International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 463-467
Published Online June 2020 in IJEAST (http://www.ijeast.com)

GLOBAL FORECASTING OF COVID-19 USING ARIMA BASED FB-PROPHET

Dr. Shikha Gaur
Department of Mathematical Sciences
NMIMS, Mumbai, Maharashtra, India

Abstract— **In this paper, global outbreak of COVID-19 is analyzed. Forecasting is a task that helps system with measurements planning and irregularity detection. Regardless of its importance, there are encounters associated with producing reliable and high -quality forecasts, especially when there are a variety of time series and analysts. To address these encounters, we describe a practical approach to forecasting "at scale" that combines configurable models with analyst-in-the-loop performance analysis. We propose a modular regression model with interpretable parameters that can be intuitively adjusted by analysts with domain knowledge about the time series. We describe performance analyses to compare and evaluate forecasting procedures. Comparison of diverse forecasting schemes are depicted, outcome of the same is discussed.**

Keywords—**Forecasting, ARIMA, Prophet, COVID-19, Global.**

I. INTRODUCTION

The preciseness of traditional forecasting largely depends on the availability of data to base its forecasts and estimates of ambiguity. In outbreaks of epidemics there is no such data at all in the commencement and then limited as time passes, making predictions widely uncertain. On February 18, 2020, a New York Times article(BBC(2020-02-18)) cautioned against excessive optimism about the crisis peaking, even though there were close to 50 days since the virus had been identified.

Besides, there are concerns that the data may not be reliable, as was t0he case of bird flu and SARS when the number of affected people and deaths were misreported to hide the extent of the epidemic. Similarly, in the case of COVID-19, the reporting did not reflect the correct numbers as well when on the February 13 a new category of "clinically diagnosed" was added to "lab-confirmed" ones(BBC(2020-02-18)). Such problems decrease forecasting accuracy and increase uncertainty, making the drawing of definite conclusions more difficult.

Related to forecasting accuracy and uncertainty, there is a more severe problem that has to do the perception of epidemics and pandemics. Systems are concerned with regards to the measures to be taken while the general population fears about the impact on the epidemic on their health/lives. Furthermore, the pharmaceutical firms are working on vaccinations for the new virus with considerable commercial interest. This was the case with SARS when governments persuaded on the severity of the virus bought large numbers of vaccines that were never used as its spread stopped without the need to vaccinate people.

Of course, the big problem is the asymmetry of risks and the irrational fear of a pandemic with its possible catastrophic consequences, as happened with the 1918 Spanish flu that killed an estimated 50 million worldwide. In contrast, the SARS killed a total of 774 in 2003 and the bird flu around 100 in 1997. COVID-19 has resulted in an estimated 253218 deaths until now (05/05/2020). At the same time, there is much less concern over the seasonal flu that kills about 646,000 people worldwide each year Medicine Net(2020-02-19).

Medical predictions are often not accurate while their uncertainty is seriously underestimated(Makridakis S, Wakefield A, Kirkham R(2019)). Predicting the future of epidemics and pandemics is much more difficult as the number of cases to be studied can be measured in one hand. At one end of the scale is the case of SARS where the fear of becoming a pandemic was overblown, resulting in overspending and the application of restrictive measures to be contained that it turned out to be unnecessary. At the other end is the Spanish flu that turned out to be a serious pandemic with catastrophic consequences, arguably in a different era when communication and the ability to raise public awareness (and possibly exaggerated fear) were limited.

Despite the inaccuracies associated with medical predictions, still forecasting is invaluable in allowing us to better understand the current situation and plan for the future. In this paper, we provide statistical forecasts for the confirmed cases of COVID-19 using ARIMA by FB-Prophet model. The comparison of same is done with other models to depict the best forecasting by proposed model.

II. ANALYSIS AND FORECASTING

We focus on the cumulative daily figures aggregated globally of the three main variables of interest: confirmed cases, deaths and recoveries. These were retrieved by the GitHub data set (https://github.com/datasets/covid-19) and are presented in

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 463-467
Published Online June 2020 in IJEAST (http://www.ijeast.com)

Fig-1. The data refer to daily cumulative cases and cover the period from January 22, 2020 until May 19, 2020. While all three data patterns show an exponential growth, the trends all three cases and the deaths were reduced in the mid of March; a second exponential increase is observed in late March and April as a result of the increased number of cases in USA, Spain, and Italy. At the same time, the number of recovered cases is steadily increasing.

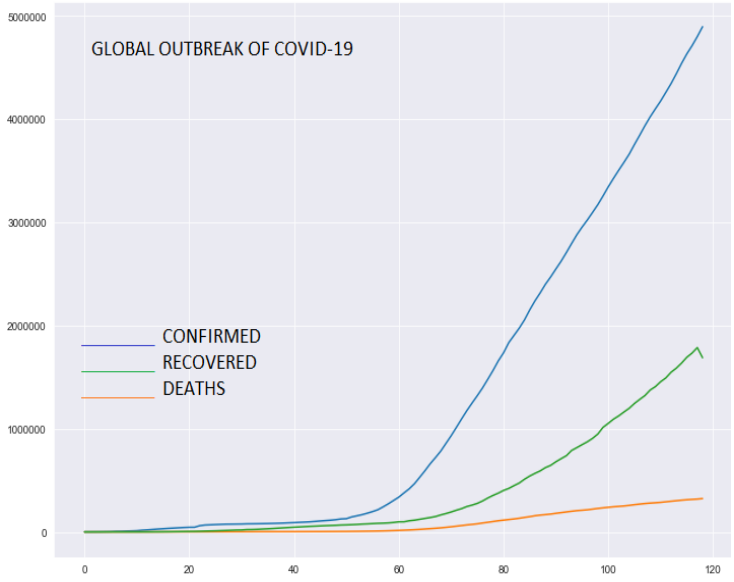

Fig. 1. Day-wise Global Analysis of COVID-19

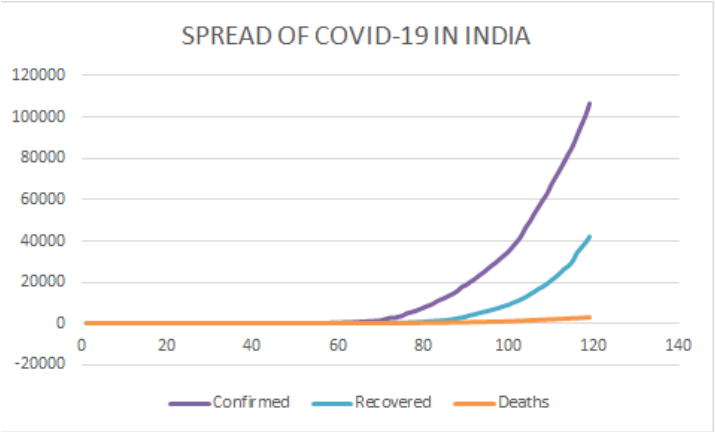

Fig. 2. Spread of COVID -19 in India

To understand the current status of COVID-19 spread during the first two weeks of March, model-based estimates were developed for India based on India centric data (Arni S. R. Srinivasa Rao et. al. (10 APRIL 2020)). It can be easily observed in Fig-2 that all three curves that is confirmed, recovered and deaths curves seems only stretched out in time but not changed its nature, which shows that effective application of quarantine.

It is significant to project the future to understand how much population are expected to get affected by the disease. This could help better prepare and ensuring there are resources to handle disease growth. We can use historical data points to help us understand. When we try and make such projections based only on time and values this is known as a time series analysis. We first look for the general trend in our time series to understand whether there has been an increase or decrease in values over time. Secondly, we look at cyclical fluctuations. You can think of these as patterns of changes. The duration of its nature is dependent on the type of spread. We thirdly evaluate the seasonal trends in our time series, these are oscillations that occur daily, month or year- depending on the nature of the forecast. Lastly when the seasonality, trend and cyclical fluctuations are removed from our time series, we are left with the residual effects or irregularities. You can think of these as unpredictable events resulting in outlier highs or lows. By their nature these are unsystematic and have no clear patterns but need to be factored into our forecasting model.

A. ARIMA

For this analysis, chose a dataset on global data of COVID-19. Disease spread in this context refers to the number of confirmed cases per day over the course of the defined period. My aim is to forecast disease over the few months. This would presumably help in the preparation to face this Pandemic. Here we chose to use a regression model i.e. Autoregressive Integrated Moving Average (ARIMA) because of the relative simplicity of the model, its flexibility and its ability to perform relatively well in forecasting tasks. Let's start off by breaking apart the shortening to its individual components, AR, I and MA. The AR part simply speaks about autocorrelation using the past to explain current data. We assume that disease spread is dependent, at least to a certain extent, on already infected population. With autocorrelation we are trying to find out how many lags/periods of time best predict our current numbers. The 'I' refers to using differencing to attain stationarity. Stationarity means the statistical properties of the time series such as the median, variance and correlation remain stationary over time. It is more difficult to carry out forecasting if every parameter varies. It is important for us to find common statistical properties to best look into the future. Differencing refers to the number of times we need to difference (subtract an observation from an observation at the previous time step) the time series against itself to attain stationarity. When we try and forecast a value using a forecasting model, there will be a delta between our prediction and the actual value. The MA part can be thought of as a collection of these error terms. We want to incorporate these error terms to factor in random fluctuations/irregularities when making our forecast.

B. IMPLEMENTING ARIMA

From my research, python seemed to be best suited for implementing time-series forecasting so we chose to this for my forecasting task. After loading the relevant libraries and my dataset, I grouped the time series by date. It is then filtered out dates after the 22 January 2020, as this was the first case reported officially. It is assumed that this would not be valuable for time series analysis. Then converted the numerical values in my dataset to a time series, specifying daily confirmed cases and the beginning date. This returns a time series of daily values from January 2020 to June 2020.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 463-467
Published Online June 2020 in IJEAST (http://www.ijeast.com)

C. TRAINING AND TEST SPLIT

We need to be able to test the accuracy of predictions, in order to do this, we created a training and test data. Test data will be used to gauge how well predictions made by the model then trained with ARIMA perform against observed customer flow.

D. FORECASTING WITH ARIMA

Three parameters need to be tuned to find the best ARIMA fit-p (AR),q(MA),d(I). Since this can be a time consuming, I instead opted to use auto ARIMA . This function returns the best ARIMA parameters. After adding the prescribed parameters, we created a forecast for three weeks into the future, we then proceeded to plot these values, also making sure to plot test observations to get a rough idea of the accuracy of this model.

E. TESTING MODEL ACCURACY

The accuracy function is available to diagnose the fit of the ARIMA model by printing out multiple measures of accuracy. To assess model fit chose to use the Root Mean Squared Error for Prophet Model

$$\frac{1}{N}\sqrt{\sum_{t=1}^{N}|Actual_t - Forecast_t|^2}$$

For both these values the lower the values the better the model fit. The Root Mean Squared Error for Prophet Model 8145.463234015535. I can compare this with a traditional Linear Regression forecast to assess whether the ARIMA is an improvement over naive forecasting. Naive forecasting approach produces forecasts equivalent to the last observed value. Seasonal naive goes a step further by factoring in seasonality. Each prediction becomes equivalent to the last observed value of the same season. This visibly performs worse on my test series further validating the goodness of fit of my ARIMA model.

F. FB-PROPHET

An interesting alternative to my forecasting problem is using an open source package created by Facebook that makes the task of forecasting more accessible and easier to carry out. The great part of fb-prophet is that it automatically detects change points in your time series and allows you to factor in hourly, daily, weekly, monthly, yearly trends and even allows you to factor in changes of trends during pre-defined holidays to make you forecast more accurate. After a few trials, I got the best model from turning on weekly seasonality and setting seasonality mode as multiplicative. This improved the accuracy of my model because seasonality grows with the trend in my time series, it is not a constant additive.

III. FORECAST

The forecasted value reveals both the trends, the upper and lower bound forecasts (lowest and highest forecasted values). To forecast confirmed cases of COVID-19, we adopt simple time series forecasting approaches. We produce forecasts using models from the exponential smoothing family (Hyndman RJ, et. al. 2002)), (Taylor JW(2003)).This family has shown good forecast accuracy over several forecasting competitions(Makridakis S, et. al. (2000) Carbone R, et al(1982), Makridakis S, et. al. (2020)) and is especially suitable for short series. Exponential smoothing models can capture a variety of trend and seasonal forecasting patterns (such as additive or multiplicative) and combinations of those. We limit our attention to trended and non-seasonal models, given the patterns observed in Fig-1. Note that we follow a pragmatic approach in that we assume that the trend will continue indefinitely in the future. This approach contradicts other modelling approaches to COVID-19 using an S-Curve model (logistics curve) that assumes convergence. While statistical approaches to model selection (such as information criteria, which measure the maximum likelihood of a model while penalizing for its complexity) could be used, we judgmentally select a model(Petropoulos F, et. al. (2018)) to better reflect the nature of the data. We opt for an exponential smoothing model with multiplicative error and multiplicative trend components. Even if in some cases an additive trend model gave lower information criteria values, we opted for the multiplicative trend model given the asymmetric risks involved as we believe that it is better to err to the positive direction. We produce 30-days-ahead point forecasts and prediction intervals and update our forecasts every 30 days.

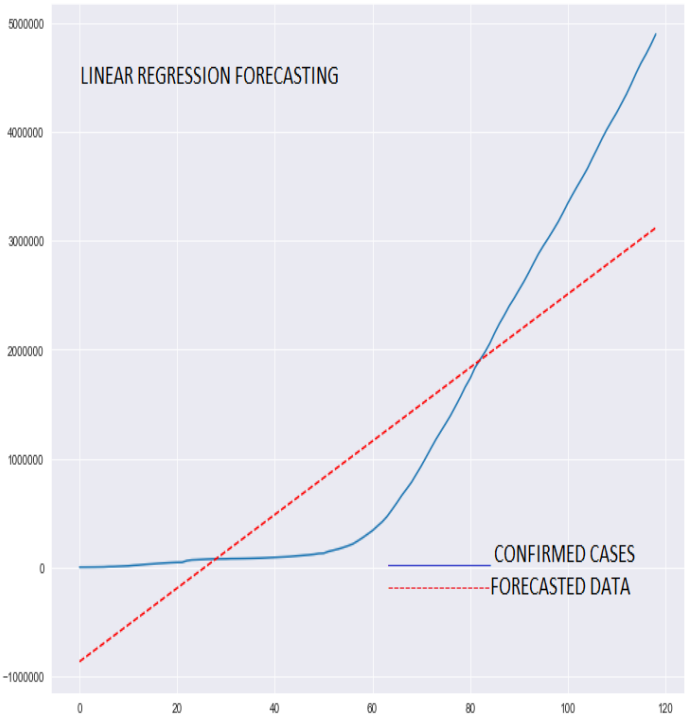

Fig. 3. Traditional Linear Regression Forecasting

We produced a set of forecasts and prediction intervals using the most recent data, up until 04/05/2020. These are presented in Fig-3 with red dotted and blue solid lines. The trend of our forecasts is not fitted with traditional linear regression. In Fig-4 ARIMA using FB-Prophet is used. We predict the spread will reach up-to 50 Lakhs by third week of May, new cases for this round (a total of 15 Lakhs cases). The associated levels of

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 463-467
Published Online June 2020 in IJEAST (http://www.ijeast.com)

uncertainty are also increased: There is a chance that the total confirmed cases will exceed 50 Lakhs by the end of May 2020.

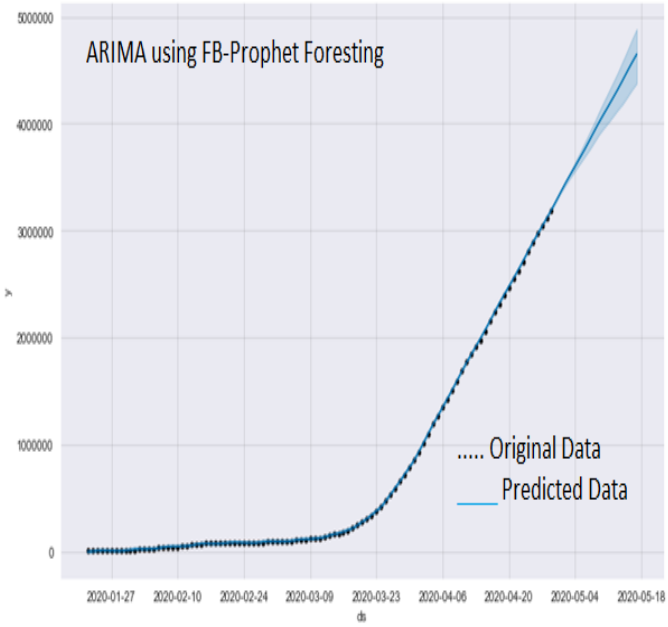

Fig. 4. Forecasting By Arima Using FB-Prophet Model.

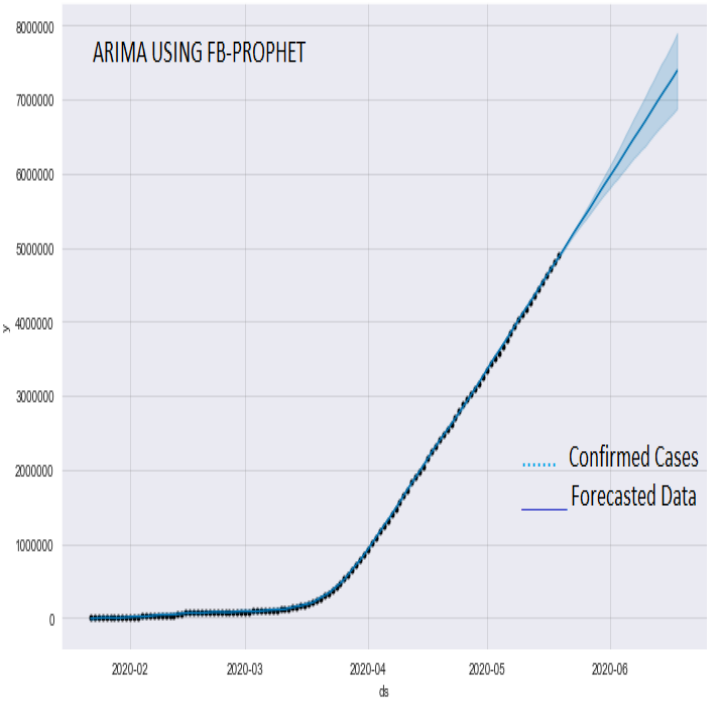

Fig. 5. SATEG -II FORECASTING BY ARIMA USING FB-PROPHET MODEL.

We also attempted to produce analysis within India and compared with the key countries. As the trends into these two groups are different. We notice that using the strategies adopted by India shows effective control in spread of virus at larger scale this approach. However, the estimated uncertainty by splitting the data is considerably lower, possibly since the confirmed cases outside India have significantly increased only recently. Fig- 6 clearly depict that timely lockdown in India had drastically controlled the spread of virus a large scale. It can be observed that with the time stretch the confirmed cases are not shooting whereas in other key countries the control has to take place by adopting good strategies.

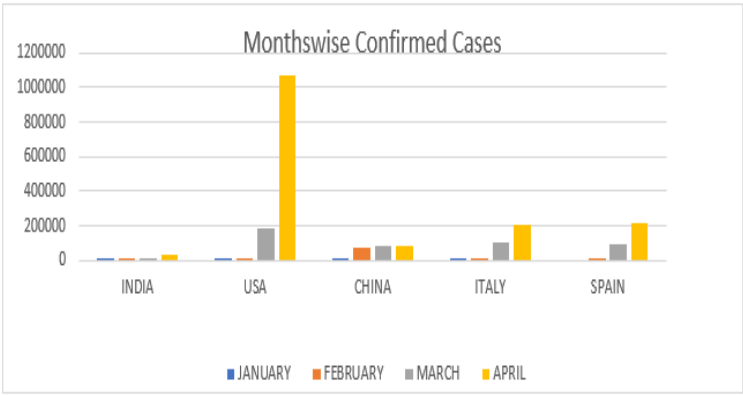

Fig. 6. Comparison of Month-wise Confirmed Cases in Key Countries

IV. CONCLUSION

The uncertainty surrounding an unknown, novel coronavirus can spark a global alarm, leading a Harvard Professor stating that 40-70% of the global population might be infected in the coming year (Nash C. Mediaite(2020-02-18)) which matches Chancellor Angela Merkel's warning regarding the effects of the novel coronavirus in Germany (BBC News(2020-03-15)) . Norman J, Bar-Yam Y, Taleb NN(January 26, 2020) discuss the systemic risk of pandemics, the existence of fat-tailed processes due to global interconnectivity and the negatively biased estimates of spread, reproduction and mortality rates. On the opposite side, others are arguing about people overly panicking (Salzberg S. Forbes(2020-03-14)) and neglecting the probabilities (Sunstein CR. Bloomberg(2020-03-14), Della Cava M(2020- 03-14)) with the new virus being the first "info emic" as a result of the hyper-connectivity offered by today's social media BBC News(2020-03-14). The polarisation of the opinions globally can be summarised by the quotes of three renowned personalities.

We believe that the significant forecast error at the end of the forecast period (from 22/01/2020 to 20/06/2020) as depicted in Fig-4 could be the result of two factors:

- While the forecasts that we produced using the data up until 20/06/2020 would be a good estimate in the scenario of pandemic, they disregard the fact that the world *will* act to get the virus under control. The World Health Organisation helped in creating

awareness of the novel virus. So, the decline in the spread of the COVID-19 during this could well be linked with these attempts from local and global authorities.

- There may be a "garbage-in, garbage-out" situation. As mentioned above, our analysis and forecasts assumed that the data are accurate. It could be the case that the positive bias of the linear regression forecasts is not as significant as it seems dues to potential inaccuracies in the actual data and the under-accounting of confirmed cases. This is especially true given the delay effects in diagnosing COVID-19.

Our ARIMA sets of forecasts that cover the period 22/01/2020 to 20/06/2020 were very close to the recorded confirmed cases (the forecast error was lower than 5 thousand cases at the end of each 30-day period). The slowing down of the trend during this period suggested that COVID-19 would not cause any serious problems if this lockdown and social distancing maintained properly. Though the sets of forecasts that cover the period 02/03/2020 to 20/06/2020 show a significant increase in the trend of cases globally coupled with an increase in the associated uncertainty. We hope that our forecasts will be a useful tool for governments and individuals towards making decisions and taking the appropriate actions to contain the spreading of the virus to the degree possible.

ACKNOWLEDGEMENT- I thank School of Mathematical Sciences, NMIMS for constructive comments which helped me in revising the article.

V. REFERENCE

[1] Wang V(2020-02-19). Coronavirus epidemic keeps growing, but spread in China slows. New York Times. [https:// www.nytimes.com/2020/02/18/world/asia/china-coronavirus cases.html?referringSource=articleShare].

[2] BBC(2020-02-18). Coronavirus: Sharp increase in deaths and cases in Hubei. [https://www.bbc.co.uk/news/worldasia-china-51482994].

[3] Medicine Net(2020-02-19). Flu kills 646,000 people worldwide each year: Study finds. [https://www.medicinenet. com/script/main/art.asp?articlekey=208914].

[4] Makridakis S, Wakefield A, Kirkham R(2019). Predicting medical risks and appreciating uncertainty- Foresight: The International Journal of Applied Forecasting.; 52:28–35.

[5] Arni S. R. Srinivasa Rao, Steven G. Krantz, Thomas Kurien, Ramesh Bhat, Sudhakar Kurapati(10 APRIL 2020), Model-based retrospective estimates for COVID-19 or coronavirus in India: continued efforts required to contain the virus spread, Current Science, Scientific Correspondence, VOL. 118, NO. 7, 1023-1025.

[6] Hyndman RJ, Koehler AB, Snyder RD, Grose S(2002). A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting. 18(3):439–454.

[7] Taylor JW(2003). Exponential smoothing with a damped multiplicative trend. International Journal of Forecasting. 19(4):715–725.

[8] Makridakis S, Andersen A, Carbone R, Fildes R, Hibon M, Lewandowski R, et al(1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting. 1(2):111–153.

[9] Makridakis S, Hibon M(2000). The M3-Competition: results, conclusions and implications. International Journal of Forecasting. 16(4):451–476. https://doi.org/10.1016/S0169-2070(00)00057-1.

[10] Makridakis S, Spiliotis E, Assimakopoulos V(2020). The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting. 36(1):54–74.

[11] Petropoulos F, Kourentzes N, Nikolopoulos K, Siemsen E(2018). Judgmental selection of forecasting models. Journal of Operations Management. 60:34–46.

[12] Nash C. Mediaite(2020-02-18) Harvard Professor Sounds Alarm on 'Likely' Coronavirus Pandemic: 40% to 70% of World Could Be Infected This Year. [https://www.mediaite.com/news/harvard-professor-sounds-alarmon-likely-coronavirus-pandemic-40-to-70-of-world-could-be-infected-this-year/].

[13] BBC News(2020-03-15). Coronavirus: Up to 70% of Germany could become infected—Merkel [https://www.bbc.co. uk/news/world-us-canada-51835856].

[14] Norman J, Bar-Yam Y, Taleb NN(January 26, 2020). Systemic Risk of Pandemic via Novel Pathogens—Coronavirus: A Note. New England Complex Systems Institute.

[15] Salzberg S. Forbes(2020-03-14). Coronavirus: There Are Better Things To Do Than Panic. [https://www.forbes.com/ sites/stevensalzberg/2020/02/29/coronavirus-time-to-panic-yet].

[16] Sunstein CR. Bloomberg(2020-03-14). The Cognitive Bias That Makes Us Panic About Coronavirus. [https://www. bloomberg.com/opinion/articles/2020-02-28/coronavirus-panic-caused-by-probability-neglect].

[17] Della Cava M(2020- 03-14). USA Today. Coronavirus and its global sweep stokes fear over facts. Experts say it's unlikely to produce 'apocalyptic scenario'. [https://eu.usatoday.com/story/news/nation/2020/02/27/ coronavirus-experts-urge-sharing-facts-covid-19-spreads-worldwide/4892422002/].

[18] BBC News(2020-03-14). Coronavirus in Europe: Epidemic or 'infodemic?. [https://www.bbc.co.uk/news/worldeurope-51658511].