The Prediction of the Development on the Novel Coronavirus (COVID-19) in Hong Kong

DENG Xindi 18081072d

Abstract

In this article, I made the use of both plot and mathematical analysis to predict the development on the Novel Coronavirus (COVID-19) in Hong Kong. The prediction is focus on two sub-questions. The first question is the prediction towards the features of susceptible population, and the second question is prediction of daily additional cases in the third wave of outbreaks.

In the first question, I discussed whether age and gender are the features that influence the infection rate, and obtained the result that we could see obviously different by using age-disaggregated data. In the second question, I used the curve or normal distribution to fit with the line of daily additional confirmed cases. According to the analysis towards the first two waves, I predicted the rough end date of the third wave of outbreak.

Introduction

Since January 23rd in 2020，the first confirmed case of COVID-19 was recorded in Hong Kong, the novel coronavirus has been spread for eleven months till these days. According to the official data supported online, I drew the line chart of daily additional confirmed cases (in Hong Kong) shown below via R language.

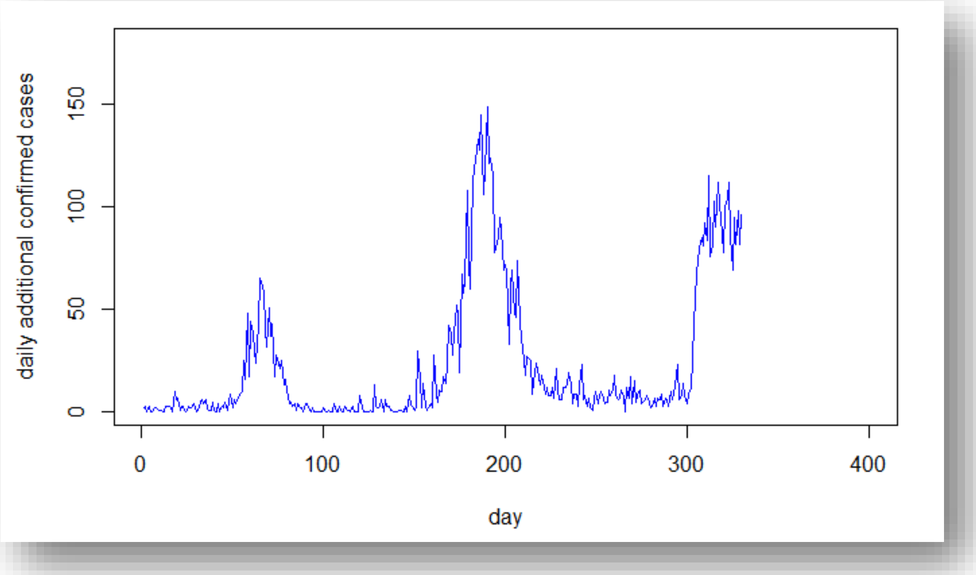


Figure 1. Daily additional confirmed cases on each day

According to the chart, we could notice that Hong Kong is experiencing the third wave of outbreaks. Two of the problems that people care about the most is that which kinds of people will be more possible to be infected, and how the number of daily additional confirmed cases will change in the near future. These two questions are the main topics discussed in this paper and they could be collectively referred to as the prediction of the development on the Novel Coronavirus (COVID-19) in Hong Kong.

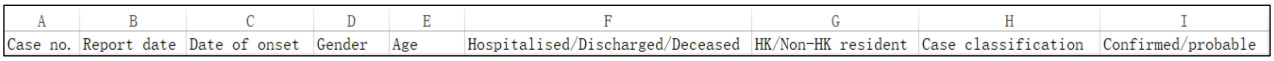
For the first question, I did the analysis on the features of susceptible population. It is reported that the "children, particularly those younger than 10 years old, may be less susceptible and contagious than adults." (Fontanet, A., Cauchemez, S. 2020). Thus, I conjecture that age may be one of the features that influences the infection rate. Another report claimed that more men are dying from this disease than women in the case of the ongoing COVID-19 pandemic (Gausman et al. 2020). Thus, we may have some new findings in the sex-disaggregated data. According to these two points, I split the data based on the age and the gender respectively, and analysis the number of disaggregated daily additionally confirmed cases. The main conclusion is older people have more possibility to be infected in the near future in Hong Kong, and the gender differences could not be obviously observed.

Turning to the second question, I did the development prediction to the third wave of outbreaks according to the first two waves. I first did the analysis towards the first two waves. According to the normal logic, the additional confirmed cases should resemble a bell curve with a normal distribution during the outbreaks. Because the daily additional cases would keep increasing in the beginning, and after people take some effective actions, the increasing data would climb to the peak and then begin to decrease. To verify this assumption, in the first step, I did the image fitting and just made the judgment just by my naked eyes. In this step, the simulation seems to perform pretty well. In the second step, I tested the goodness fit to the normal distribution. In the first outbreak, the data approximate a normal distribution with 5% significant level. However, in the second outbreak, the sum of χ^2 is so large that it is easily to reject the null hypothesis (the data approximate a normal distribution). The results seems to be different from the image fitting. After the observation, I found that image fitting could ignore some local extreme values (mainly occurred in the latter part of period) but just focus on the main values. While in the goodness-of-fit test, the extreme points could make huge influence to χ^2. Combined with multiple factors and using the outcome in first two waves, I did the prediction of the third wave of outcomes. The main result is that the number of additional confirmed cases will begin to become stable roughly in early January in 2021.

Data Sources

I downloaded the data relating to the Novel Coronavirus (COVID-19) in Hong Kong from https://data.gov.hk, which is a website coordinated by the Office of the Government Chief Information Officer (OGCIO) for free distribution of public information to the general public. The website for download is https://data.gov.hk/en-data/dataset/hk-dh-chpsebcddr-novel-infectious-agent/resource/24f08145-6c16-4de9-83a0-ee52b67eeedb (Data.GOV.HK. 2020).

The downloaded data is stored in ".csv" file with 9 columns and 7900 rows (except the header). The file records all confirmed cases from January 23 to December 17 in 2020. Each row represents to one confirmed cases with features shown below.



A	B	C	D	E	F	G	H	I
Case no.	Report date	Date of onset	Gender	Age	Hospitalised/Discharged/Deceased	HK/Non-HK resident	Case classification	Confirmed/probable

Methods and Results

For both of these two questions, I used the method of analogy. That is I used the condition in the first two waves of outbreaks to predict the development direction in the third wave.

Question 1: prediction towards the features of susceptible population

For this question, I obtained the results mainly with the use of line chart.

One of the features that may influence the infection rate is age. In the first step, I split the data into 6 subsets, respectively with the age "<15", "15-30", "30-45", "45-60", "60-80" and ">80". Considering the different population base at different ages, I used the value of r = $\frac{daily\ additional\ confirmed\ cases}{population\ ratio(\text{2016 Population By}-\text{census Office,2017})}$ to represent the relative relationship of infection rate among different subsets. (Because the demographic census of Hong Kong in 2020 is hard to find, I used the data in 2016 to simulate.) Then I drew the line chart of r for each subsets in the same plots.

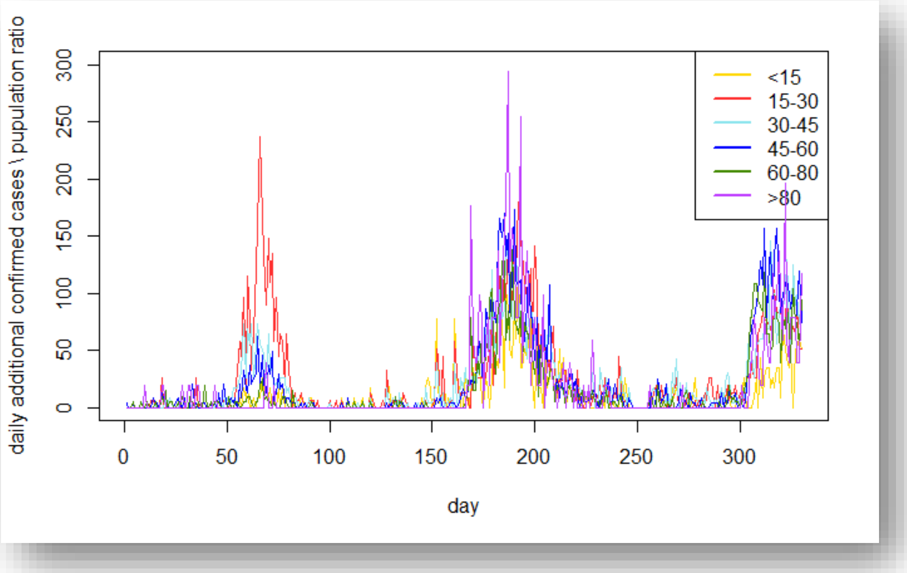


Figure 2. Daily additional confirmed cases for people in 6 different age groups on each day

However, the lines are too irregular and unsmooth to distinguish. To solve this problem, I added the use of “spline” function to make the lines more smooth and improve their visualization effects.

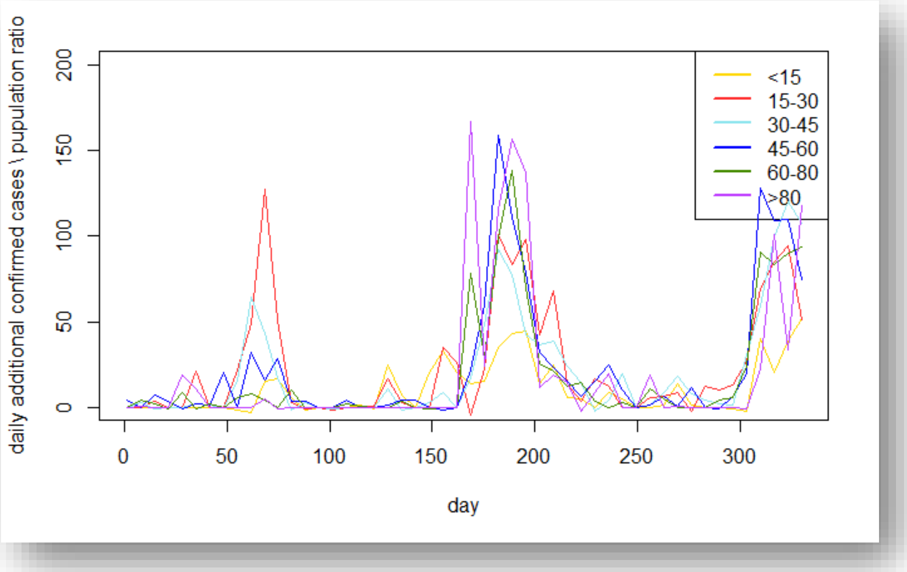


Figure 3. Daily additional confirmed cases for people in 6 different age groups on each day with smoother data

From this improved plot, we could easily find several results below.

Result 1. The people less than 15 years old always have relative lower infection rate of COVID-19 in Hong Kong.

Result 2. The people from 15 to 45 years old have obviously higher infection rate of COVID-19 in the first wave of outbreak, but have relative lower infection rate in the following outbreaks.

Result 3. The people larger than 45 years old have relative lower infection rate in the first wave of outbreak, and have obviously higher lower infection rate in the following outbreaks.

To make the comparison more clearly, I re-split the data into two subsets, respectively with the age lower than 45 and equal or greater than 45.

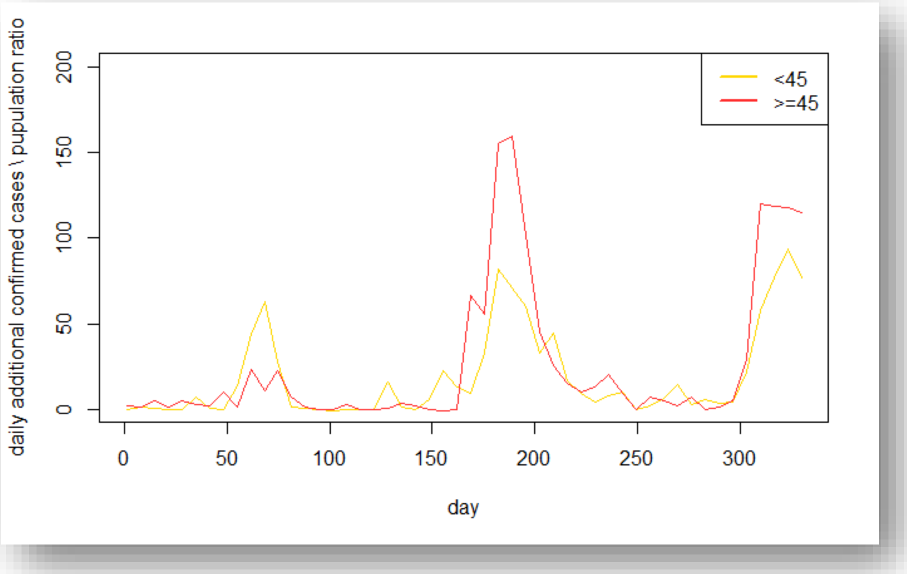


Figure 4. Daily additional confirmed cases for people in 2 different age groups on each day with smoother data

According to the relevant scientific research, elder people are more possible to be infected by the Novel Coronavirus, which is coincident with the latter part of Result 2 and Result 3. The result that younger people have higher infection rate in the first wave could be explained by the social factors. According to the young people's habits, they have higher frequency to appear in the crowded place. At the beginning of epidemic, people have lower protection awareness, thus, young people, who always attach with more people, are more likely to be infected at the earlier stage with fewer protection methods. After all people in Hong Kong improve their protection awareness, the infection rate could better react the true difference made by the age, and we could use this conclusion to predict the susceptible population in the near future.

Result 4. Elder people have more possibility to be infected in the coming days.

Another one of the features that may influence the infection rate is the gender. The line chart is shown below.

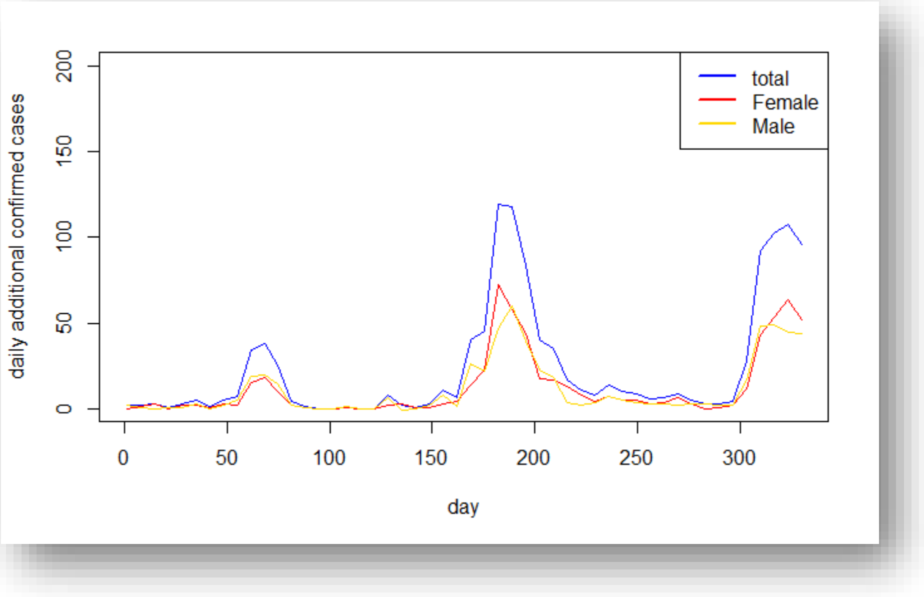


Figure 5. Daily additional confirmed cases for people in different gender on each day with smoother data

However, I could not see obvious differences between line of female and line of male.

Result 5. There is no obviously difference of infection rate between females and males.

Question 2: prediction of daily additional cases in the third wave of outbreaks

For this question, I obtained the results with the use of both plot and mathematical analysis.

First, let do some pre-definitions.

Definition 1. Define S[k] be the sum of additional confirmed cases in the k^{th} wave of outbreaks.

Definition 2. Define u[i] be the daily additional confirmed cases in day I, while u[i,j] be the sum of additional confirmed cases from day i to day j.

I divided the period of k^{th} wave of outbreaks into n classes. Let i be the start day of each class, while let j be the end day of corresponding class.

Definition 3. Define S[k] be the sample size k^{th} wave of outbreaks.

Definition 4. Define u[i,j] be the observed frequency of each class.

Definition 5. Define the probability of day i in the k^{th} wave of outbreaks equal to $\frac{u[i]}{S[k]}$.

Definition 6. Define the scales-up normal distribution be the normal distribution that multiply S[k] to the probability that the variable corresponding to.

Under such pre-definitions, I assumed that the day in each wave of outbreaks follows the normal distribution.

In the first step, I used image fitting to obtain an intuitive sense of degree of fitting. To improve the visualization effect of the plot, I multiplied S[k] for each probability. Thus, the meaning of vertical coordinate is not the probability but the daily additional confirmed cases. I added three curves of scaled-up normal distribution (in Definition 6) to do the image fitting. The plot is shown below.

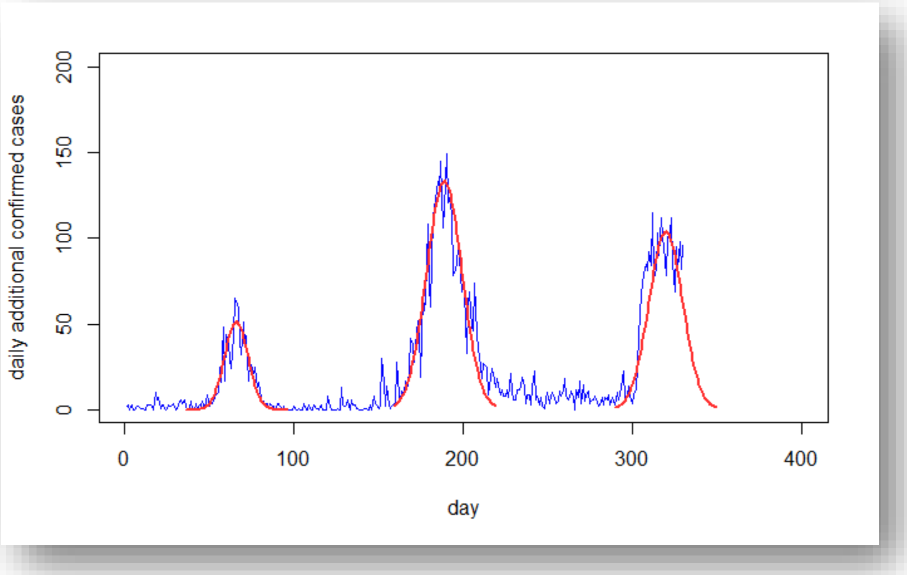


Figure 6. Daily additional confirmed cases on each day fitting with the curves with normal distribution

Just make the judgment by our naked eyes, the simulation seems to perform pretty well.

Result 6. The curve scaled-up normal distribution (in Definition 6) visually well fits the line of daily additional confirmed cases in each wave of outbreaks.

To further verify the assumption, I did the mathematical analysis. That is testing the goodness of fit to the normal distribution. (He, 2020) Goodness of fit is a well-known method to test whether a given set of data approximate a continuous distribution such as the normal in this assumption.

First, I used goodness of fit to test the performance in the first wave of outbreaks. Let H_0 be the data approximate a normal distribution. Let H_1 be the data don't approximate a normal distribution. Columns 1 to 4 are the summary data obtained from the data source. We divided the data into 8 classes. As mentioned above, End-point Yi refers to the end day of each class. Oi refers to the sum of additional confirmed cases from in i^{th}class. Let $\bar{Y}$ = 66, S = 7.5. I tested to see if these data can be considered to be normally distribution, and the procedure is summarized in the following table and the computational steps follow the table.

Class	Mid-point	End-point Yi	Oi	Est. Zi=$\frac{Y_i-\bar{Y}}{S}$	Cum. Prob.	Int. Prob.	Ei	Contribution to χ^2
1	50.5	47.875	11	−2.4167	0.0078	0.0078	7.2130	1.9883
2	56.375	53.75	33	−1.6333	0.0512	0.0434	39.9417	1.2064
3	62.25	59.625	125	−0.8500	0.1977	0.1465	134.8925	0.7255
4	68.125	65.5	245	−0.0667	0.4734	0.2758	253.9759	0.3172
5	74	71.375	284	0.7167	0.7632	0.2898	266.8934	1.0964
6	79.875	77.25	158	1.5000	0.9332	0.1700	156.5541	0.0134
7	85.75	83.125	52	2.2833	0.9888	0.0556	51.2093	0.0122
8	91.625	89	13	3.0667	0.9989	0.0101	9.3233	1.4499
sum			921					6.8093

1. Standardize the class interval end points, $\frac{Y_i-\bar{Y}}{S}$, where $\bar{Y}$ = 66, S = 7.5 are given.
2. Determine the cumulative probability for each standardized value from Standard Normal Table. In R language, Cum. Prob. = pnorm(Z_i)
3. Calculate the probability for each class interval. Int. Prob. [i] = Cum. Prob. [i] when i = 1. Int. Prob. [i] = Cum. Prob. [i] – Cum. Prob. [i-1] when i>1.
4. The expected frequency for each class is calculated by multiplying the interval probability by the sample size (sample size = sum of Oi = the sum of additional confirmed cases in the k^{th} wave of outbreaks). In this case, sample size equals to 921.

5. The contribution of each class to overall χ^2 is equal to $\frac{(Observed\, frequency - Expected\, frequency)^2}{expected\, frequency} = \frac{(Oi-Ei)^2}{Ei}$

6. The calculated χ^2 is the sum of each class contribution. χ^2 = 6.8093. The number of class is 8 in this case. After 1 degree for using the total, 1 degree for using $\bar{Y}$ and 1 degree for using S, we got the degree of freedom = 8 – 1 – 1 – 1 = 5.

7. Compared with χ^2_{cr} = 11.0705 ($\alpha = 5\%, \text{d.f.} = 5$). Because $\chi^2 < \chi^2_{cr}$, the calculated χ^2 is not significant at 5% level, we could not reject the null hypothesis and the data approximate a normal distribution.

Result 7. The days in first wave of outbreak approximate a normal distribution with 5% significant level.

Then, I used goodness of fit to test the performance in the second wave of outbreaks. The hypothesis and procedure is highly similar to the first one. Let $\bar{Y}$ = 189, S = 12. Using the same step, we got the following table.

Class	Mid-point	End-point Yi	Oi	Est. Zi=$\frac{Y_i-\bar{Y}}{S}$	Cum. Prob.	Int. Prob.	Ei	Contribution to χ^2
1	159	162	49	-2.25	0.0122	0.0122	44.2159	0.5176
2	165	168	81	-1.75	0.0401	0.0278	100.6781	3.8462
3	171	174	246	-1.25	0.1056	0.0656	237.2413	0.3234
4	177	180	389	-0.75	0.2266	0.1210	437.5759	5.3925
5	183	186	675	-0.25	0.4013	0.1747	631.7681	2.9584
6	189	192	763	0.25	0.5987	0.1974	714.0416	3.3568
7	195	198	542	0.75	0.7734	0.1747	631.7681	12.7552
8	201	204	374	1.25	0.8944	0.1210	437.5759	9.2370
9	207	210	274	1.75	0.9599	0.0656	237.2413	5.6955
10	213	216	124	2.25	0.9878	0.0278	100.6781	5.4025
11	219	222	100	2.75	0.9970	0.0092	33.4381	132.4980
sum			3617					181.9831

Unlike the well-performance in the first case, this time I obtained a pretty large χ^2, which is so large that it is easily to reject the null hypothesis.

Result 8. The days in second wave of outbreak don’t approximate a normal distribution with reasonable significant level.

To find the reason that causes of huge different fitting performance in these two cases, I observed the data both in the table and data set. Then I found that, before the peak (in first 6 classes), Contribution to χ^2 is pretty small, but after the peak, there are some pretty large values such as 12.7552 in class 7 and 132.4980 in class 11. In class 7, the observed frequency is obviously smaller than expected one, while in class 11, the observed frequency is obviously larger than expected one. That means, in the latter period of the second wave of outbreak, the daily additional confirmed cases have higher rate to decrease than that of the rate to increase in the first half part.

Combined with mathematical knowledge and social factors, I made the following conjecture of the reason to this results. For the curve with normal distribution centers on the mean and it is asymmetric between the left and the right. In the first wave of outbreak, the rate of the increment of daily additional confirmed cases in the first half period, is approximately equal to the rate of decrement in the latter half period. Thus, the true condition could be well fitted by the normal distribution. However, in the second wave of outbreak, the people and the government has more experience to face with it, thus the variation shows a more rapidly rate to decrease and the true condition in the latter part could not be well fitted by the normal distribution. Another reason that leads to this results is that the daily additional confirmed cases after the second mass outbreak are not keep stable in a very low level as the condition after the first mass outbreak, which is still not coincident with the curve of normal distribution.

According to the analysis of the first wave of outbreaks, we could make a prediction to the third wave. From the Figure 6, we could see that Hong Kong are know just after the peak of the third outbreak. I conjecture that the variation line of daily additional confirmed cases in the third wave could still approximate to the curve of normal distribution in visualization. However, because of the more experience to response to the outbreak, I think it will still go on the development direction as the second one in the rest of the time. If the development approximate the curve of normal distribution, the number of additional confirmed cases may become stable after 20 days from December 17 (With the start at 290th day, peak at 320th day and now at 330th day). However, adding with the influence factors mentioned above, the true date would be little bit earlier.

Conclusions and Discussions

Conclusions

Overall, the prediction towards the coming development of the third wave of outbreak in Hong Kong could be concluded in two aspects. The first aspect is elder people would have relative higher infection rate. And the second aspect is the number of additional confirmed cases has high probability to begin to become stable in low level from early January in 2021.

Future Works

With regards to the second question, the simulation has many shortages and plenty of future works to be done.

First improvement could be the additional preprocessing towards the local extreme data points. Because goodness-of-fit test is sensitive to outliers, fewer local extreme values could make huge influence to χ^2. If we could do some preprocessing to the data set to make the line smoother, the degree of fitting could be better.

Another improvement is that we could try to use Susceptible-exposed-infectious-recovered model with vital dynamics (SEIR model) to make the prediction. SEIR model is one of the most common models to "describe the spread of the virus and compute the number of infected and dead individuals" (José M. et al., 2020). In the future, if I obtain more detailed data relating to COVID-19, such as the number of recovery in each day, how the factors (etc. temperature, humidity) influence the transmission rate of COVID-19, the basic reproduction number and so on, I could use this model to make a prediction and make a comparison with the conclusions in this paper.

References

Fontanet, A., Cauchemez, S. 2020. “COVID-19 herd immunity: where are we?” *Nat Rev Immunol* 20: 583–584. https://doi.org/10.1038/s41577-020-00451-5

Jewel Gausman, ScD, MHS and Ana Langer. 2020. “Sex and Gender Disparities in the COVID-19 Pandemic.” *MD JOURNAL OF WOMEN’S HEALTH Volume 29, Number 4, Mary Ann Liberty,* Inc. DOI: 10.1089/jwh.2020.8472

Data.GOV.HK. December 17, 2020. “Details of probable/confirmed cases of COVID-19 infection in Hong Kong.” *Hong Kong Government.* https://data.gov.hk/en-data/dataset/hk-dh-chpsebcddr-novel-infectious-agent/resource/24f08145-6c16-4de9-83a0-ee52b67eeedb

2016 Population By-census Office. April 10, 2017. “Usual Residents by Age, Year and Sex”. *Hong Kong Government.* http://data.chinabaogao.com/gonggongfuwu/2019/0Q9441E62019.html

Daihai, He. 2020. “AMA488 Simulation and Risk Analysis.” *Lecture notes*: 209-214

Carcione José M., Santos Juan E., Bagaini Claudio, Ba Jing. 2020. “A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model.” *Frontiers in Public Health* 8: 230.