Time Series Analyses and Clustering for Chicago Covid-19

CS 418
Liang Liu
Seungbin Yang
Hongcheng Wu

Problem Statements

Task1: predict the weekly case rate, weekly death rate, weekly cumulative case rate in the future.

Task2: cluster the Chicago area to varies risk levels in map to have a more direct intuition.

Data Sources

Dataset: "COVID-19 Cases, Tests, and Deaths by ZIP Code" from 03/01/2020 to 11/21/2020

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2280 entries, 0 to 2279
Data columns (total 21 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   ZIP Code                               2280 non-null   object
 1   Week Number                            2280 non-null   int64
 2   Week Start                             2280 non-null   object
 3   Week End                               2280 non-null   object
 4   Cases - Weekly                         2105 non-null   float64
 5   Cases - Cumulative                     2105 non-null   float64
 6   Case Rate - Weekly                     2105 non-null   float64
 7   Case Rate - Cumulative                 2105 non-null   float64
 8   Tests - Weekly                         2250 non-null   float64
 9   Tests - Cumulative                     2280 non-null   int64
 10  Test Rate - Weekly                     2280 non-null   int64
 11  Test Rate - Cumulative                 2280 non-null   float64
 12  Percent Tested Positive - Weekly       2280 non-null   float64
 13  Percent Tested Positive - Cumulative   2280 non-null   float64
 14  Deaths - Weekly                        2280 non-null   int64
 15  Deaths - Cumulative                    2280 non-null   int64
 16  Death Rate - Weekly                    2280 non-null   float64
 17  Death Rate - Cumulative                2280 non-null   float64
 18  Population                             2280 non-null   int64
 19  Row ID                                 2280 non-null   object
 20  ZIP Code Location                      2242 non-null   object
dtypes: float64(10), int64(6), object(5)
memory usage: 374.2+ KB
```

ɔn

ɔic

04

Week Number

A sequential count of weeks, starting at the beginning of 2020.

05

Death Rate – Weekly

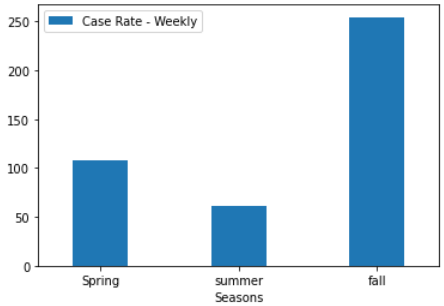
Case rate per 100,000 population in the week.

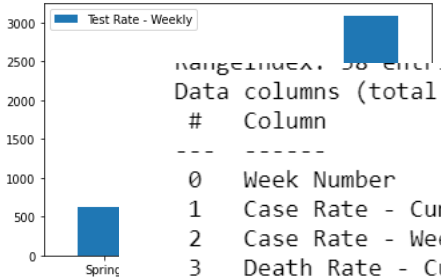
06

ZIP Code

Home ZIP Code of the cases and people tested.

Data Exploration and Data Reshape






```
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Week Number             38 non-null     int64
 1   Case Rate - Cumulative  38 non-null     float64
 2   Case Rate - Weekly      38 non-null     float64
 3   Death Rate - Cumulative 38 non-null     float64
dtypes: float64(3), int64(1)
memory usage: 1.3 KB
```

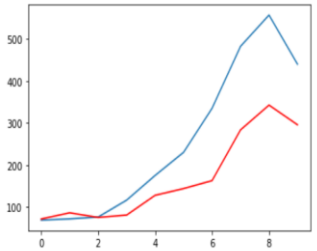
result

1: Some relationship between temperature and Covid-19.
2. More people get tested
3. Death rate decreases

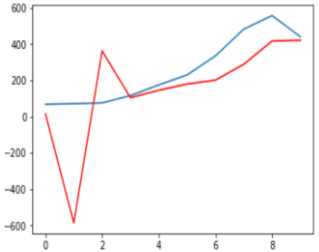
Reshape the weekly data based on zip to the whole Chicago

Time Series Analyses on Chicago Weekly Case Rate

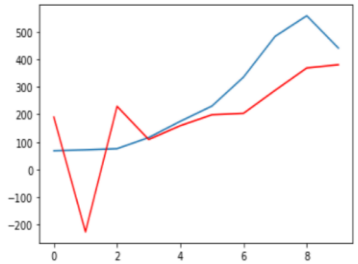
01 MA model



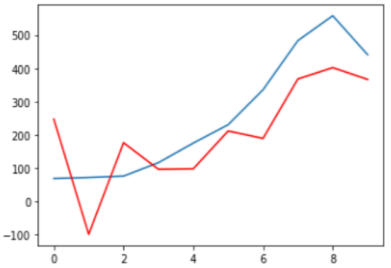
02 AR model



03 ARMA model



04 ARIMA mode



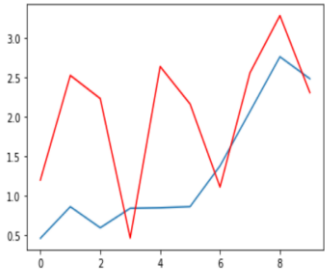
Use autocorrection and partial autocorrection to get the p and q.
Blue line: real weekly case rate
Red line: predicted weekly case rate

result:

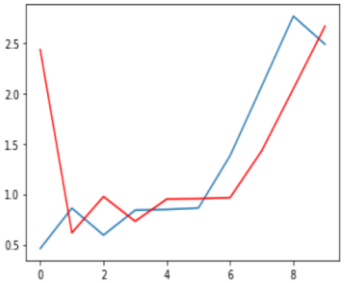
MSE: 14763 Correlation: 0.979 MA
MSE: 59634 Correlation: 0.601 AR
MSE: 22273 Correlation: 0.687 ARMA
MSE: 14204 Correlation: 0.790 ARIMA

Time Series Analyses on Chicago Weekly Death Rate

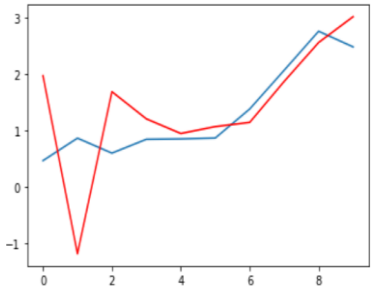
01 MA model



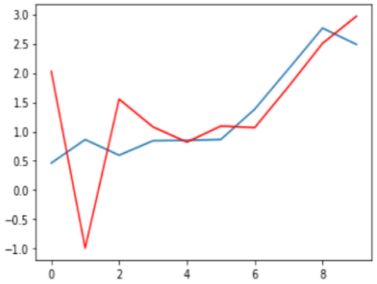
02 AR model



03 ARMA model



04 ARIMA mode



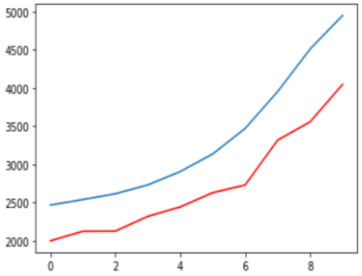
Use autocorrection and partial autocorrection to get the p and q.
Blue line: real weekly case rate
Red line: predicted weekly case rate

result:

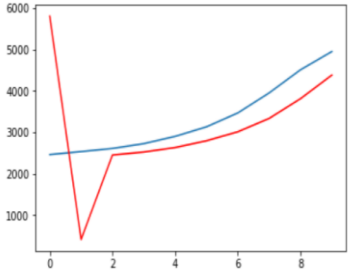
MSE: 1.164 Correlation: 0.506 MA
MSE: 0.524 Correlation: 0.531 AR
MSE: 0.831 Correlation: 0.571 ARMA
MSE: 0.737 Correlation: 0.582 ARIMA

Time Series Analyses on Chicago Weekly Cumulative Case Rate

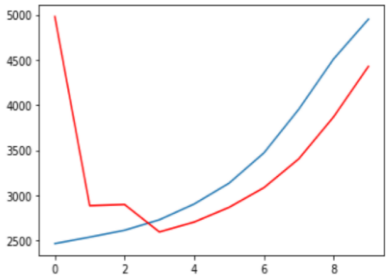
01 MA model



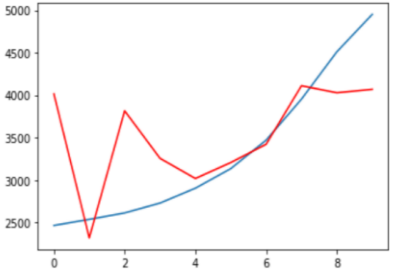
02 AR model



03 ARMA model



04 ARIMA mode

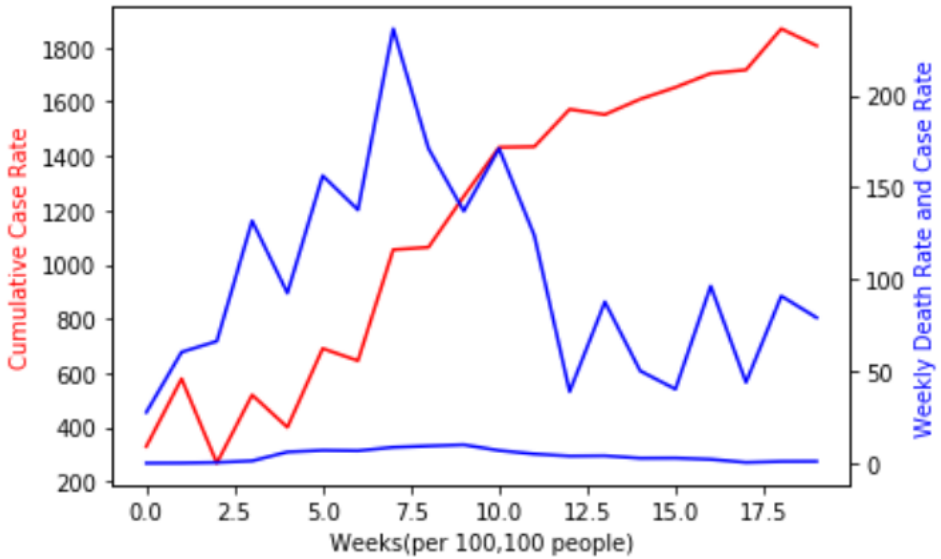


Use autocorrection and partial autocorrection to get the p and q.
Blue line: real weekly case rate
Red line: predicted weekly case rate

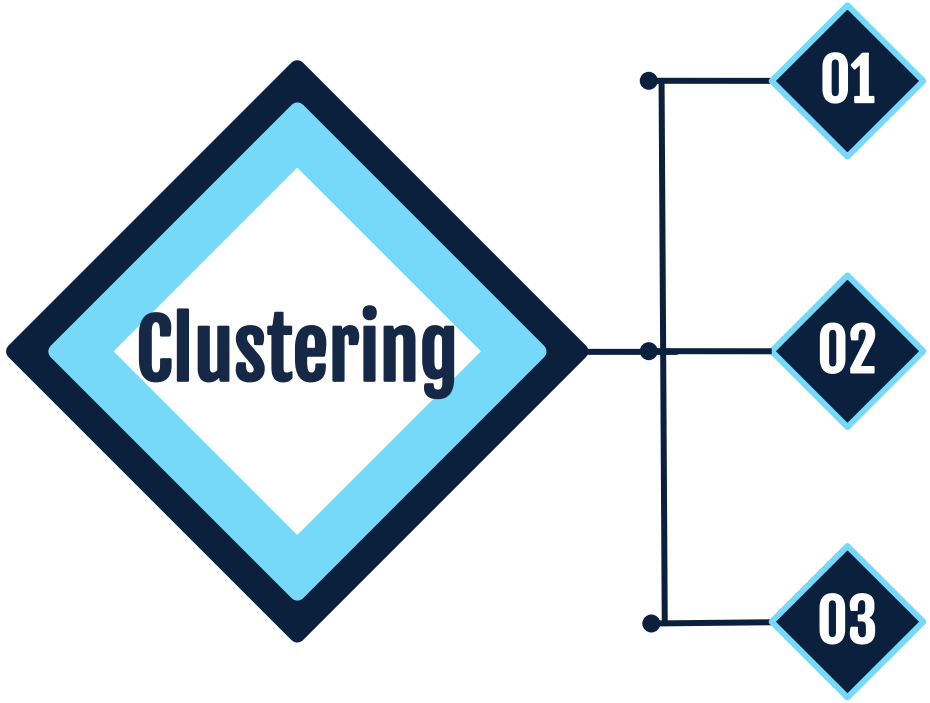
result:

MSE: 395562 Correlation: 0.994 MA
MSE: 1727847 Correlation: 0.353 AR
MSE: 778487 Correlation: 0.390 ARMA
MSE: 522733 Correlation: 0.559 ARIMA

Predict Situation in next 20 weeks



we can see the weekly death rate is always low, weekly case rate keep increase in the next 7 weeks then go down, cumulative case rate increase also get slower after 7 weeks in the future.



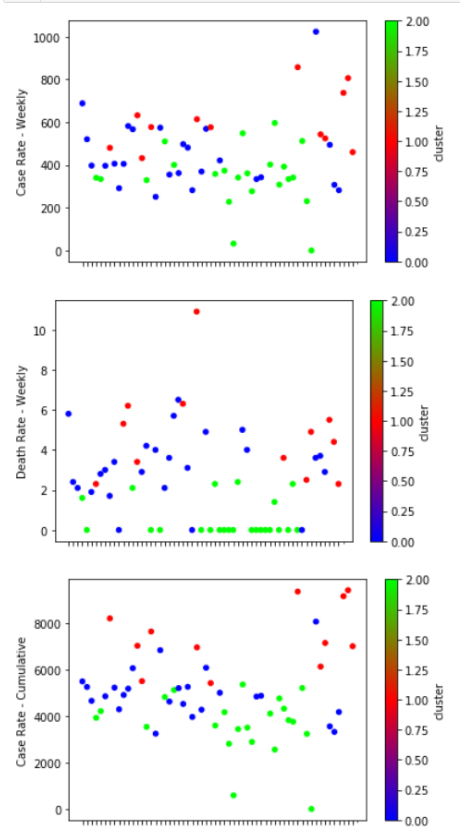
K-means clustering

Single linkage hierarchical clustering

Complete linkage hierarchical clustering

Based on the latest case rate. Death rate and Cumulative case rate

How to define risk level

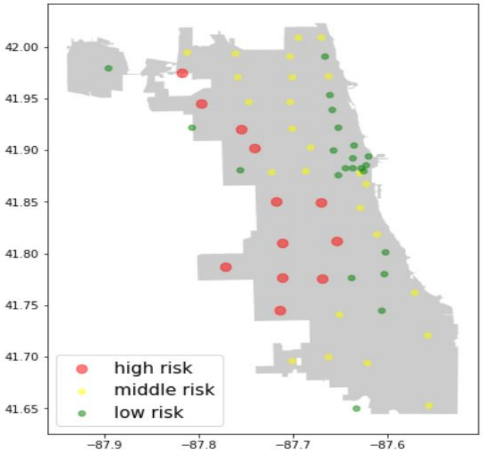


We define the cluster which has the highest weekly case rate, highest death rate, highest weekly cumulative case rate to high risk. Same way to define other cluster risk level.

Time Series Analyses on Chicago Weekly Cumulative Case Rate

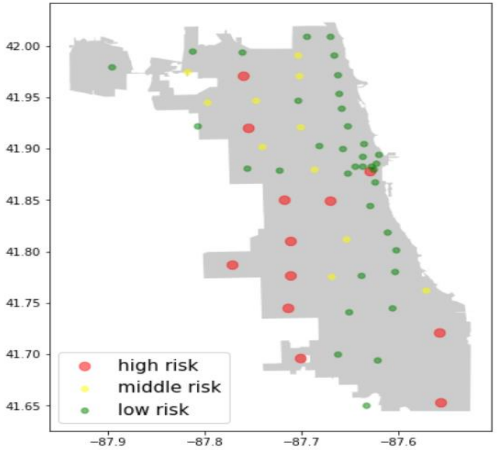
01

K-means



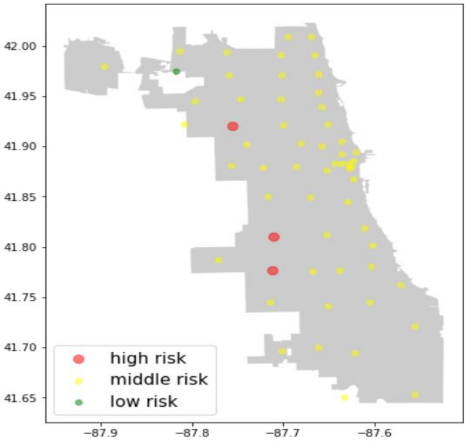
02

Single linkage



03

Complete linkage



the area close to downtown has less risk than in suburbs

Thanks for watching