Introduction:

On 31st December 2019, WHO was alerted to several cases of pneumonia in Wuhan City, Hubei Province of China. The virus did not match any other known virus. This raised concern because when a virus is new, its effect on people was unknown.

Coronavirus disease 2019 (COVID-19) is a contagious respiratory and vascular (blood vessel) disease. It is caused by becoming infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is a specific type of coronavirus. Common symptoms include fever, cough, fatigue, shortness of breath or breathing difficulties, and loss of smell and taste. The incubation period, which is the time between becoming infected with the virus and showing symptoms, may range from one to fourteen days. While most people have mild symptoms, some people develop acute respiratory distress syndrome (ARDS), multi-organ failure, septic shock, and blood clots. Longer-term damage to organs (in particular, the lungs and heart) has been observed, and there are concerns about a significant number of patients who have recovered from the acute phase of the disease but continue to experience a range of side/after-effects—including severe fatigue, memory loss and other cognitive issues, low grade fever, muscle weakness, breathlessness, and other symptoms—for months afterwards.

Recommended measures to prevent infection include frequent social distancing, quarantine, covering coughs and sneezes, hand washing, and keeping unwashed hands away from the face. The use of face masks or cloth face coverings has been recommended by health officials in public settings to minimize the risk of transmissions, with some authorities requiring their use in certain settings, such as on public transport and in shops. There are no proven vaccines or specific treatments for COVID-19 yet, though several are in development. Management involves the treatment of symptoms, supportive care, isolation, and experimental measures. The World Health Organization (WHO) declared the COVID-19 outbreak a pandemic on 11 March 2020.

COVID-19 has been recognized as a global threat, and several studies are being conducted using various mathematical models to predict the probable evolution of this epidemic. These mathematical models are based on various factors and analysis are subject to potential bias. Here, we propose a simple econometric model that could be useful to predict the spread of COVID-19. We performed Auto Regressive Integrated Moving Average (ARIMA) model prediction on the Johns Hopkins epidemiological data to predict the epidemiological trend of the prevalence and incidence of COVID-19. For further comparison or for future perspective, case definition and data collection have to be maintained in real time.

These data are useful because they provide a forecast for COVID-19 epidemic, thus representing a valid and objective tool for monitoring infection control. All institutions involved in public health and infection control can benefit from these data because by using this model, they can construct a reliable forecast for COVID-2019 epidemic on an everyday basis. The benefit of these data lies in their easy collection and in the possibility to offer a valid forecast for COVID-19 daily monitoring after applying the ARIMA model. These data represent an easy way to evaluate the transmission dynamics of COVID-2019 to verify whether the strategy plan for infection control or quarantine is efficient.

Forecasting future COVID-19 case counts would allow governments, businesses, and individuals to prepare in advance and plan their responses accordingly. For instance, local governments can use the results to anticipate the need for emergency supplies and health-related public services. Hospitals can use the results to foresee medical supply and personnel shortages. Universities and companies can also use the results to plan for adequate protection for students and employees.

Previous work

The first paper1 analyses the potential relationship between climate and the spread of COVID-19. Studies have suggested that the spread of COVID-19 is expected to be more in the cold and temperate climate as compared to the warm and tropical climate, consistent with the behavior of a seasonal respiratory flu virus. Multiple viruses from the Coronaviridae family, including the SARS CoV-1 and MERS CoV, also demonstrate seasonality and preference for low temperature and humidity. It then discusses the attempts to clarify the role of weather parameters on the possible viral spread in the US. Through statistical analysis, vulnerable ranges of weather parameters were identified and validated for different time intervals, and against worldwide findings. Using these results, predictions have been made on the risky states in India, which could possibly be more vulnerable for the transmission of COVID-19 in the recent and upcoming months in 2020. Absolute humidity was the main parameter used, which is determined by the formula :

(1)

The vulnerable AH range determined for the US states was used to predict which Indian states would be more vulnerable for the transmission of the disease.

In the US, once the number of new cases in a 10-day interval went over 10,000, the majority of the cases were found to be reported in states experiencing absolute humidity (AH) in a narrow range of 4 to 6 g/m3, and temperature (T) in a wider range of 4 °C – 11 °C. Considering AH as a better parameter than T to study the relationship between weather and COVID-19 spread, it was found that the vulnerable 4 < AH < 6 g/m3 range for 10-day intervals were generalizable to monthly study intervals. Also, these results lied within the literature reported AH ranges worldwide. Assuming the US model to represent the world findings, risky Indian states were classified with expected 4 < AH < 6 g/m3 through all months of 2020. This study has a few limitations. Out of a large number of possible weather parameters experienced around the globe, only a few were considered specific to the US. Caution should be utilized when extrapolating these results beyond the US based weather parameters considered in this study. It should be mentioned that the role of weather in affecting the transmission rates can only be termed as correlation and not causation at this point in time. Also this study was conducted when the number of active COVID-19 cases was exceptionally low.

The second paper2 talks about the Advanced Autoregressive Integrated Moving Average (ARIMA) model. The comparison of recent cumulative and predicted cases was done for the top 15 countries with confirmed cases, deaths, and recoveries from COVID-19. The spatial map is useful to identify the intensity of COVID-19 infections in the top 15 countries and the continents. The recent reported data for confirmed cases, deaths, and recoveries for the last 3 months was represented and compared between the top 15 infected countries. The ARIMA model provides a weight to past values and error values to correct the model prediction, so it is better than other basic regression and exponential methods. The data from the top 15 countries were preprocessed with their spatial locations to collect and create spatial attributes for the overall available data sets to forecast the trajectory of COVID-19 cases.

The third paper3 focuses on the outbreak and spread of COVID-19 in China and majorly in Wuhan and non-Hubei areas. The goal of the paper was to predict when the situation will be stabilized and the inflection points in the above-mentioned areas. This paper adopts 3 kinds of mathematical models, namely, Logistic model, Bertalanffy model and Gompertz model. The regression coefficient(R2) is then calculated to find the fitting ability of each model.

Logistic Model :

(2)

Bertalanffy model :

(3)

Gompertz model:

(4)

Qt is the cumulative confirmed cases (deaths); a is the predicted maximum of confirmed cases (deaths). b and c are fitting coefficients. t is the number of days since the first case. t0 is the time when the first case occurred. Among them, the Logistic model was better than the other two models in fitting the data and both the Logistic and Gompertz models outperformed the Bertalanffy model in that sense. According to the predictions based on the three models, the total number of people expected to be infected is between 49852 and 57447 in Wuhan,12972-13405 people in non-Hubei areas and 80261-85140 in China, respectively. It was also predicted that the pandemic will end in late-April 2020 in Wuhan and before late-March in other areas, respectively.

The fourth paper4 used the modeling of COVID-19 for a cumulative number of infected cases using data available in its early phase and used it to predict the cases for the next two months. A Time series model was adopted for short term predictions using data points collected from the past in time domain. The prophet algorithm was used to train the model. The model predicted that COVID-19 infected cases would be 1.6 million and 2.3 million by the end of May and June, respectively.

Additionally, another paper5 which aimed to identify common features of cases so as to better understand what factors promote super spreading events, wherein an extraordinarily large number of secondary transmissions are produced by a single primary case, was also studied. A total of 110 cases among eleven clusters were examined. The clusters included four in Tokyo and one each in Aichi, Fukuoka, Hokkaido, Ishikawa, Kanagawa, and Wakayama prefectures. All clusters were associated with close contact in indoor environments, including fitness gyms, a restaurant boat on a river, hospitals, and a snow festival where there were eating spaces in tents with minimal ventilation rate. It was concluded that closed environments largely contribute to secondary transmission of COVID-19 and promote super spreading events.

The above papers implement only one model whereas this project attempts to implement various time-series models and compares them to find the best model.

Problem statement and approach to problem:

Approach to the problem:

Since this data is collected over a period of time, time-series forecasting has been used for forecasting.

The time-series models that are implemented in this project are:

1. Holts Winter Model Prediction
2. Holts Linear Model
3. AR Model Prediction
4. MA Model Prediction
5. ARIMA Model Prediction
6. SARIMA Model Prediction

As part of this project non-time series models are also implemented to check if they gave better predictions.

The non-time series models used were:

1. Linear Regression
2. Polynomial Regression
3. Support Vector Machine

Results:

Root Mean Square Error (RMSE) values are as follows:

|  |  |
| --- | --- |
| Models | RMSE Values |
| Linear Regression | 7830571.030601998 |
| Polynomial Regression | 985900.2801263804 |
| Support Vector Machine | 9793495.12240128 |
| Holts Linear | 172021.58712565133 |
| Holts Winter | 119882.62182205531 |
| AR | 37251.69663112116 |
| MA | 62945.04574070125 |
| ARIMA | 116811.54273309378 |
| SARIMA | 105588.60269296769 |