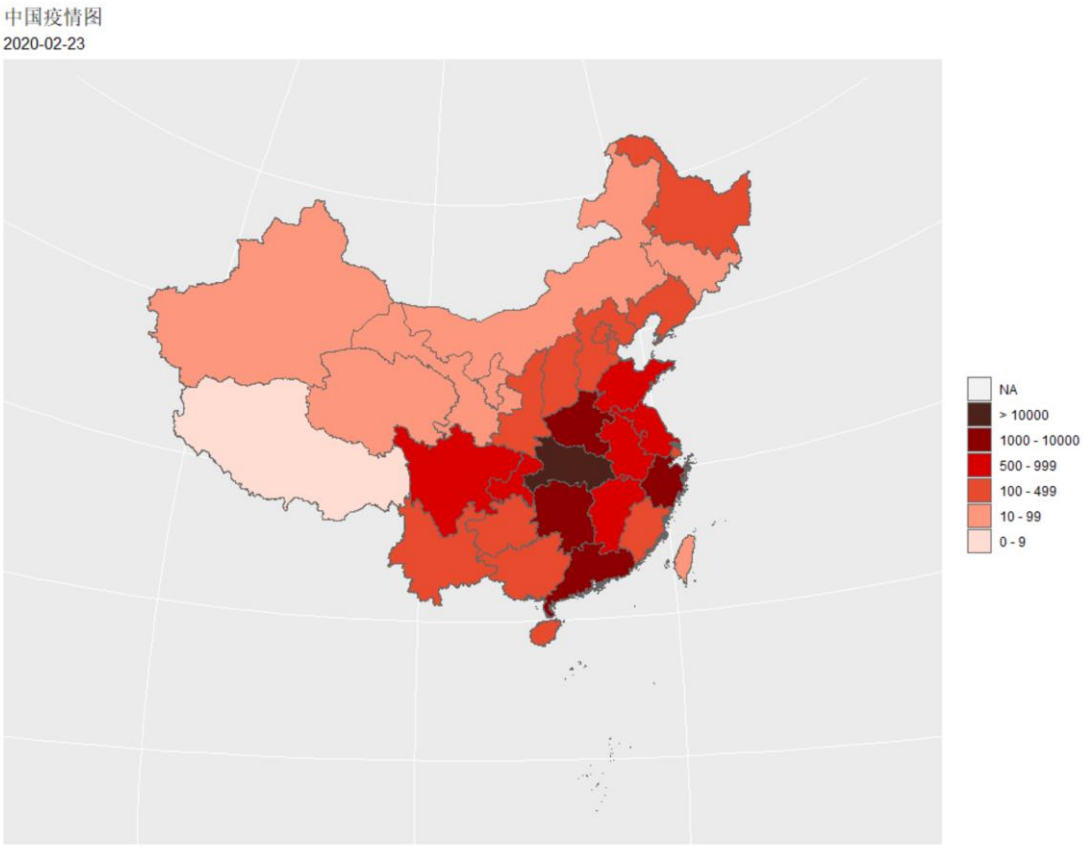
2020-2-23 日湖北省各市确诊人数占比饼状图:



从 2.23 日湖北各市确诊数占比中，可以看出湖北省 72%的确诊案例在武汉市，武汉市的疫情最严重。

2020-2-23 日中国疫情图：

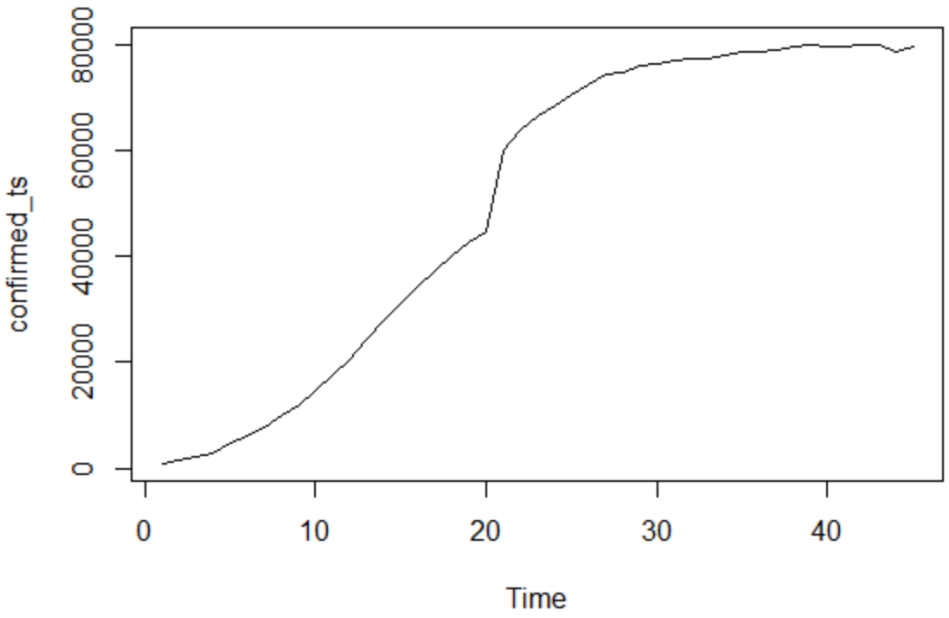
导入地图文件将疫情数据与地图数据关联整合，得到 2020-2-23 日的中国疫情地图:



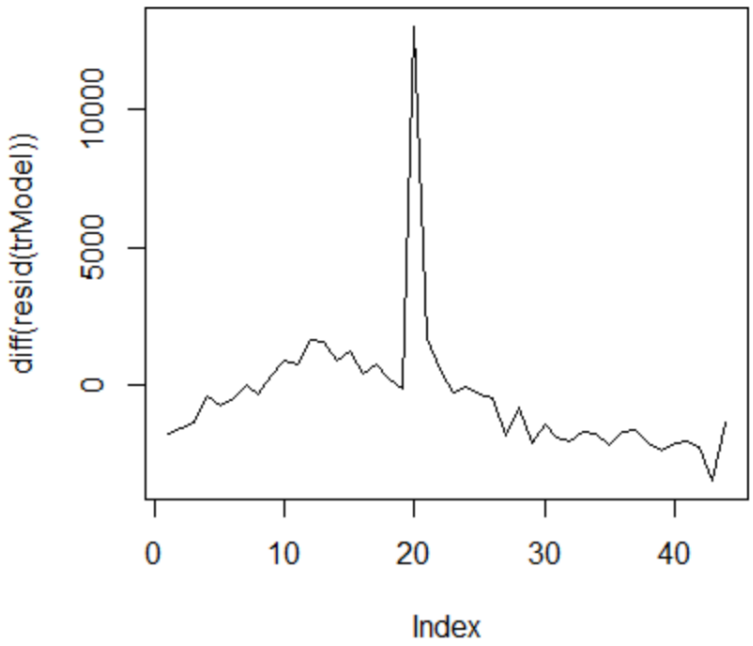
从图中可以看出，湖北疫情最严重，其次是湖南，河南，广东，江苏，34 个省份中西藏疫情状况最佳。

第三部分：时间序列预测

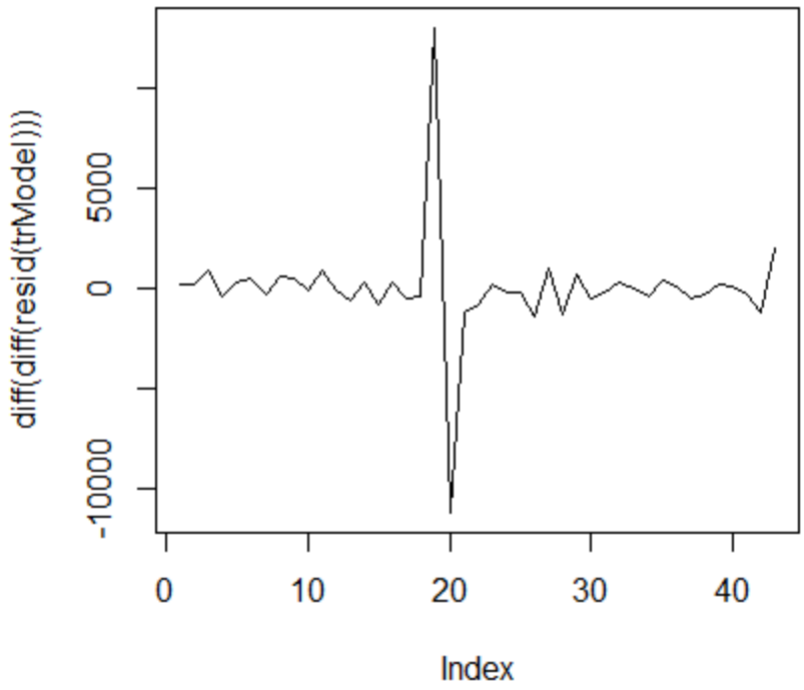
挑选数据较准确的 45 天，将确诊人数转化成时间序列
前 45 天确诊人数图



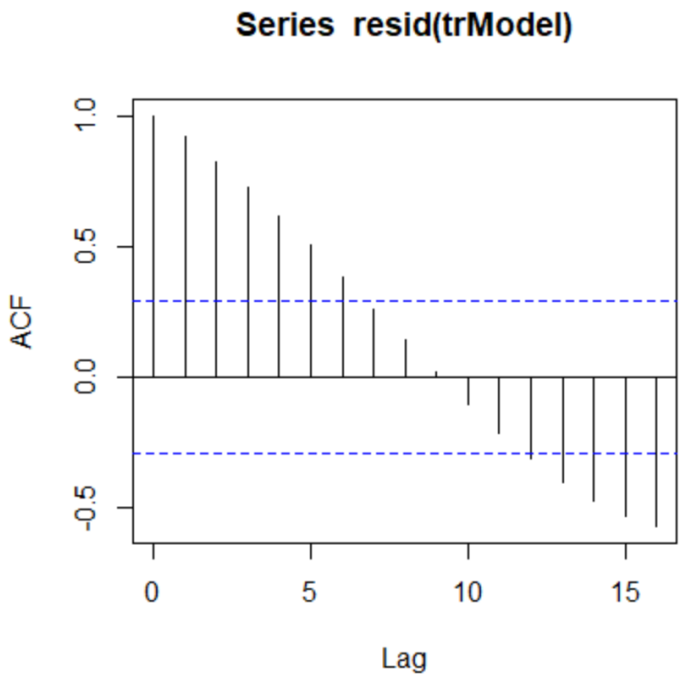
一阶差分



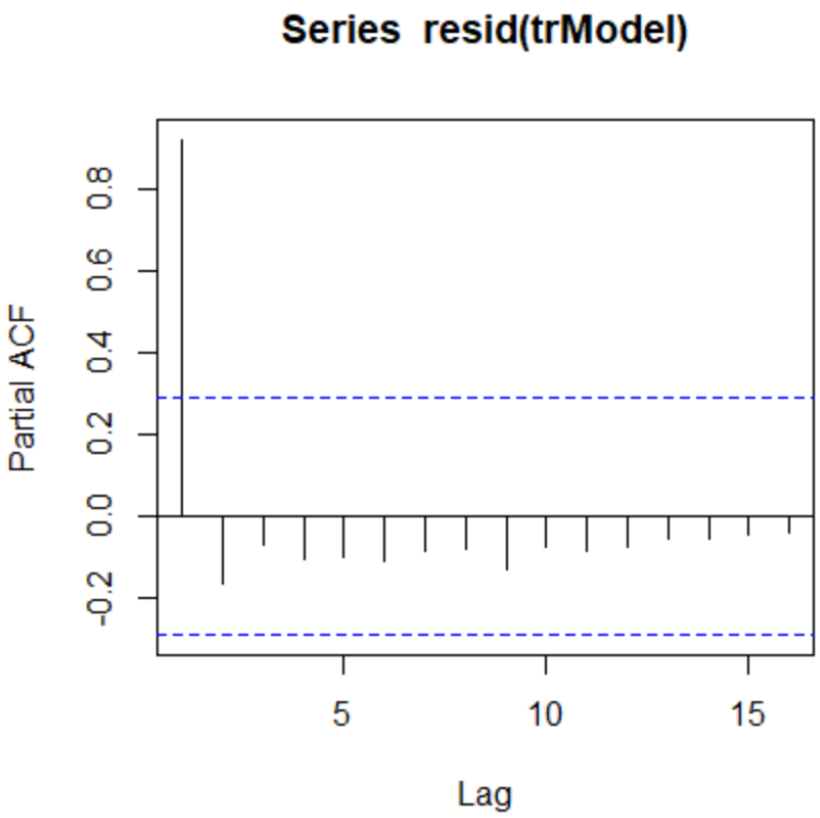
二阶差分：数据中有一个异常值，因为那日统计了所有之前未完全统计的确诊人数，新增确诊人数剧增。



ACF



PACF

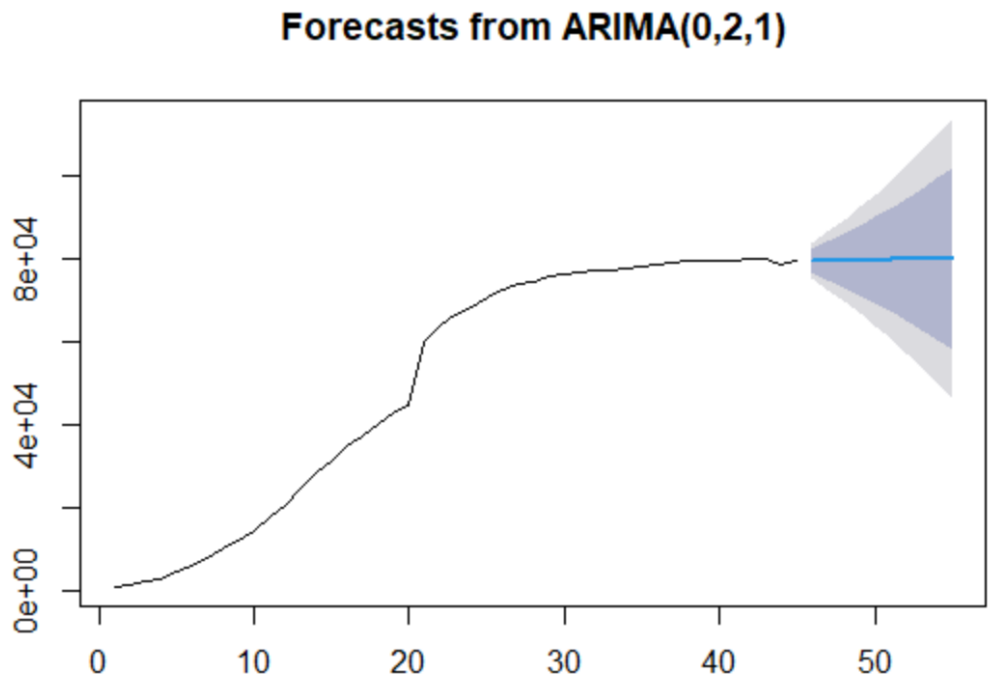


建模并总览模型:

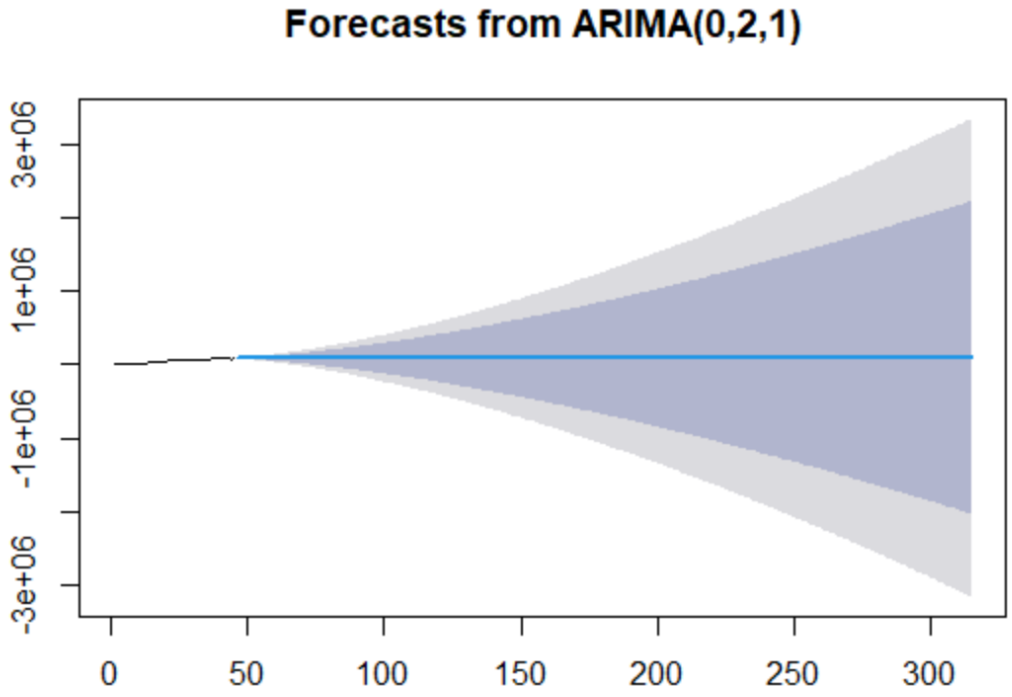
```
> summary(model)
Series: confirmed_ts
ARIMA(0,2,1)

Coefficients:
          ma1
      -0.7129
s.e.   0.1081

sigma^2 estimated as 4944624:  log likelihood=-392.26
AIC=788.52   AICc=788.82   BIC=792.04

Training set error measures:
                     ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set -46.58912 2148.251 1052.958 1.676043 3.294831 0.5694465 -0.007100158
```

该模型为 ARIMA(0,2,1)的模型

预测 10 天后疫情状况：



预测 270 天后疫情状况：



10 天后疫情状况的置信区间：

```
> forecast_10$lower[10,2]
      95%
46467.45
> forecast_10$upper[10,2]
      95%
113600.9
```

可以从结果看出，10 和 270 天的预测都基本水平，即控制住了疫情，但是置信区间特别大所以需要有更多更准确的数据才能完成较准确的预测（数据越多越准确是数据分析及预测的王道）