CS418 Final Project:

Chicago Covid-19 Analysis and Prediction

Fall 2020

Group Members:

Liang Liu **(661748538)**

Seungbin Yang **(675230833)**

Hongcheng Wu **(672171388)**

Submitted Date: 29 Nov 2020

Problem Selection

More than 100 years ago, in 1918, the Spanish flu swept the world, causing pain that is unforgettable. Today, more than 100 years later, COVID-19 became another challenge for human survival. But we will use data analytics and data prediction transforming passive prevention into active prevention to overcome the pandemic.

COVID-19 is one of the most important problems in the real world. Now, according to the WHO, over 9.2 million people have confirmed cases and over 230,000 deaths in the United State. Also Illinois has the fifth-largest number of confirmed cases in all U.S. states. Since we currently live in Chicago, which belongs to Illinois, we will study about the COVID-19 by ZIP code in Chicago.

We will use the time series analysis to analyze the trend of past Chicago weekly case rate, weekly death rate and weekly cumulative case rate from 03/01/2020 to 11/21/2020 then get the best model to predict next 20 weeks situation. And we will also use k-means clustering and varies hierarchical clustering methods to cluster Chicago to three different danger levels based on the most recently weekly case rate, weekly death rate, and weekly cumulative case rate, determine the risk levels then visualize the result into Chicago map.

Data Collection

We will use the dataset "COVID-19 Cases, Tests, and Deaths by ZIP Code", downloaded from Chicago Data Portal, here we only describe the data we will use.

Column Name	Description	Type
ZIP Code	Home ZIP Code of the cases and people tested.	Plain Text
Case Rate - Weekly	Case rate per 100,000 population in the week.	Number
Week Number	A sequential count of weeks, starting at the beginning of 2020. These numbers are aligned to CDC MMWR weeks.	Number
Case Rate - Cumulative	Total case rate per 100,000 population through the week.	Number
Death Rate - Weekly	Death rate per 100,000 population in the week.	Number
ZIP Code Location	A point within the ZIP Code to allow for geographic analysis. The precise point shown has no other meaning.	Point

URL: https://data.cityofchicago.org/Health-Human-Services/COVID-19-Cases-Tests-and-Deaths-by-ZIP-Code/yhhz-zm2v

Data Preparation

1.Deal with the missing values and data description

```
1 data.isnull().sum()
```

```
ZIP Code                                0
Week Number                             0
Week Start                              0
Week End                                0
Cases - Weekly                        175
Cases - Cumulative                    175
Case Rate - Weekly                    175
Case Rate - Cumulative                175
Tests - Weekly                         30
Tests - Cumulative                      0
Test Rate - Weekly                      0
Test Rate - Cumulative                  0
Percent Tested Positive - Weekly        0
Percent Tested Positive - Cumulative    0
Deaths - Weekly                         0
Deaths - Cumulative                     0
Death Rate - Weekly                     0
Death Rate - Cumulative                 0
Population                              0
Row ID                                  0
ZIP Code Location                      38
dtype: int64
```

```
1 data.dtypes
```

```
ZIP Code                                 object
Week Number                               int64
Week Start                               object
Week End                                 object
Case Rate - Weekly                      float64
Case Rate - Cumulative                  float64
Test Rate - Weekly                        int64
Test Rate - Cumulative                  float64
Percent Tested Positive - Weekly        float64
Percent Tested Positive - Cumulative    float64
Death Rate - Weekly                     float64
Death Rate - Cumulative                 float64
Population                                int64
Row ID                                   object
ZIP Code Location                        object
dtype: object
```

There are 21 variables, int64(6), float64(10), object (5), and there are several null values. We changed Nan value to 0, since most of missing value appear at the beginning of record, and we also dropped some variables that we don't use for analyzing.

2. Transform weekly zip data to monthly Chicago data for season analyze

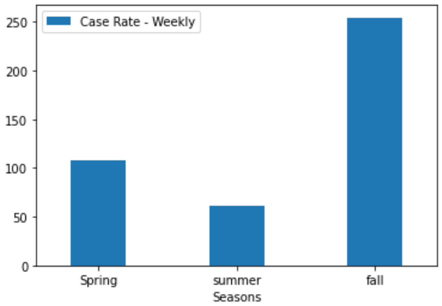
In order to have a better intuition about how season change affect Covid-19. We create new variable "Month" according to the variable "Week Start"

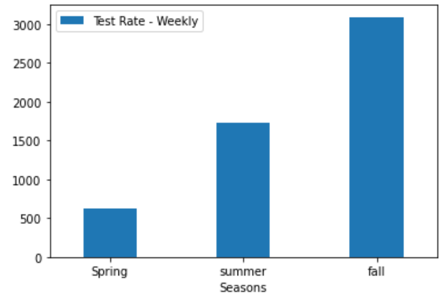
Month	Week Number	Case Rate - Weekly	Case Rate - Cumulative	Test Rate - Weekly	Test Rate - Cumulative	Percent Tested Positive - Weekly	Percent Tested Positive - Cumulative	Death Rate - Weekly	Death Rate - Cumulative	Population
3	12.0	40.036667	70.231667	160.083333	294.595667	0.127333	0.110333	1.057000	1.304000	46230.216667
4	16.5	161.366667	578.600000	625.075000	2164.183750	0.231250	0.235417	7.789167	23.430000	46230.216667
5	21.0	123.286667	1299.659000	1085.080000	6540.155333	0.128000	0.193333	6.966333	60.476333	46230.216667
6	25.5	44.137500	1594.243333	1417.670833	12076.105000	0.033750	0.141250	2.509167	78.409583	46230.216667
7	29.5	64.758333	1824.582083	1824.520833	18972.335833	0.034167	0.105000	0.773750	83.393333	46230.216667
8	34.0	74.950000	2146.974333	1918.960000	27376.599333	0.038667	0.088333	0.682667	86.475667	46230.216667
9	38.5	68.020833	2461.683333	2058.754167	36224.480417	0.027917	0.077083	0.590417	89.183333	46230.216667
10	42.5	210.579167	3008.524583	3078.204167	46858.276667	0.076250	0.073333	0.983333	92.429167	46230.216667
11	46.0	485.350000	4396.642778	4131.605556	59958.565556	0.131111	0.090000	2.437778	98.848889	46230.216667

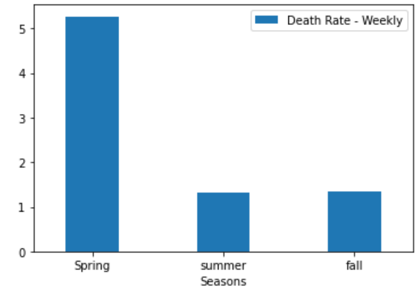
3. Transform weekly zip data to weekly Chicago data for time series analyse and clustering

Data Exploration.

We suppose Month 3-5 is Spring, 6-8 is Summer, and 9-11 is Fall, since data is start on March 1st, and end on Nov 21st. then visualize the data







For the weekly case rate, we can see that autumn has the highest case rate than other seasons. summer has the lowest. We doubt there exist some relationship between temperature and Covid-19.

For the weekly test rate, we can see that people are getting more tests over time. It was a little over 500 per 100,000 in autumn there are 3,000 per 100,000 people get tested.

For about death rate, spring is highest rate than other seasons. and summer and autumn has similar rate.

We calculate the p-value of weekly case rate between different seasons to figure out whether it is possible the case rate of different seasons is equal.

```
[statistic, pvalue] = st.ttest_ind(spring['Case Rate - Weekly'],summer['Case Rate - Weekly'],equal_var = False)
print(pvalue*2)
```

```
0.6384281788527542
```

```
[statistic, pvalue] = st.ttest_ind(summer['Case Rate - Weekly'],fall['Case Rate - Weekly'],equal_var = False)
print(pvalue*2)
```

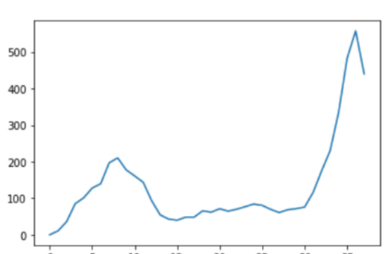
```
0.5094128255454738
```

```
[statistic, pvalue] = st.ttest_ind(spring['Case Rate - Weekly'],fall['Case Rate - Weekly'],equal_var = False)
print(2*(pvalue))
```

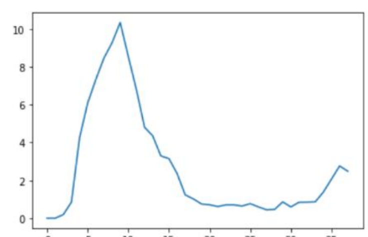
```
0.7098481982179874
```

All P-value are far above 0.05, so there is no sufficient data to reject the null hypothesis that Case Rate by month of varies seasons are equal. In other words, we fail to reject the null hypothesis.

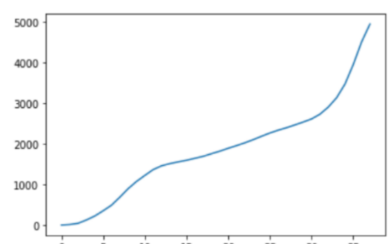
We also plot the Chicago weekly case rate, weekly death rate and weekly cumulative case rate for an advanced analysis.



Chicago Weekly case rate



Chicago weekly death rate



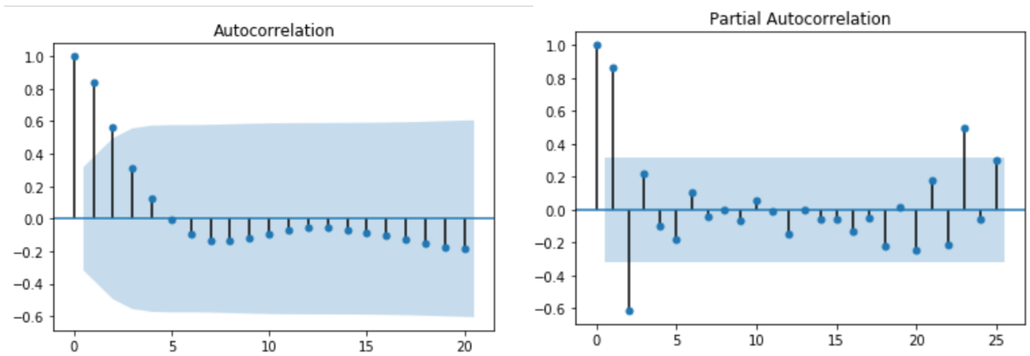
Chicago Cumulative case rate

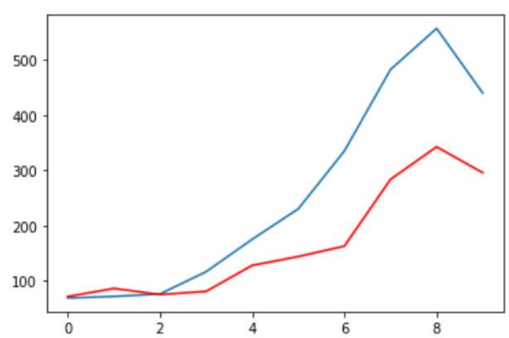
Data Modeling.

1: time series analysis on Chicago weekly case rate, death rate and cumulative case rate.

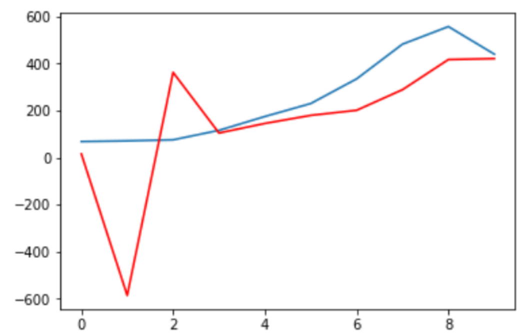
We use Autocorrelation plot and partial Autocorrelation plot to get the best order of AR part and MA part, then partition the whole data to about 75% training set and 25% test set, use varies time series model trained by training set to test on the test set in order to get the best model.

Time series analyse of Chicago weekly case rate

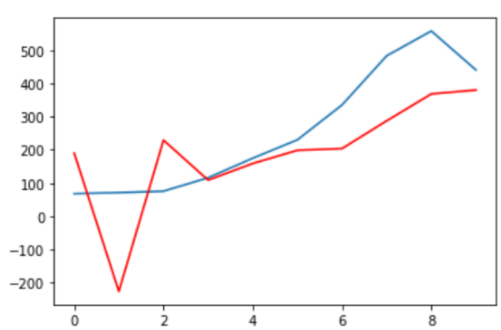




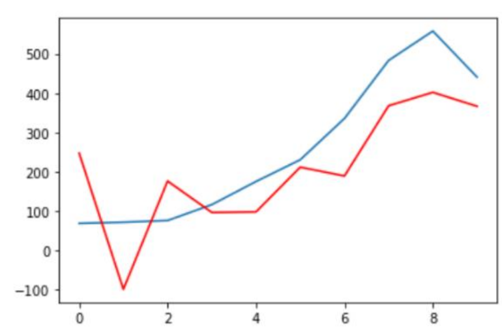
MA model (bule real line, red predict line)



AR model(bule real line, red predict line)



ARMA (bule real line, red predict line)



ARIMA(bule real line, red predict line)

result:

MSE: 14763 Correlation: 0.979 MA

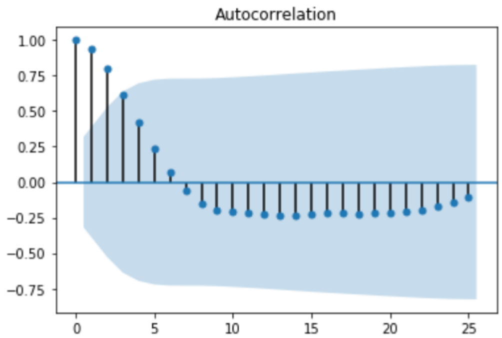
MSE: 59634 Correlation: 0.601 AR

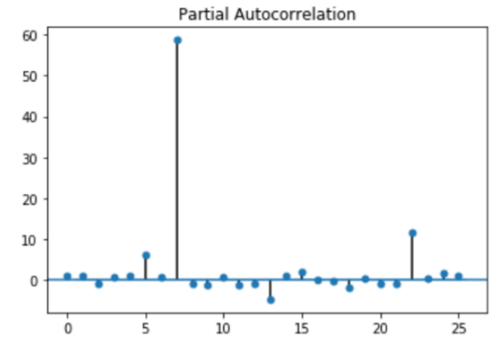
MSE: 22273 Correlation: 0.687 ARMA

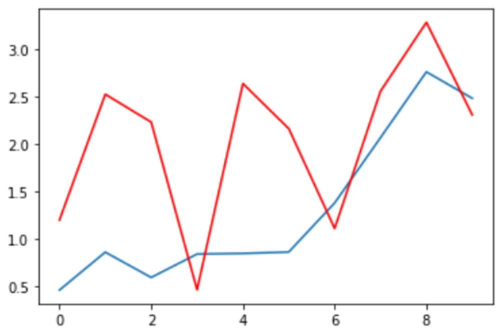
MSE: 14204 Correlation: 0.790 ARIMA

MA model has much lower MSE and higher Correlation, we choose MA as our model to predict weekly case rate

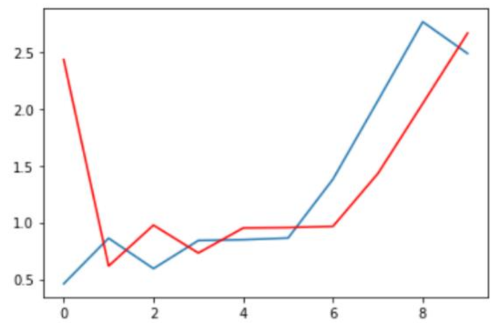
Time series analyse of Chicago weekly death rate



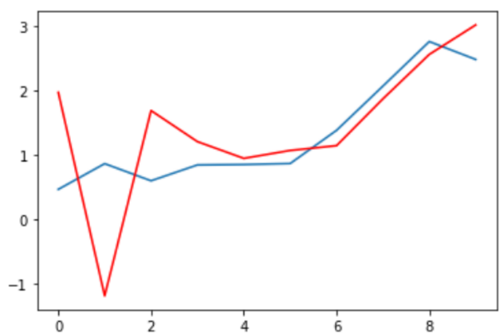




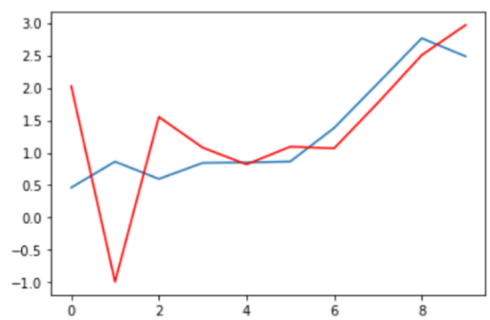
MA model (bule real line, red predict line)



AR model(bule real line, red predict line)



ARMA (bule real line, red predict line)



ARIMA (bule real line, red predict line)

result:

MSE: 1.164 Correlation: 0.506 MA

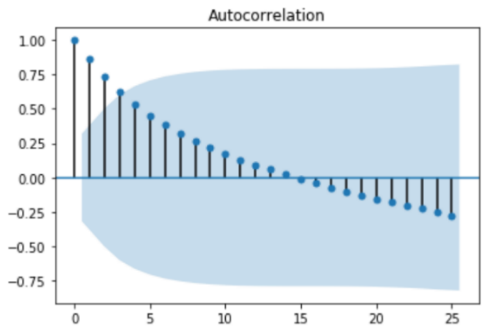
MSE: 0.524 Correlation: 0.531 AR

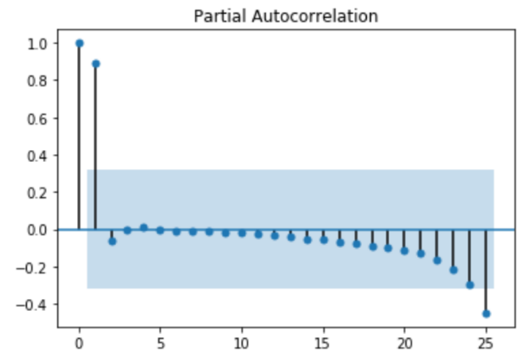
MSE: 0.831 Correlation: 0.571 ARMA

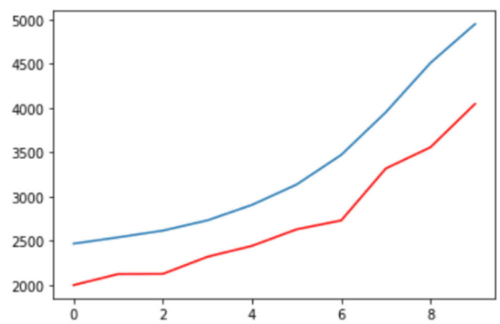
MSE: 0.737 Correlation: 0.582 ARIMA

Since they have similar correlation ratio, but AR model has much lower MSE, we choose AR as our model to predict weekly death rate

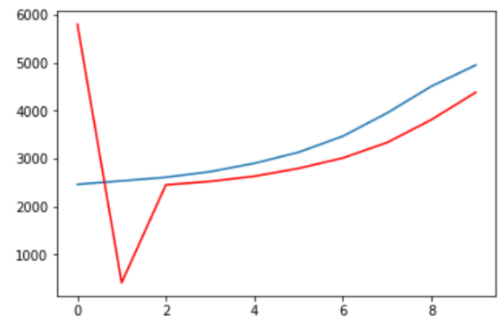
Time series analyse of Chicago cumulative weekly case rate



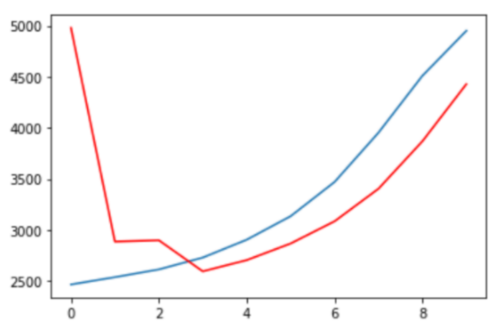




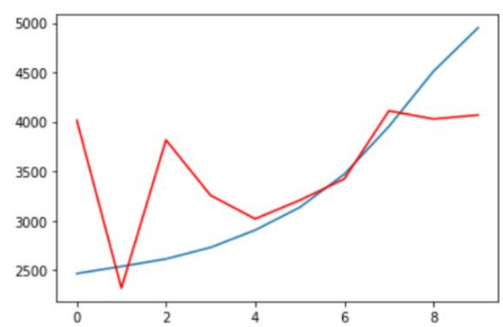
MA model (bule real line, red predict line)



AR model(bule real line, red predict line)



ARMA (bule real line, red predict line)



ARIMA (bule real line, red predict line)

result:

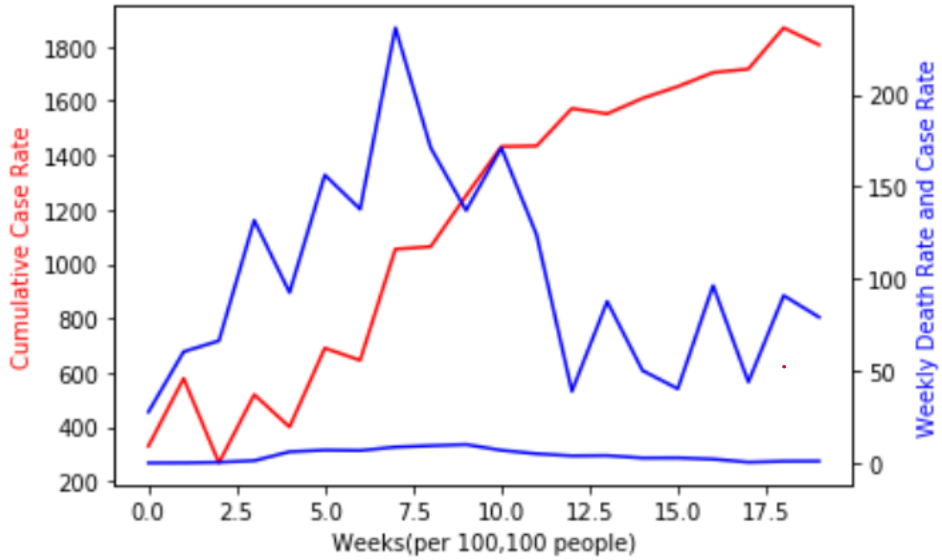
MSE: 395562 Correlation: 0.994 MA

MSE: 1727847 Correlation: 0.353 AR

MSE: 778487 Correlation: 0.390 ARMA

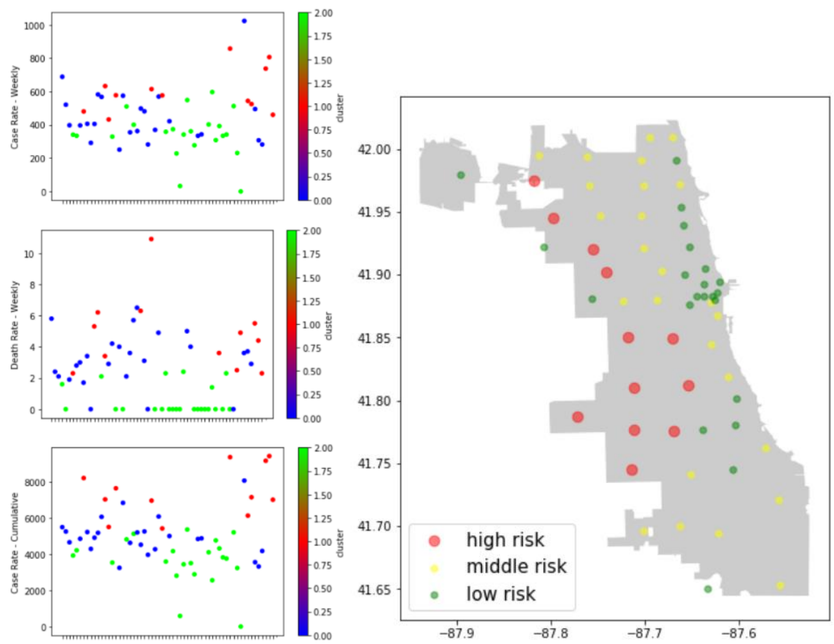
MSE: 522733 Correlation: 0.559 ARIMA

MA model has much lower MSE and higher Correlation, we choose MA as our model to predict cumulative case rate.

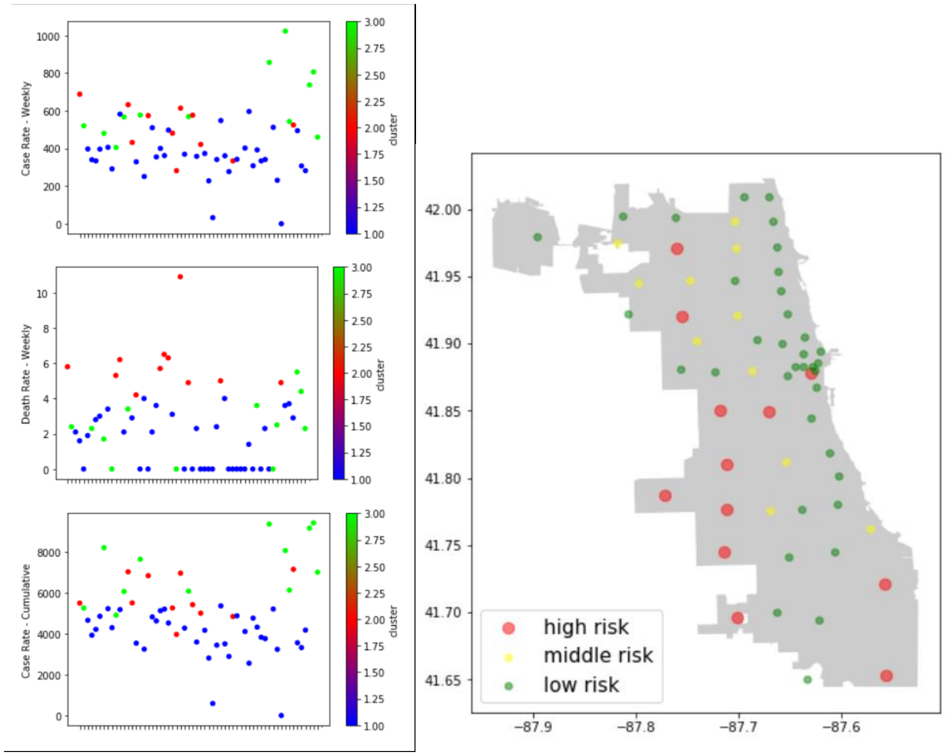


Use the best model to predict next 20 weeks Chicago weekly case rate, death rate and cumulative case rate. From figure, we can see the weekly death rate is always low, weekly case rate keep increase in the next 7 weeks then go down, cumulative case rate increase also get slower after 7 weeks in the future.

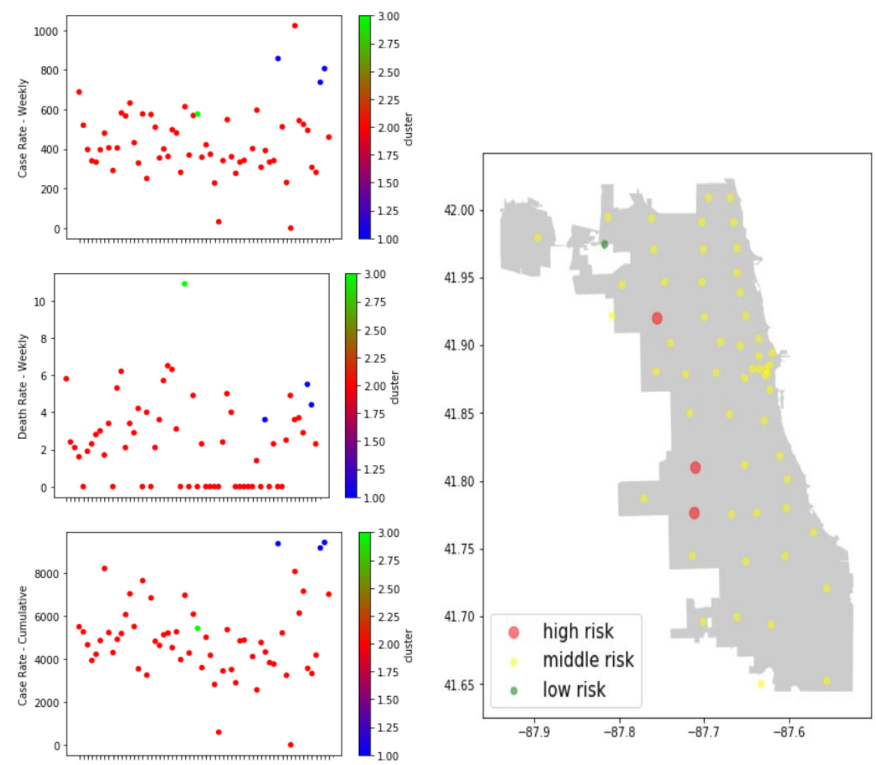
2.clustering Chicago area into three risk levels



K-means clustering method cluster Chicago area based on weekly case rate, death rate and cumulative case rate, from left figure we determine cluster 1 to high risk, cluster 0 to middle risk and cluster 2 to low risk.



Single linkage hierarchical clustering method cluster Chicago area based on weekly case rate, death rate and cumulative case rate, from left figure we determine cluster 3 to high risk, cluster 2 to middle risk and cluster 1 to low risk.



complete linkage hierarchical clustering method cluster Chicago area based on weekly case rate, death rate and cumulative case rate, from left figure we determine cluster 2 to high risk, cluster 3 to middle risk and cluster 1 to low risk.

We can see the area close to downtown has less risk than in suburbs