PUSAN NATIONAL UNIVERSITY

전 세계 COVID-19 현황 파악과 Machine Learning과 Time Series Analysis를 통한 미래 동향 모델 구현

Do-Hee Kim
Department of Statistics and Information Computer Engineering,Pusan National University, Busandaehak-ro 63beon-gil, Jangjeon-dong, Geumjeong-gu, Busan
* More details in https://github.com/kheedogg/COVID-19_Analysis

Rationale / Objectives

2020년초 전세계적 신종바이러스 전염병 "COVID-19" 유행 시작 → [1] 코로나의 퍼지는 양상 및 특징 파악 필요 → [2] COVID-19 감염자 수 예측 → 적절한 대비체계 마련 도움

Methods / Framework

- [1] COVID-19 바이러스 전파 현황 및 특징 파악
  - Datewise Analysis
  - Countrywise Analysis
  - Clustering of Countries
- [2] COVID-19 바이러스 Confirmed Cases 예측
  - Machine Learning
    - Linear Regression
    - Polynomial Regression
  - Time Series Analysis
    - Holt's Linear
    - Holt's Winter Model
    - Auto Regressive Model
    - Moving Average Model
    - SARIMA
    - ARIMA

* [2] COVID-19 바이러스 Confirmed Cases 예측은
  - 발병 이력이 존재하는 전세계 국가에 대한 Confirmed Cases
  - [1] Clustered Countries에 대한 Confirmed Cases

크게 2번 시행되며, 각 케이스별로 적합률이 가장 높은 모델을 확인 후 예측된 값을 비교하였다.

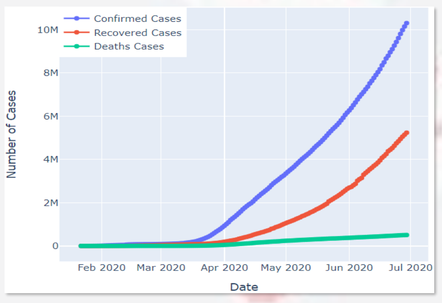
Data Introduction

	Province/State	Country/Region	Last Update	Confirmed	Deaths	Recovered	SNo	ObservationDate
0	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0	1	01/22/2020
1	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0	2	01/22/2020
2	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0	3	01/22/2020
3	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0	4	01/22/2020
4	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0	5	01/22/2020

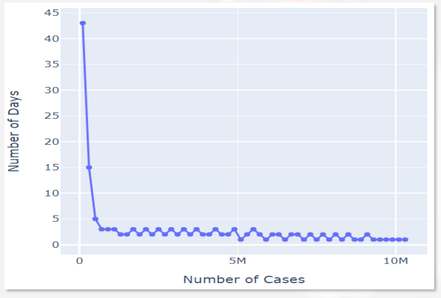
* SNo – Serial number
* ObservationDate - MM/DD/YYYY

EDA (Exploratory Data Analysis)

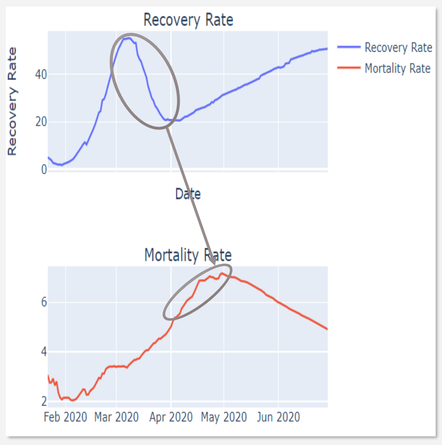
〈 Datewise Analysis 〉

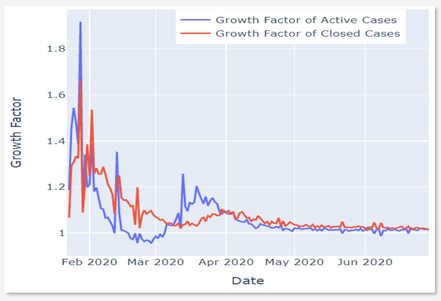


➢ Cumulative Growth of different types of cases



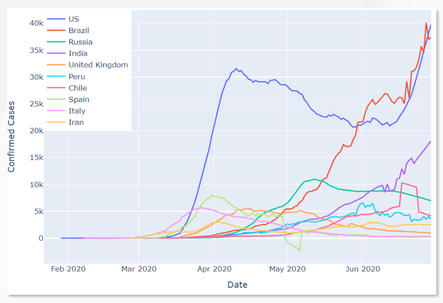
➢ Number of Days required for increase in number of cases



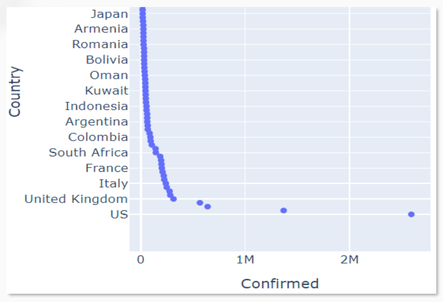


➢ Datewise Growth Factor of Active and Closed Cases
- Active case = Confirmed case – Recovered case – Death case
- Closed case = Recovered + Death case

〈 Countrywise Analysis 〉

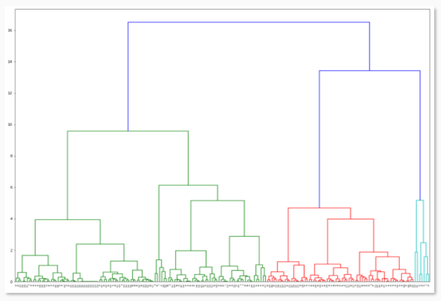


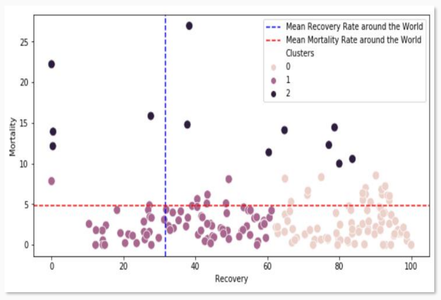
➢ 7 Days Rolling Average of Daily increase of Confirmed cases



➢ The number of Confirmed per Countries

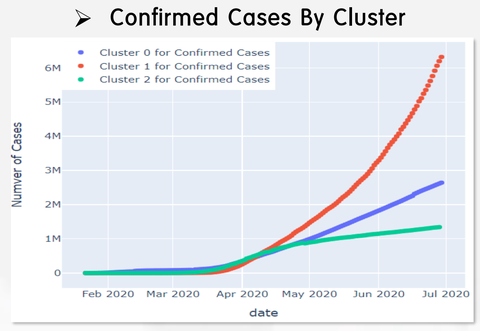
〈 Clustering of Countries 〉





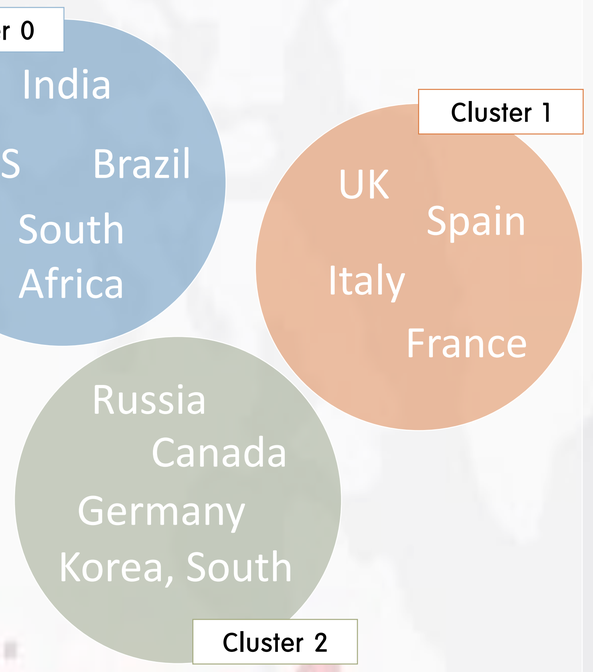
All methods namely Elbow Method and Hierarchical Clustering shows K=3 will correct number of clusters.

➢ Confirmed Cases By Cluster



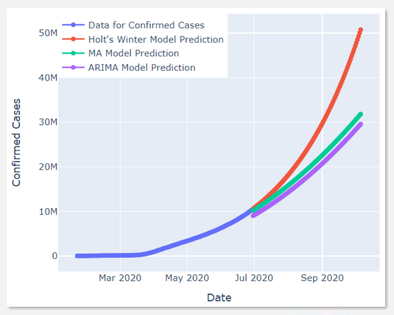
It seems to be each group shows explicit pattern.
So, each should try to implement a predictive model.
The group characteristics are as follows;

Cluster 0
"Low Mortality Rate & really Low Recovery Rate"
Cluster 1
"High Mortality Rate & considerably Good Recovery Rate"
Cluster 2
"Low Mortality Rate & really High Recovery Rate"

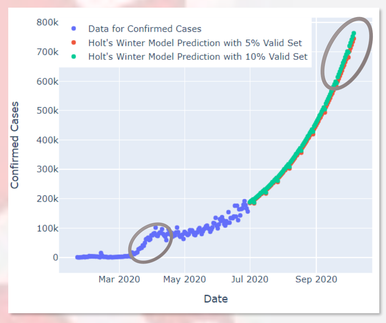


Predictions

〈 The nations of the world 〉



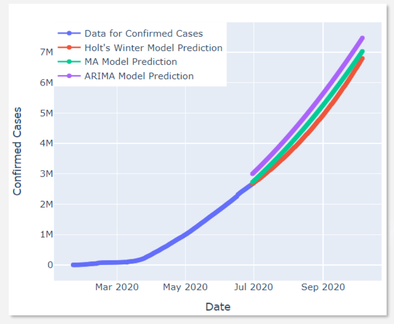
	Model Name	Root Mean Squared Error
3	Holt's Winter Model	52727.719735
5	Moving Average Model (MA)	68691.384006
6	ARIMA Model	70212.495132
7	SARIMA Model	70212.495132
2	Holt's Linear	182429.611669
4	Auto Regressive Model (AR)	199901.214255
1	Polynomial Regression	1983232.816777
0	Linear Regression	2598361.511152



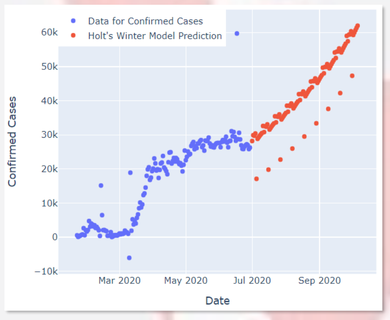
➢ Compare Prediction by the best model with 5% Validation Set

➢ Amount of Change by a day

〈 Cluster 0 countries 〉



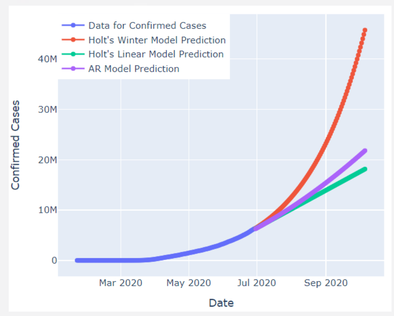
	Model Name	Root Mean Squared Error
3	Holt's Winter Model	7306.711512
5	Moving Average Model (MA)	17017.861655
6	ARIMA Model	18416.033889
7	SARIMA Model	18416.033889
2	Holt's Linear	22695.646030
4	Auto Regressive Model (AR)	41156.823807
1	Polynomial Regression	100799.369166
0	Linear Regression	498568.202112



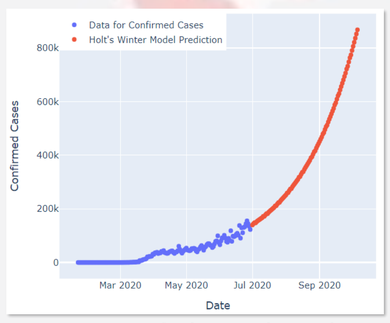
➢ Compare Prediction by the best model in Cluster 0

➢ Amount of Change by a day

〈 Cluster 1 countries 〉



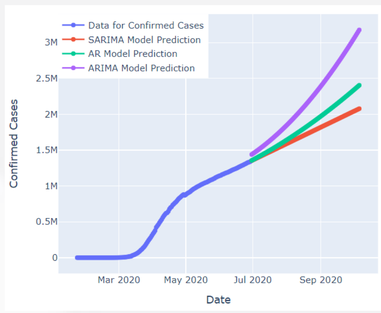
	Model Name	Root Mean Squared Error
3	Holt's Winter Model	44001.721669
2	Holt's Linear	76177.823534
4	Auto Regressive (AR)	100230.132317
5	Moving Average (MA)	176939.392096
6	ARIMA Model	255456.527738
7	SARIMA Model	255456.527738
1	Polynomial Regression	355280.018103
0	Linear Regression	2158644.361009



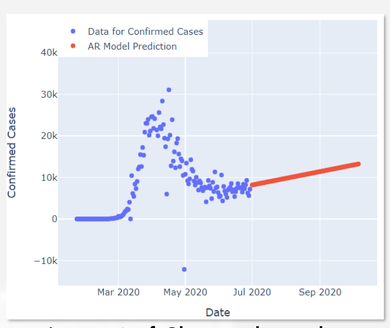
➢ Compare Prediction by the best model in Cluster 1

➢ Amount of Change by a day

〈 Cluster 2 countries 〉



	Model Name	Root Mean Squared Error
4	Auto Regressive Model (AR)	2298.004181
6	ARIMA Model	2987.212972
7	SARIMA Model	2987.212972
2	Holt's Linear	4870.127846
5	Moving Average Model (MA)	7313.003092
3	Holt's Winter Model	37779.753229
1	Polynomial Regression	60825.623238
0	Linear Regression	62290.676479



➢ Compare Prediction by the best model in Cluster1

➢ Amount of Change by a day

Conclusion

COVID-19 does not have very high mortality rate as we can see. Also the healthy Recovery Rate implies the disease is curable. The only matter of concern is the exponential growth rate of infection. Plus, Countries like USA, Spain, UK, and Italy are facing some serious trouble in containing the disease showing how deadly the negligence can lead to. Through the best prediction model, we can know that there is a possibility the second coronavirus pandemic will come in autumn, especially cluster 0 and 1. Therefore, it seems necessary for countries including cluster 0 and 1 to prepare appropriate countermeasures accordingly.

References

Johns Hopkins Github repository, https://github.com/CSSEGISandData/COVID-19

"COVID-19 Visualizations, Predictions, Forecasting" , https://www.kaggle.com/neelkudu28/covid-19-visualizations-predictions-forecasting (2020년 8월 19일)

이재길, 『R 프로그램에 기반한 시계열 자료 분석』, 황소걸음 아카데미, 2017