# Review of COVID-19 outbreak prediction models and datasets

## 1. Prediction Models

**Paper1:** 'Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries' https://www.medrxiv.org/content/10.1101/2020.04.21.20074732v1

### Goal

- Predict the trajectory of cumulated and daily death rate
- Predict the health service needs (hospital admission, ICU admission, length of stay, ventilator need)

### Data sources

- Covid-19 Death info where obtained from the following:
  - https://github.com/CSSEGISandData/COVID-19 -> and aggregated data set for covid19 from different sources around the world that contains worldwide daily reports containing lat, long, confirmed, deaths, and recovered data for each country/state, also it contains a worldwide time series data set for confirmed, recovered, and death cases. The data is available in English and CSV format, also the data is updated daily.
  - https://github.com/pcm-dpc/COVID-19 -> data set from Italian government contained detailed information about covid19 status in Italy in different formats such as csv and json, however all the data is in Italian and there is no English version of the data.
  - https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Gesamt -> data set from the RKI German governmental institution that reports the covid19 statistics in the federal states of Germany such as total number of cases, number of new cases, cases in the last 7 days, and number of deaths. The data is available in both English and German, and the data is available only in pdf format.
  - http://wjw.hubei.gov.cn/fbjd/dtyw/index_1.shtml -> data set from the health commission of Hubei province (that contains Wuhan city) providing information about the situation updates in Hubei province such as the discovered new cases, the total number of cases, the data is in Mandarin only and there is no English version, also the data is available as html page only.
- Data about social distancing policies where gathered from official government declarations
- Data about medical where collected from different governmental and institutional sources, but the sources was not specified.

### Social distancing policies

- school closures
- partial non-essential business closures
- complete non-essential business closures
- restricting group gatherings

- stay-at-home recommendations
- severe local travel restrictions including public transport closures

### Predicting model.

Using curve fitting techniques for nonlinear fitting, with the assumption that the cumulative death rate follows Gaussian error. And for data reach interties a linear extension was developed to increase the accuracy of the model. For predicting the hospital service utilization, the authors used individual level microsimulation model based on location specific ration of admission to death, or prediction of that ratio. Also, the author developed a covariant to measure the effect of social distancing policies on the spread of covid19.

## Paper 2: Application of the ARIMA model on the COVID-2019 epidemic dataset (Work submitted on 14th of February for Data in Brief)

**Objective:** Applying Auto Regressive Integrated Moving Average (ARIMA) models (Auto Regressive, Moving Average, Seasonal AutoRegressive Integrated Moving Average) to predict the epidemiological trend of the prevalence and incidence of COVID-2019.

Prevalence: is the proportion of cases **in the** population at a given time

Incidence: is the number of new cases occurring within a period of time.

Reporting the 95% confidence interval: 95% sure that the mean prediction falls within that interval.

## Paper 3: COVID 19 outbreak reproduction number estimations and forecasting in Marche, Italy

**Objective:** compute different daily transmission rates ($R_t$), I.e. the reproduction number.

It's the average number of people who become infected by an infectious person. If $R_t$ is above 1.0, the virus will spread quickly. When $R_t$ is below 1.0, the virus will stop spreading

Approach: Not clear

## Paper 4: Early forecasting of the potential risk zones of COVID-19 in China's megacities (Science of the Total Environment IF 5.5 )

**Objective:** Estimate the risk of disease based on known case locations and a set of environmental variables that describe some of the factors that likely influence the spread of the epidemic: Population density, demand for public transport (bus stops, subway stations, length of roads), demands of daily life ( rent of rental house, supermarkets, shopping malls) and medical resources ( hospitals). These information are fed into a species distribution model (Maxnet) to determine the areas with high infection risks.

This approach is interesting but requires the appropriate data.

Maxnet package is available with R

## Paper 5: Effect of Weather on COVID-19 Spread in the US: A Prediction Model for Indian 2020 (Science of the Total Environment IF 5.589)

**Objective:** understand the role of weather in transmission of the disease in the highly populated countries, such as India. The study is from **January first to 9th of April**

**Approach**: Plotting the evolution of the number of new cases vs the Temperature and Absolute Humidity. The vulnerable ranges of AH (in g/m3) and T (in C), which led to the highest number of new cases, were identified.

**Findings**:

- From Mar 10, 2020 to Apr 9, 2020, 298,893 new cases are reported. Over 50% of these new cases were observed in states with a **narrow range of AH between 4 and 6g/m3**.
- From Jan 1 to Feb 29, the distribution as per the ranges of T indifferent states did not follow any trend.
- During Mar 1-Mar 30, 2020, over 70% of the new cases were in states with temperatures between 4 and 11 C
- From Mar 31to Apr 9, 2020 interval, approximately 60% new cases were in the states within 4oC –11 C range ==> **No narrow temperature range is identified: Temperature is a poor metric to study the effect of weather on the spread**
- **Map the Absolute Humidity Range to India states to identify the most vulnerable ones.**

## Paper 6: Estimation of COVID-19 prevalence in Italy, Spain, and France (Science of the Total Environment IF 5.589)

**objective:** Auto-Regressive Integrated Moving Average (ARIMA) models were developed to predict the epidemiological trend od prevalence of COVID-19 prevalence of Italy, Spain, and France

Data obtained from John's Hopkins website.

## Paper 7: Forecasting the novel coronavirus COVID-19

**Objective:** Forecasting the confirmed cases as well as analysis of the number of deaths and recoveries globally

Approach: Using Exponential smoothing models

Data obtained from John's Hopkins website.

## Paper 8: Forecasting the prevalence of COVID 19 outbreak in Egypt using nonlinear autoregressive artificial neural networks (Process Safety and Environmental Protection IF 4.3)

**Objective:** Predicting the prevalence of the disease in Egypt

**Approach:** Using ARIMA and Nonlinear autoregressive neural networks to forecast the cumulative number of cases for different time horizons (ten days, one week, 17 days and 20 days)

**Data:** Confirmed cases from 1 March to 10 May 2020.

## Paper 9: ONLINE FORECASTING OF COVID-19 CASES IN NIGERIA USING LIMITED DATA (Data in Brief)

**Objective:** an online forecasting mechanism that streams data from the Nigeria Center for Disease Control to update the parameters of an ensemble model which in turn provides updated COVID 19 forecasts every 24hours.

**Approach**: The ensemble combines ARIMA, Prophet –an additive regression model developed by Facebook, and a Holt-Winters Exponential Smoothing model. Each forecasting methods established an upper and lower bound prediction. The aggregation is conducted by taking the minimum lower bound and the maximum upper bound.

## Paper 10: Real-time estimation and prediction of mortality caused by COVID-19 with patient information-based algorithm (Journal of Total Environment IF 5.5)

**Objective:** Estimating the death rate of a disease in real-time using publicly available data collected during an outbreak.

Approach: Propose a new algorithm, Patient Information Based Algorithm (PIBA):

1- Data collection: patients' information

2- Estimate the average number of days from hospitalization to death by fitting Normal distribution on patient's data (not clear which data). Two parameters are obtained: days from onset of symptoms to death and days from admission to the intensive care unit (ICU) to death (their respective means, mean +- standard deviation and mean +- 2* standard deviation)

3- Use these parameters to calculate the daily mortality.

## Paper 11 Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020

**Objective:** Short-term forecasts in real-time using three phenomenological models using few data. Authors provide 5, 10, and 15 days forecasts for five consecutive days, February 5th through February 9th.

**Approach:** Using models previously applied for SARS, Ebola, pandemic influenza, and dengue: The generalized logistic growth model (**R**efer to **"Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world"** for technical details), The Richards model (Refer to "**Modelling fatality curves of COVID-19 and the effectiveness of intervention strategies**" for technical details) and a sub-epidemic wave model (Refer to "**A novel sub-epidemic modeling framework for short-term forecasting epidemic waves**" for technical details).

These approaches attempt to model the epidemic using differential equations.

Key findings:

- More data are required.
- The width of the prediction intervals decreases on average as more data are included.
- The sub-epidemic model outperforming the other models in terms of mean squared error (MSE)

## Paper 12: Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil (Chaos, Solitons & Fractals IF 3.0)

**Objective:** Forecast one, three, and six-days ahead of the COVID-19 cumulative confirmed cases in ten Brazilian states with a high daily incidence.

**Approach**: Using ARIMA, CUBIST, Random Forest, Ridge Regression, Support Vector Regressor and Ensemble Learning.

**Data**: the cumulative confirmed cases of COVID-19 that occurred in Brazil until April, 18 or 19 of 2020 collected from an API (URL:https://brasil.io/api/dataset/covid19/caso/data/?place_type=state).

**Findings:**

- SVR and stacking-ensemble235learning model are suitable tools to forecast COVID-19 cases.

## Paper 13: Time series forecasting of COVID-19 transmission in Canada using LSTM networks (Chaos, Solitons & Fractals IF 3.0)

**Objective:** forecasting COVID-19 outbreak in Canada for the 2, 4, 6, 8, 10, 12 and 14th day ahead.

**Approach**: Using LSTM, train the network with data until March 31, 2020 reported by Canadian health authority. The motivation of applying deep learning models is that there is no prior assumption compared to well-known statistical model (paper 11 for instance)

Findings:

- Model trained on Canada data: The RMSE error is 34.83 with an accuracy of 93.4% for short term predictions in Canada. Meanwhile, based on our testing/validation data set the RMSE error is about 45.70 with an accuracy of 92.67% for long term predictions.
- Model trained on Italy: the RMSE error is about 51.46 which is higher than previous model. According to this second model within 10 days, Canada is expected to see exponential growth of confirmed cases.

### Paper 14: Time series modelling to forecast the confirmed and recovered cases ofCOVID-19 (Travel Medicine and Infectious Disease IF 4.4)

**Objective**: Forecasting the global confirmed and recovered cases.

**Approach**: Using Autoregressive time series models based on two-piece scale mixture normal distributions

### Paper 15: Multiple-Input Deep Convolutional Neural Network 1Model for COVID-19 Forecasting in China

**Objective**: predict the number of confirmed cases in the next day using data from the previous 5 days for China using CNN.

**Approach**: Feed the six-time sequences composed of factors influencing the cumulative number of cases: Total confirmed cases, Total confirmed new cases, Total cured cases, Total cured new cases, Total deaths and Total new deaths

**Data**: Confirmed cases of COVID-19 from January 23, 2020 to March 2, 2020

### Paper 16: SutteARIMA: Short-term forecasting method, a case: Covid-19 and stock market in Spain (Journal of Total Environment IF 5.5)

**Objective**: predict the short-term of confirmed cases of covid-19 and IBEX in Spain

**Approach**: Using SutteARIMA, a variant of ARIMA model.

Findings: the closing price of the IBEX stock market has decreased from the beginning of Covid-19 in Spain (12 February 2020) and began to stabile on 24 March 2020 in around 6900 per share.

## 2. COVID 19 Datasets

- https://covid-19-apis.postman.com/ -> big list of APIs providing information about covid19 such as the number of infected cases, recovered, death, number of tests, and it has some APIs that provide news from governments, newspapers, WHO, and other health institution. The API is developed and maintained by third parties not postman, and most of the API covers a wide range/All the countries, also there is some API for specific countries such as USA and India. Some interesting API are:
    - https://covid19api.com/ -> Data sourced from Johns Hopkins CSSE
    - https://covid-api.com/ -> Based on public data provided by Johns Hopkins CSSE
    - https://documenter.getpostman.com/view/8854915/SzS7NkAS?version=latest -> Twitter and YouTube API feeds for WHO, and their RSS feeds
- https://github.com/CSSEGISandData/COVID-19 -> git hub repository is owned and updated by Johns Hopkins University CSSE contains an aggregated data set for covid19 from different sources around the world that contains worldwide daily reports containing **lat**, **long**, **confirmed**, **deaths**,

and **recovered**, **Active**, and more data for each country/state, also it contains a worldwide time series data set for **confirmed**, **recovered**, and **death cases**. The data is available in English and CSV format, also the data is updated daily.

- https://www.kaggle.com/sudalairajkumar/covid19-in-india -> data set about covid19 statistics in India, it includes information about infection in different age groups, hospital beds ( **# of district hospital** , **# of public health centers**, **# of beds in public health care centers** , etc. ), ICMR testing labs (**location** , **pin code**, etc. ), personal details for infected/recovered/dead cases (**date of diagnosis**, **age** , **gender** , **city** , **current state**), and the number of cases in India based on state and time, and finally the India census. The data is gathered mainly from governmental sources.
- https://www.kaggle.com/roche-data-science-coalition/uncover -> This dataset is composed of a curated collection of over 200 publicly available COVID-19 related datasets from sources like Johns Hopkins, the WHO, the World Bank, the New York Times, and many others. It includes data on a wide variety of statistics and indicators, like local and national infection rates, global social distancing policies, geospatial data on movement of people, and more. it is very rich data set of size around 1 GB. That data is updated by the respective sources and can be accessed directly on namara https://app.namara.io/#/marketplace which provide an API option for the data. Most of the data on this data set is about USA and Canada, and the global data contains information like daily and cumulated cases information.
- https://www.kaggle.com/imdevskp/corona-virus-report -> a data set containing information about the **number of infected**, **recovered**, **deaths**, **active**, **new cases** and **new deaths** for each country, also it has information about the **critical cases** and **death/infection ration** per 1 million. The data is updated daily, and it is published and updated by an individual.
- https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset -> a data set analyzing the relation between food diets and covid19, the data set include information about **fat quantity**, **energy intake** (kcal), **food supply quantity** (kg), and **protein** for different categories of food such as **Alcoholic Beverage**, **Animal product**, **Aquatic products**, **eggs** , **fish** , etc., also the data set include **obesity** and **undernourished** rate along with covid19 **confirmed**, **recovered**, and **death**. The data is collected from various source such as Food and Agriculture Organization of the United Nations and USDA Center for Nutrition Policy, the data is created maintained by a single individual, and the data set is updated weekly.
- https://www.kaggle.com/kimjihoo/coronavirusdataset -> a very detailed data set for covid19 in Korea, the data set include highly detailed statistics such as **overall infection summary**, all patient information (**sex**, **birth year**, **age group** , **city** , **infection case**), patient movement route (**global number**, **date**, **city**, **long**, **lat**), policy taken by Korean government (**type** , **start** , **end** , **details**), searching trends of the internet from 2016 to current date, the distribution of disease according to age/gender , and weather information in Korea from 2016 to current time. The data is from the Korean government and is updated every 2 weeks.
- https://www.kaggle.com/ishivinal/covid-19-analysis-visualizations-predictions/notebook -> analysis on the relation between covid19 and different features such as lookdown, mean age, human freedom index, fatality rate, tourism and more. the data for this analysis was extracted from the following data sets
    - https://www.kaggle.com/ishivinal/covid19-useful-features-by-country -> offline data for useful covid19 related features such as **Country region, population size**, **tourism**, **first fatality**, **Lockdown type** and more.

    - https://www.kaggle.com/lin0li/covid19testing -> data set containing information about daily and cumulated covid19 data, and the number of tests made, the data set is updated daily and have over 100 country.
    - https://www.kaggle.com/gsutters/the-human-freedom-index -> an offline data set containing information about human freedom index for countries in the world such as indexes for **civil justice**, **criminal justice**, **rule of law**, **homicides**, and **disappearance**.
- https://www.kaggle.com/dheerajmpai/hospitals-and-beds-in-india -> a data set containing information about hospitals in India such as number of health centers, number of governmental hospital and beds, Number of AYUSH hospitals and beds , Number of Hospitals and beds maintained by Ministry of Defence, Number of Railway Hospitals and beds, Number of Employees State Insurance Corporation Hospitals and beds. The data is from the Indian government archives and it is offline data (last updated in March).
- https://www.kaggle.com/rohanrao/covid19-mobility-data -> data set about mobility trends from google and Apple , the mobility trend can be used to measure effect of look down on the movement behavior of individuals, the data is provided by google and Apple , and it was last updated in 15th of May.
- https://covidtracking.com/ -> a web site hosting covid19 data set about the united states, the data is on estate level and it contains current and historical USA covid19 information such as **tests conducted** and their outcome (positive/negative), **recovered**, **deaths**, and more. Also, it has racial information about covid19, such as the distribution of infected/recovered/deaths over the different races in USA at state level. The data is available as csv or json, and it can be fetched through their API, and the data is updated daily.
- https://www.covidexitstrategy.org/ -> contains a data about the performance of US states towards facing covid19, the data captures health system recourse utilization, number of tests made, covid19 statistics (infected, recovered, deaths, growth, etc.). the data is online and being frequently updated and can through a google spread sheet beside the dashboard.
- http://ghdx.healthdata.org/record/ihme-data/gbd-2015-smoking-prevalence-1980-2015 -> data set contains data about smoking between males/females/all for all counties and among the different age groups. The data set is offline and covers from 1980 to 2015. It can be useful as some sties suggest a correlation between smoking and covid19.
- http://www.healthdata.org/data-visualization/mdg-viz -> data set about progress achieved among 188 countries in meeting each of the health-related Millennium Development Goals (MDGs), it is offline data that capture data for 188 countries between 1990 to 2013.
- https://midasnetwork.us/covid-19/ -> a collection of covid19 recourse includes cases line listing for different country (offline most ends at March 2020), parameter estimation data set (such as effective reproduction estimation) from different papers, and a list of dashboards and modeling tools mostly in written in R for coivd19.
- https://www.unacast.com/covid19/social-distancing-scoreboard -> Dashboard showing scores for social distancing rules in different USA states from mobility data, we can use similar techniques with google and Apple mobility data to measure the social distance behavior, the data currently available in USA only, and they are working on developing similar dashboard for worldwide.
- https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/ -> a database by WHO that includes international multilingual scientific findings and knowledge on COVID-19. We

can use this database for searching about relation between different factors (such as humidity) and covid19.

- https://www.semanticscholar.org/cord19 -> a database of free and publicly available scholarly article about covid19, like the WHO database we can use this data set to search for relation between different factors and covid19
- https://discovid.ai/search -> COVID-19 Open Research Dataset (CORD-19), which contains over 57,000 scholarly articles about COVID-19 and related coronaviruses.
- https://arxiv.org/ftp/arxiv/papers/2003/2003.12347.pdf -> A paper describing different techniques and measures for using Mobile data for the analysis of covid19, however the paper assumes the existence of GPS mobile data from crowed but does not provide a data set. Can be useful incase we have raw GPS data and we want to extract information from it.
- https://arxiv.org/pdf/2004.11430.pdf -> A paper studying the relation between travel distance decay and increase of stay at home time with the spread of covid19 using mobile data, the data they used is not publicly available and can be purchased from https://datarade.ai/data-providers/quadrant/datasets/mobile-location-data-6858075c-92d2-4fed-8a85-079cc91a7045 and https://shop.safegraph.com/.
- https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker -> data set about covid19 government response tracker which is developed and maintained by the university of oxford which systematically collects information on several different common policy responses that governments have taken to respond to the pandemic on 17 indicators such as school closures and travel restrictions. It now has data from more than 160 countries. The data is available as csv or through API interface.
- https://ourworldindata.org/coronavirus -> a data set developed and maintained by our word in data organization, the data set It is updated daily and includes data on **confirmed cases**, **deaths**, and **testing**, **GPD per capital**, **extreme poverty**, **CVD death rate**, **smokers**, **hospital beds**, and more. The data can be accessed through this GitHub -> https://github.com/owid/covid-19-data/tree/master/public/data.
- https://datasetsearch.research.google.com/ -> a covid19 dataset search engine by google that can be useful for finding different data sets rending from covid19 statistics to global temperature.
- https://covid19datahub.io/ -> a comprehensive data set about covid19 including information about covid19 statistics (**deaths**, **confirmed**, **tests**, **recovered**), policy measure (**policy name**, **degree and date of activation**) and geographical data (**long**, **lat**).
- https://github.com/covid19-data/covid19-data -> a data repository containing daily covid19 data obtained from our world in Data, and income classification for countries in the world obtained from the world bank.
- https://rs-delve.github.io/data_software/global-dataset.html-> a comprehensive covid19 data set that captures different data that are related to covid19 pandemic and country specific data such as 17 None-pharmaceutical inventions (**school closure**, **travel restrictions**, etc.) , covid19 statistics (**total cases**, **new cases**, **total deaths**, **new deaths**, **etc**.),country specific data (**population**, **median age**, **GDP per capital, smoking, hospital beds per 1k ,** ), mobility data from different sources including google and Apple (such as mobility trends in **grocery, parks, workplaces, etc.**), and weather Data such as (**precipitation, humidity , temperature mean/min/max, wind speed**). The data is updated on daily bases and available as csv and python API.

# 3. COVID-19 Dashboards

## 3.1. Closed Source Dashboards

1. WHO Coronavirus Disease (COVID-19) Dashboard (https://covid19.who.int/)

A React.js based dashboard from the world health organization with the following features:

- World Map shows the following data for each country:
    - a. Confirmed cases
    - b. Confirmed deaths
- Confirmed Cases Over Time graph
- Deaths Over Time graph
- Case Comparison graph for each continent
- Daily Cases graph for each continent
- Situation by Country, Territory or Area
    - a. Daily graph
    - b. Cumulative graph
    - c. Cases and deaths tabs
- Data is refreshed regularly

2. IHME, COVID-19 Projections (https://covid19.healthdata.org/)

A React.js based projection dashboard from the Institute for Health Metrics and Evaluation organization with the following features for each country:

- Total deaths numbers and graph along with its projections for extra 1 month
- Daily deaths and its projections for extra 1 month
- Estimated daily infections
- Confirmed infections
- Tests and projected tests for extra 1 month
- Projected hospital recourses usage for extra 1 month

3. Government Finance Officers Association (https://gfoa.org/dashboard-covid19-prediction-model)
    - A projection dashboard from the Government Finance Officers Association with the following features for the USA:
    - USA 30 Day Forecast of Total and daily COVID-19 deaths for extra 1 month with 95th upper bound percentile and 5th lower bound percentile

4. COVID-19 Dashboards (https://covid19dashboards.com/covid-progress-projections/):
    - A dashboard to reflect the projection of ICU needs by each country.
    - Countries sorted by current ICU demand, split into Growing and Recovering countries by current transmission rate.

5. COVID ActNow dashboard (https://covidactnow.org/us/tx/?s=48058)
    - A dashboard to predict the following rates for the state of Texas, US:

- Infection rate for the upcoming 10 days
- Future Hospitalization (both ICU and non-ICU) Projections for 2 months ahead

6. QCRI GCC Situation Dashboard (https://covid-research.qcri.org/situation/):
   - A dashboard from Qatar Computing Research Institute that reflects the following:
   - General situation of the GCC countries, including the cumulative confirmed, recovered, active, deaths, incident, and mortality.
   - The use of SEIR model to predict the number of confirmed cases from the beginning of the pandemic till the end of September.
   - Comparison between the data of the GCC countries.

7. Honorable mentions:
   - https://www.worldometers.info/coronavirus/
   - https://ncov2019.live/
   - https://www.data.gov.qa/pages/dashboard-covid-19-cases-in-qatar/
   - https://nssac.bii.virginia.edu/covid-19/dashboard/

## 3.2. Open Source Dashboards

1- John Hopkins University & Medicine (https://github.com/CSSEGISandData/COVID-19):

A React.js based dashboard from John Hopkins University & Medicine with the following features:

- World map shows the following data for each country and its states in different tabs:
  a. Cumulative confirmed cases
  b. Active cases
  c. Incident rate
  d. Case-fatality ration
  e. Testing rate
  f. Hospitalization rate
- Data is updated each hour
- Table of confirmed cases for each country
- Deaths vs Recovered comparison
- Confirmed cases Daily and logarithmic graphs

2- https://github.com/soaple/corona-board

3- https://github.com/stevenliuyi/covid19

4- https://github.com/trekhleb/covid-19

5- https://github.com/pabloVinicius/covid-19-dashboard

## References