# Global Deaths prediction due to Covid-19

Akhilesh Chauhan(170070), Chirag Jindal(170225), Shivam Arya(170666), Amit Badoni(170091) and Dhawal Raturi(170244)

November 30,2020

# Objective

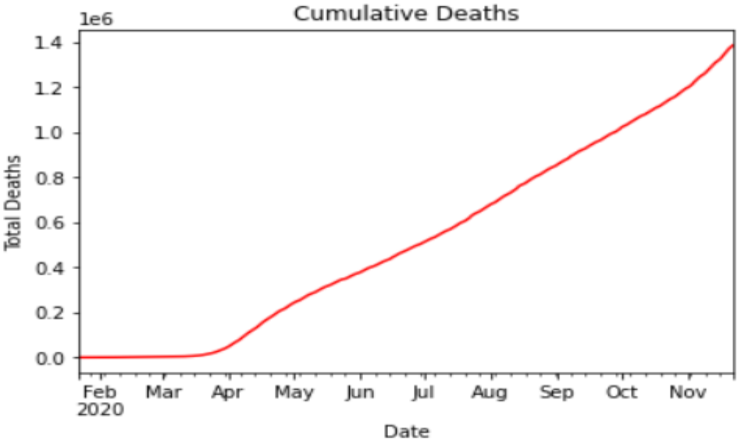
- Model the time series using standard time series models.
- Predict the future observations after fitting the model.
- Tune the parameters of the fitted model well to get good predictions

# Motivation

- Machine learning and statistical methods, which time series forecasting is a subset of, have been successfully implemented in the past in the area of infectious diseases. For example- modeling leptospirosis and its relationship to rainfall and temperature.
- Similar approaches have also been followed to model diseases that occur in cyclic or repeating patterns, such as the seasonal influenza, for which a number of studies that use time-series modeling to predict future outbreaks have been published.
- Regarding COVID-19 forecasting, there has been a surge in scientific work published during the last few months. The majority of these studies focus on predicting coronavirus-related metrics such as active cases and deaths across the world.

# Data Set

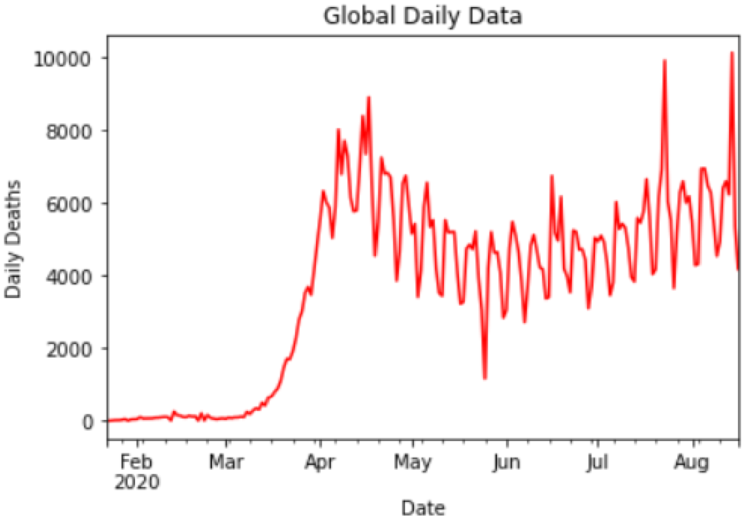
- We are using the data set available at John Hopkins University's Github Page.
- It contains the country wise data for cumulative deaths due to Covid-19.



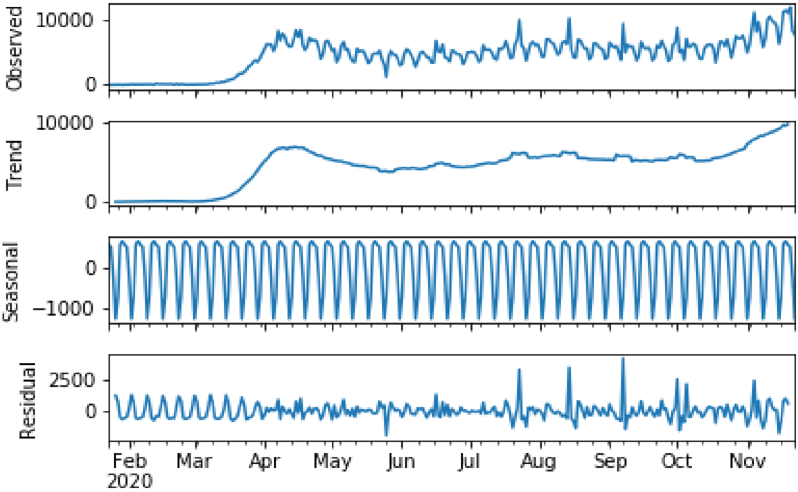
# Methodology

- Organised the data and took the sum of data from all the countries.
- Took the first difference to get the daily data.
- Made the series stationary by taking an appropriate order of difference.
- Investigated the Auto-Correlations(ACF) and the Partial Auto-Correlations(PACF).
- Tried to fit different time series models and evaluated the model based on the Akaike Information Criterion(AIC).
- Compared the prediction performance of different model fits.

# Daily Data

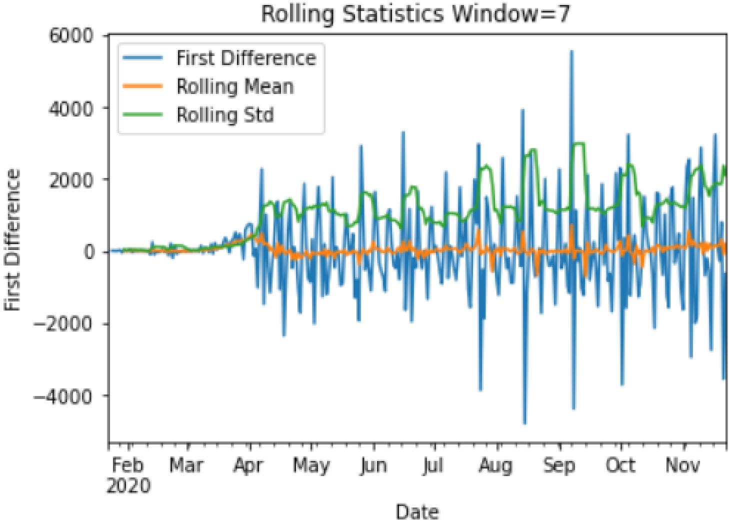

Global Daily Data
Daily Deaths
10000
8000
6000
4000
2000
0
Feb
2020
Mar
Apr
May
Jun
Jul
Aug
Date

# Seasonal Decomposition



# Stationarity of first difference

- We tested the stationarity using the Augmented Dicky Fuller Test.

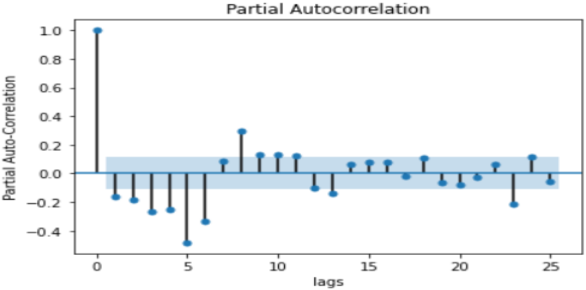


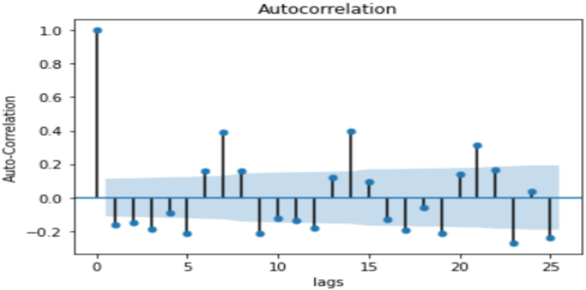
# Dickey-Fuller Test for Trend on First Difference

$H_0$ : Series has a trend.
$H_A$ : Series is stationary.
Results of Dickey-Fuller Test :
Test Statistic : -2.914002
p-value : 0.043744
Lags Used : 12.000000
Number of Observations Used : 193.000000
Critical Value (1%) : -3.464694
Critical Value (5%) : -2.876635
Critical Value (10%) : -2.574816

Here, Test Statistic is less than Critical Value at 5% and thus we reject the Null Hypothesis and thus at a confidence interval of 95%, Trend is not present in first difference.

# Auto Correlation Plots



# Parameter Estimation

- The auto-correlation plot suggests a seasonality of order 7, which is also evident from the plot.
- To begin with, we split the data into 2 parts(Test and Train). We tried fitting different models on the Train data and evaluated the performance on the test data.
- Then we started with the ARIMA Model.

# ARIMA Model

- ARIMA offers a high level of interpretability, as, based on the assumptions of the model, the relationship between the independent variables and the dependent variables are well-understood and therefore easily explained.
- This enables researchers to gain a deep understanding not only of the relationship between the current state as a function of the past states (endogenous variables), but also of any influence inputs outside the state of the series might have (exogenous variables).
- ARIMA(Auto Regressive Integrated Moving Average) Model($X_t$ is the time series,$\epsilon_t$s are i.i.d. residuals )

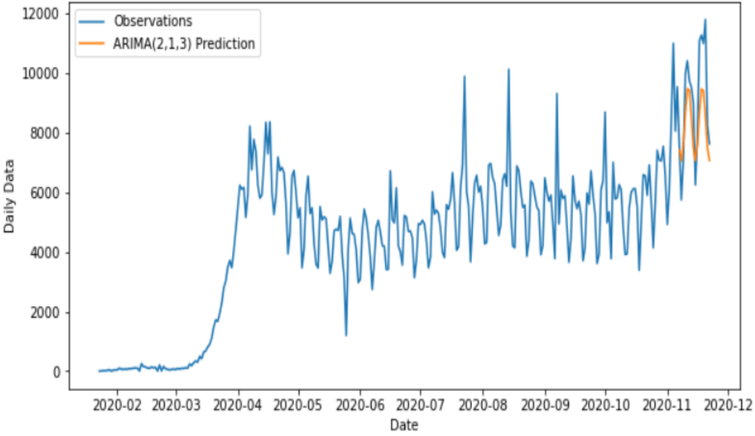
  $$\phi(B)(1-B)^d X_t = \theta(B)\epsilon_t$$

  where B denotes the lag operator and $\phi(B)$ and $\theta(B)$ are the AR and MA polynomials of order p and q respectively.
- $$AIC = -2 log \hat{L} + 2(p+q+d+1)$$

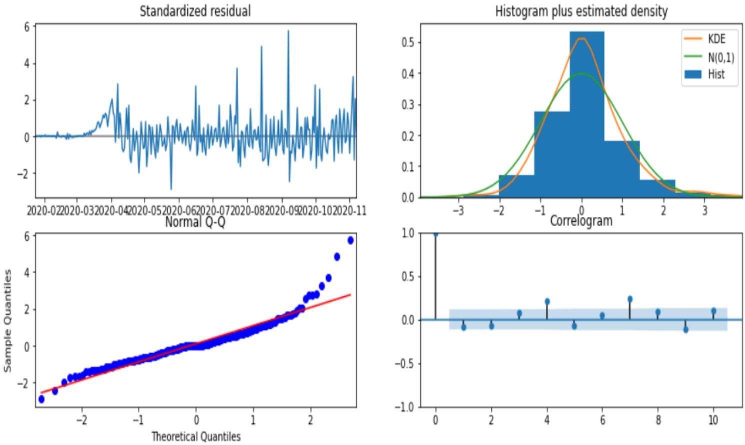
  where $\hat{L}$ denotes the likelihood and the second term contains the number of parameters being used.

# ARIMA Model

- Varying p and q, and using the AIC as the evaluation parameter, we get p =2,q=3,d=1.



# Residual Analysis



# Results of ARIMA Model

```
                               SARIMAX Results
==============================================================================
Dep. Variable:             first diff   No. Observations:                  290
Model:               SARIMAX(2, 1, 3)   Log Likelihood               -2364.199
Date:                Thu, 26 Nov 2020   AIC                           4740.398
Time:                        13:37:55   BIC                           4762.397
Sample:                    01-23-2020   HQIC                          4749.213
                         - 11-07-2020
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.2457      0.005    226.592      0.000       1.235       1.256
ar.L2         -0.9977      0.005   -191.815      0.000      -1.008      -0.987
ma.L1         -1.8813      0.038    -48.952      0.000      -1.957      -1.806
ma.L2          1.7177      0.057     30.170      0.000       1.606       1.829
ma.L3         -0.5955      0.041    -14.657      0.000      -0.675      -0.516
sigma2      7.506e+05   3.22e+04     23.314      0.000    6.88e+05    8.14e+05
===================================================================================
Ljung-Box (Q):                      187.00   Jarque-Bera (JB):               486.70
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               3.18   Skew:                             1.37
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.74
===================================================================================
```

# SARIMA Model

- SARIMA Model for a time series $X_t$ is given by,

$$\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d X_t = \Theta(B^s)\theta(B)\epsilon_t$$

  where $B$ denotes the lag operator , $\phi(B)$ and $\theta(B)$ are the AR and MA polynomials of order p and q, and $\Phi(B)$ and $\Theta(B)$ are the seasonal AR and MA polynomials of order P and Q respectively, and d and D are the orders of difference and seasonal difference.

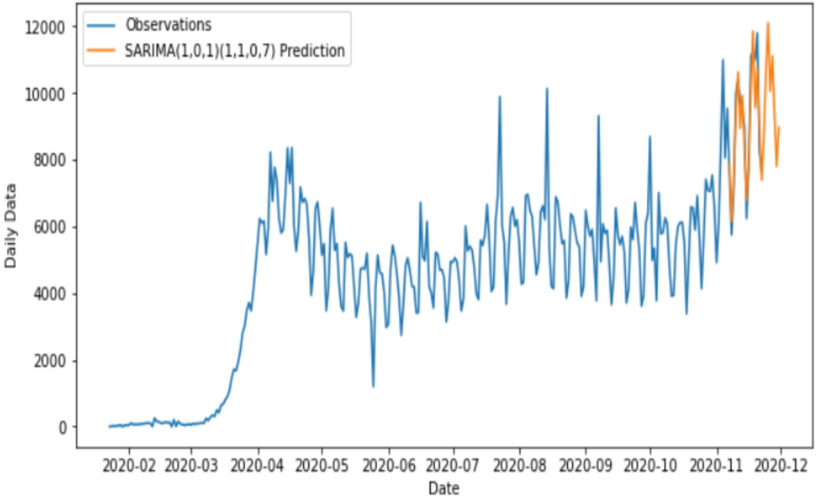
- 

$$AIC = -2log\hat{L} + 2(p+q+P+Q+D+d+2)$$

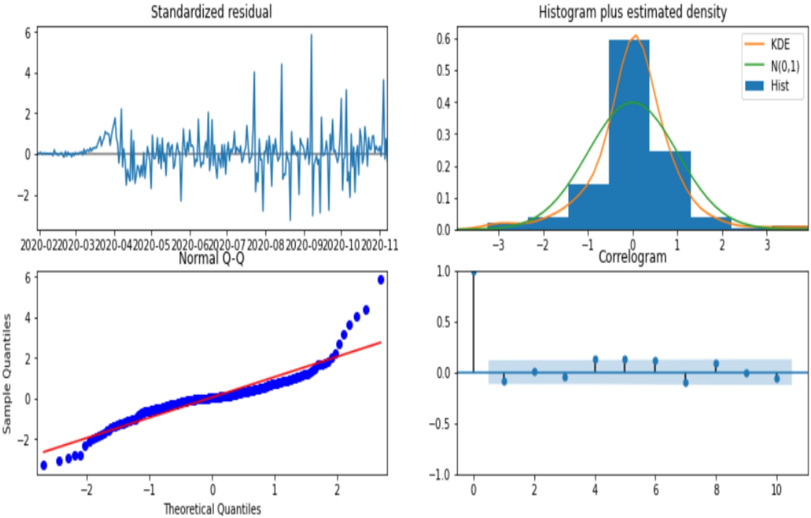
  where $\hat{L}$ denotes the likelihood and the second term contains the number of parameters being used.

- Varying p and q, and using the AIC as the evaluation parameter, we get p =1,d=0,q=1,P=1,D=1,Q=0,S=7.

# SARIMA Model



# Residual Analysis

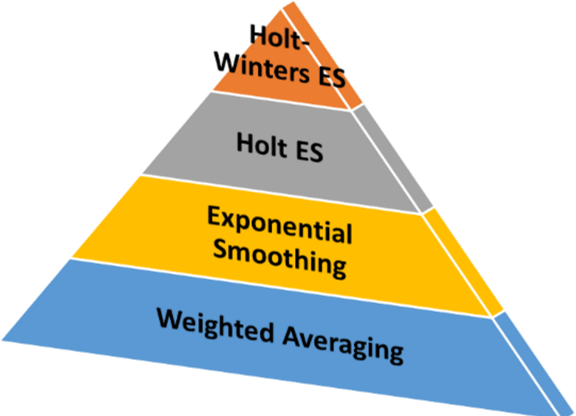


# Results of SARIMA Model

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                         first diff   No. Observations:                  290
Model:             SARIMAX(1, 0, 1)x(1, 1, [], 7)   Log Likelihood               -2323.928
Date:                            Thu, 26 Nov 2020   AIC                           4655.857
Time:                                    13:12:36   BIC                           4670.439
Sample:                                01-23-2020   HQIC                          4661.704
                                     - 11-07-2020
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9612      0.022     44.385      0.000       0.919       1.004
ma.L1         -0.6975      0.044    -15.830      0.000      -0.784      -0.611
ar.S.L7       -0.4502      0.029    -15.470      0.000      -0.507      -0.393
sigma2      7.927e+05   3.19e+04     24.838      0.000     7.3e+05    8.55e+05
===================================================================================
Ljung-Box (Q):                      110.79   Jarque-Bera (JB):               578.91
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               3.88   Skew:                             0.91
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.77
===================================================================================
```

# Holt-Winters Exponential Smoothing

- Holt-Winters Exponential Smoothing is used for forecasting time series data that exhibits both a trend and a seasonal variation. The Holt-Winters technique is made up of the following four forecasting techniques stacked one over the other:



# Holt-Winters Exponential Smoothing

Forecast at step *(i+k)*

Estimated Seasonal variation of period length=m, at step (i+k)

$$F_{(i+k)} = \underbrace{(L_i + k * B_i)}_{\text{Estimated Level at step (i+k)}} * S_{(i+k-m)}$$

Estimated Trend at step *i*

Estimated Trend at step *(i-1)*

$$B_i = \beta * \underbrace{[L_i - L_{(i-1)}]}_{\text{Estimated rate of change of level from step (i-1) to step i}} + (1-\beta) * B_{(i-1)}$$

Estimated Level at step *i*

Time series value at step *i*

Seasonal component at seasonal step *(i-m)*

$$L_i = \alpha * \frac{T_i}{S_{(i-m)}} + (1-\alpha) * \underbrace{[L_{(i-1)} + B_{(i-1)}]}_{\text{Estimated Level at i as a function of: level at (i-1), and change of level from step (i-1) to step i}}$$

Estimated Seasonal component at step *i*

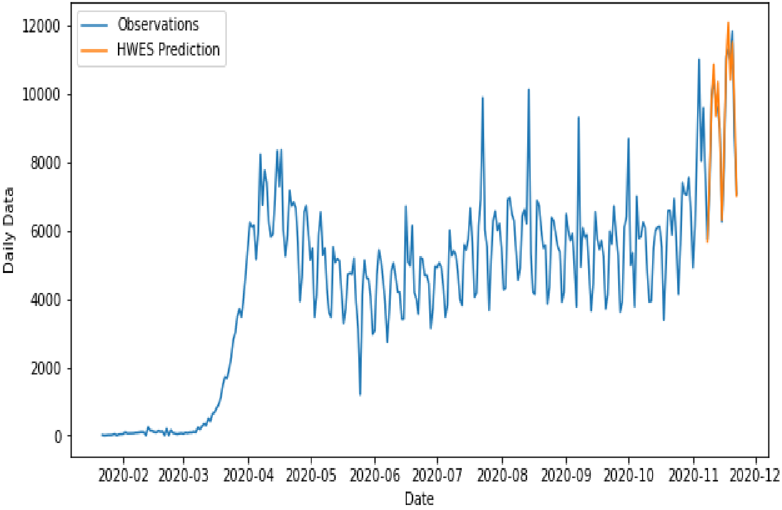
Time series value at step *i*

$$S_i = \gamma * \frac{T_i}{L_i} + (1-\gamma) * S_{i-m}$$

Level at step *i*

Seasonal component at the previous seasonal step *(i-m)*

# Holt-Winters Exponential Smoothing



# Results of HWES

```
                        ExponentialSmoothing Model Results
================================================================================
Dep. Variable:                          0   No. Observations:                  291
Model:               ExponentialSmoothing   SSE                      314026579.701
Optimized:                           True   AIC                           4064.474
Trend:                     Multiplicative   BIC                           4104.881
Seasonal:                  Multiplicative   AICC                          4065.789
Seasonal Periods:                       7   Date:                 Sun, 29 Nov 2020
Box-Cox:                            False   Time:                         15:48:04
Box-Cox Coeff.:                      None
=================================================================================
                          coeff                 code              optimized
---------------------------------------------------------------------------------
smoothing_level               0.2878571                alpha                 True
smoothing_trend               0.0575714                 beta                 True
smoothing_seasonal            0.3560714                gamma                 True
initial_level                 4831.4762                  l.0                 True
initial_trend                 1.1558217                  b.0                 True
initial_seasons.0             0.0035186                  s.0                 True
initial_seasons.1             0.0002070                  s.1                 True
initial_seasons.2             0.0016558                  s.2                 True
initial_seasons.3             0.0033116                  s.3                 True
initial_seasons.4             0.0028977                  s.4                 True
initial_seasons.5             0.0053814                  s.5                 True
initial_seasons.6             0.0101418                  s.6                 True
---------------------------------------------------------------------------------
```

# Comparison of Models Based on Root Mean Squared Error

| | 15 Days RMSE | 30 Days RMSE |
|---|---|---|
| **ARIMA Model** | 1414.776 | 3056.512 |
| **SARIMA Model** | 674.602 | 2883.674 |
| **HWES Model** | 465.38 | 2446.559 |

- We found out that SARIMA Model gives us a better fit for the data as compared to the ARIMA Model as our data has strong seasonality(evident from ACF plots).
- We found that HWES Model given even better fit than SARIMA Model as our data has noise which gets reduced by the exponential smoothing
- HWES also takes into account the seasonality in the data and gives more importance to recent observations.

# References

- (Data source) Github page of John Hopkins University
- (Modeling seasonal leptospirosis transmission) [1]
- (Holt-Winters Exponential Smoothing) [2]

**Thank You!**