ABSTRACT

In the final months of 2019, a cluster of viral pneumonia cases were reported in Wuhan, China and an investigation was launched. Just a month later, the WHO (World Health Organisation) concluded that these cases were caused by a new strain of a virus that caused severe acute respiratory syndrome (SARS). The outbreak caused by viral contagion was then classified as a pandemic by WHO in March 2020, and was formally named as SARS-CoV-2 (commonly known as the coronavirus disease, or simply COVID-19); it has now infected over 21.7 million people and resulted in the deaths of almost 770,000 people.

This aims of the study are three-fold: (a) model the community spread of the coronavirus; (b) to generate a real-time short-term forecast of 10 days of the top 3 countries with the highest incidence of confirmed cases (the USA, Brazil and India); (c) qualitatively determine and rank algorithms that are best suited for precise modelling of the linear and non-linear features of the time series.

The comparison of forecasting procedures for the total cumulative cases of each country was done by comparing the reported data and the predicted value, and then ranking the algorithms (Prophet, Holt-Winters, LSTM, ARIMA, DWT-ARIMA) based on their RMSE (Root Mean Squared Error).

INTRODUCTION

On the 31st of December 2019, a cluster of cases of pneumonia of unknown cause, in the city of Wuhan, Hubei province in China, was reported to the World Health Organisation. In January 2020, a previously unknown virus was identified[1], subsequently named the 2019 novel coronavirus, and samples obtained from cases and analysis of the virus' genetics indicated that this was the cause of the outbreak. This novel coronavirus was named Coronavirus Disease 2019 (COVID-19) by WHO in February 2020[2]. The virus is referred to as SARS-CoV-2 and the associated disease is COVID-19[3].

The ground-zero for the zoonotic spillover has been triangulated to the live-food markets of Wuhan[4], where the virus spread proximally due to direct exposure to animal shedding, bodily fluids, blood, and secretions[5]. The super-spreading of the virus is attributed to the widespread travel of people in the Sinosphere for the celebration of the Chinese New Year in late January.

SARS-CoV-2 presents at least six different strains of the virion that have been identified to have mutated from the original L strain[6] which was detected in Wuhan in 2019. Due to its rapid spreading potential (with an R_0 of 1.8 in India)[7] and in the absence of any treatment procedures, the contagious COVID-19 has caused a surge in the infection rates around the world and has ruptured the concept of normal life. The virion is responsible for infecting more than 21 million people, killing over 7 hundred thousand people and has forced more than 3 billion to isolate themselves in the safety of their homes[8].

Rapid and predictable upscaling of the healthcare framework is critical towards ensuring the availability of appropriate facilities during these unfortunate conditions of a pandemic. Some countries may exhibit the need to divert their manpower into the assembly of physical infrastructure, whereas others may be in the need for securing the capital for further investment in the aforementioned infrastructure. Due to these and various other geo-political reasons, true and accurate modelling of the community spread of infections is foremost in the national and international scope.

Day by day, the adverse effects of the government imposed lockdowns become more evident; with over a third of the world's population being forced into lockdown, the economy has slumped into a worldwide recession[9]. Stock markets have crashed, workers are being furloughed and entire companies are falling into the clutches of bankruptcy. Unemployment rates are soaring in countries like the USA, India and Brazil with 14.7%, 23.5% and 12.2% respectively[10]. With annual income losses due to viral and bacterium based infections being as high as 12.2% & 25.2% in urban and rural areas respectively[11], developments in the line of forecasting of epidemics and pandemics are crucial now more than ever. They are responsible for providing insights into the severity of the infection and its spreadability. This is crucial for evolving the acknowledgement of the severity of the pandemic in the common folk; simplified dashboards which summarises the gist of such predictions can prompt the lawmakers to take apt decisions in the future on these issues in question.

We propose the use of multiple time-series forecasting methods to create a hybrid model that can selectively cope with the linear and non-linear features of the time series. A comparison was made with multiple other forecasting tools and methods, which indicated the best-suited model for non-stationary time-series.

METHODOLOGY

1. **Data Collection**

   The team focused the analysis on 3 countries that were affected the worst by COVID-19, which were the USA, Brazil and India. The univariate time-series data of total cases of incidence was collected through the dataset published by John Hopkins University's Centre for System Science and Engineering[12]. The models were analysed on three different time intervals: (a) 22nd *January* - 15th *May*, (b) 22nd *January* - 30th *July* and (c) 22nd *January* - 10th *August*, and were ranked accordingly based on their performance metric (RMSE). Each interval was split into training and test datasets, with the last 10 days of each interval being reserved as the test dataset, and the rest being used for training.

2. **Proposed Model**

   The model being proposed was derived by quantitatively analyzing hand-picked models and adopting a model that is best suited for a stationary time-series with linear and non-linear features. The models that were chosen were,

   2.1. **ARIMA**

   Standing for Auto-Regresive Integrated Moving Average, it was proposed by Box and Jenkins in the 1970s[13] as a model which took varying trends, seasonal changes and random disturbances in account to predict the future values of the series, due to these reasons, today, it is one of the most popular models that is used for forecasting time-series. It is denoted as ARIMA(p,d,q) where p and q are the orders of the AR and MA models respectively, and d is the level of differencing. The model can be mathematically represented as,

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q},$$

   where y_t denotes the computed value of at the given time *t*, ϕ_i and θ_j are the coefficient of the AR and MA models respectively and ε_t is the random error occurring at time *t*.

REFERENCES

1. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/
2. https://www.who.int/dg/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020
3. https://www.gov.uk/government/publications/wuhan-novel-coronavirus-background-information/wuhan-novel-coronavirus-epidemiology-virology-and-clinical-features
4. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team (February 2020). "[The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]". Zhonghua Liu Xing Bing Xue Za Zhi = Zhonghua Liuxingbingxue Zazhi (in Chinese). 41 (2): 145–151. doi:10.3760/cma.j.issn.0254-6450.2020.02.003. PMID 32064853. S2CID 211133882.
5. Kevin Berger. “The Man Who Saw the Pandemic Coming”. Nautilus. Issue 83.
6. Mercatelli D and Giorgi FM (2020) Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front. Microbiol. 11:1800. doi: 10.3389/fmicb.2020.01800
7. Seema Patrikar, Deepti Poojary et al, “Projections for novel coronavirus (COVID-19) and evaluation of epidemic response strategies for India”, Medical Journal Armed Forces India, vol. 76, no. 3, 268-275, 2020
8. https://www.who.int/emergencies/diseases/novel-coronavirus-2019
9. "World Economic Outlook, April 2020: The Great Lockdown". IMF.
10. "Economic data, commodities and markets". The Economist. ISSN 0013-0613
11. Amrita Ghatak & S Madheswaran, 2011. "Burden of Income Loss Due to Ailment in India: Evidence from NSS Data," Working Papers 269 Classification- JEL C, Institute for Social and Economic Change, Bangalore.
12. CSSEGISandData. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. 2020. GitHub repository. https://github.com/CSSEGISandData/COVID-19
13. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.