ARTICLE

Prediction of daily COVID-19 cases in European countries using automatic ARIMA model

Tahir Mumtaz Awan, Faheem Aslam

Department of Management Sciences, COMSATS University, Islamabad, Pakistan

Abstract

The recent pandemic (COVID-19) emerged in Wuhan city of China and after causing a lot of destruction there recently changed its epicenter to Europe. There are countless people affected and reported cases are increasing day by day. Predictive models need to consider previous reported cases and forecast the upcoming number of cases. Automatic ARIMA, one of the predictive models used for forecasting contagions, was used in this study to predict the number of confirmed cases for next 10 days in four top European countries through R package "forecast". The study finds that Auto ARIMA applied on the sample satisfactorily forecasts the confirmed cases of coronavirus for next ten days. The confirmed cases for the four countries show an increasing trend for the next ten days with Spain with a highest number of expected new confirmed cases, followed by Germany and France. Italy is expected to have lowest number of new confirmed cases among the four countries.

Introduction

A growing list of countries are locked down, governments are ordering residents to self-quarantine themselves by staying inside their homes during coronavirus pandemic (COVID-19). According to recent statistics, up to the 23rd of March 2020, COVID-19 has spread over 168 countries, with 360,697 confirmed infections, 15,495 deaths and 100,471 recovered cases in 168 countries all over the world. The top countries in terms of total confirmed infections are China, Italy, USA, Spain, Germany, Iran, France, and South Korea, whereas in terms of deaths the top countries are Italy, the Hubei province of China, Spain, Iran, France, UK, The Netherlands, and Switzerland. According to the Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at John Hopkins University (JHU) Coronavirus Resource Center,[1] the highest numbers of recovered patients are found in the Hubei province of China, Iran, Italy, Spain, South Korea, France, the Guangdong and the Hunan province of China. The data is available at the repository for COVID-19 (https://systems.jhu.edu/research/public-health/ncov/) operated by JHU-CSSE supported by Esri Living Atlas Team and

Applied Physics Lab of JHU along live news dashboard available at https://visualizenow.org/corona-news. COVID-19 was initially named Novel Coronavirus (2019-nCoV) by the National Institute of Viral Disease Control and Prevention (IVDC) on 3rd of January 2020,[2] on 11th February 2020 the name given by The World Health Organization became COVID-19,[3,4] whereas the virus itself is named as SARS-CoV-2. This deadly epidemic was then declared as a pandemic by The World Health Organization.[5,6] After spreading mass destruction in China, especially in the Hubei province from where it was originated,[7] it now moved its epicenter to Europe.[8] This virus-related diseases has a history of outbreaks in 2018 (MERS-CoV) with 41 deaths in Saudi Arabia,[9] 2015 (MERS-CoV) with 36 deaths in South Korea,[10] 2012 (MERS-CoV) with over 400 deaths,[11] and 2003 (SARS-CoV) with about 774 deaths.[12] As of 23rd March 2020, the reported cases in EU/EEA and the UK account for 160,233 cases in total and 8622 fatalities, with Italy at the top with 59,138 cases and 5476 deaths followed by Spain, Germany and France with 28,572, 24,774, and 16,018 cases.[13] In these kinds of pandemic outbreaks, the importance of performing some kind of forecasting is rampant in many scientific and engineering disciplines.[14] The attempt to use statistical methods for predictions holds great importance as it helps the authorities for necessary arrangements and allowed timely response, which ultimately may reduce losses of lives in the case of this recent pandemic. Auto-Regressive Integrated Moving Average (ARIMA) is one of the forecasting models applied for future predictions using time series data. Its application is noticed in various domains, *e.g.* to predict next day electricity prices,[15] to forecast primary energy demand,[16] to predict stock prices,[17] to predict water quality,[18] to forecast traffic flow,[19] along with its application in medical science in general and specifically epidemics[14,20-25] to fulfill the purpose of prediction or forecasting various issues. Specifically, a recent article about COVID-19[26] used ARIMA model and predicted the epidemiological trend of the prevalence and incidence of the pandemic. Another article[27] showed similar kinds of results regarding COVID-19. However, this study makes predictions for the next seven days. This study primarily focuses at forecasting the confirmed cases of European countries. ARIMA technique was used for this purpose. The confirmed cases of COVID-19 data was used till 21st of March 2020 and predictions of upcoming one week. The materials and methods section below

Significance for public health

World Health Organization (WHO) and medical authorities all over the world and specially in the European countries are busy in taking appropriate measures against COVID-19. It is important to do proper planning and the success is dependent on the arrangements that will be made in near future to stop the spread of this disease. This study by prediction of upcoming cases will help the authorities to plan accordingly, i.e. to arrange appropriate number of medical facilities. Similar approximations for other parts of the world can be made following the methodology used in this paper and better medical arrangements can be ensured. Overall, such kind of research play an important role for policy making and making task forces to combat against epidemics.

discussed in detail about the forecasting mechanism. The results are discussed afterwards which are based on 80% and 95% confidence interval. The final section includes the conclusions and the implications and recommendations of the study for government departments and health ministries of the European countries, so that they can take preventive measures and quick policy decisions can be taken to overcome this deadly pandemic.

Design and Methods

In this study ARIMA technique was used to estimate the upcoming cases of COVID-19 in the European countries. For this purpose, time series data of daily confirmed cases of coronavirus emerging in the said countries was considered. The ARIMA is one of the most popular models for time series forecasting analysis, and has been originated from the combination of autoregressive model (AR), the moving average model (MA). The ARIMA model is used for stationary time series data, *i.e.* when there are no missing values. An identified underlying process based on observations is generated in ARIMA analysis to produce a precise process-generating mechanism resulting in a good model.[28] The ARIMA analysis includes identification estimation, and diagnostic checking.[29,30] It general ARIMA model is viewed as a filter that tries to separate signal from noise, and the signal further helps to extrapolate the future for obtaining forecasts.

Data

The data for this study was taken from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports, a repository maintained by Center for Systems Science and Engineering (CSSE) at John Hopkins University (JHU) Coronavirus Resource Center through GitHub pull request.

The data about the reported COVID-19 pandemic cases of four European countries was used for this study for the following reasons: i) Europe is at high risk because of its population density and its business connections all over the world; ii) European countries exhibited a high peak of cases in the recent days. The daily data for most affected countries namely, Italy, Germany, France and Spain are collected from January 22nd, 2020 to March 28th, 2020, which

corresponds to 66 observations. The selection of these four countries is done on the basis of highest daily growth (Δ [Xn-Xn-1]) as it shows as non-constant growth of the daily confirmed cases, which is calculated by taking the first difference.

Methodology

ARIMA is a frequently used technique for forecasting using the time series data, specified by three order parameters: *p, d, q*, where *p* stands for the order of auto regressive model, *d* is the order of differencing and *q* represents the order of moving average. The procedure of fitting an ARIMA model is also referred as the Box-Jenkins method,[31] where p, d and q are the orders of the AR part, the Difference and the MA part respectively. AR is a class of linear model where the variable of interest is regressed on its own lagged values. If y_t is modeled *via* AR process, it can be written as:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t \tag{1}$$

where, δ is intercept; y_{t-i} are regressors; ϕ_{t-i} are and ϵ is an error term ($\epsilon\epsilon$).

MA is another class of linear model. In MA, the output or the variable of interest is modeled via its own imperfectly predicted values of current and previous times. It can be written as follows in terms of error terms:

$$y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{2}$$

The mathematical form of ARMA (p,q) is as follows:

$$y_t = \{\delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p}\} + \{\theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}\} + \epsilon_t \tag{3}$$

In short, we can rewrite the above equation as:

$$y_t = \delta + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{4}$$

Table 1. Country-wise Best Model Selection using auto.ARIMA.

Spain		Germany		France		Italy	
Model	AIC	Model	AIC	Model	AIC	Model	AIC
ARIMA(2,2,2)	Inf	ARIMA(2,2,2)	988.5937	ARIMA(2,2,2)	931.1487	ARIMA(2,1,2) with drift	Inf
ARIMA(0,2,0)	1141.338	ARIMA(0,2,0)	1044.264	ARIMA(0,2,0)	1048.41	ARIMA(0,1,0)with drift	1007.747
ARIMA(1,2,0)	1062.727	ARIMA(1,2,0)	1023.181	ARIMA(1,2,0)	1015.389	ARIMA(1,1,0) with drift	1009.532
ARIMA(0,2,1)	Inf	ARIMA(0,2,1)	993.1037	ARIMA(0,2,1)	985.305	ARIMA(0,1,1) with drift	1009.242
ARIMA(2,2,0)	1051.972	ARIMA(1,2,2)	Inf	ARIMA(1,2,2)	936.8493	ARIMA(0,1,0)	1007.44
ARIMA(3,2,0)	1044.932	ARIMA(2,2,1)	986.5106	ARIMA(2,2,1)	945.2363	ARIMA(1,1,1) with drift	1009.462
ARIMA(4,2,0)	1046.47	ARIMA(1,2,1)	990.2094	ARIMA(3,2,2)	933.5196	*Best model*	*ARIMA(0,1,0)*
ARIMA(3,2,1)	1037.447	ARIMA(2,2,0)	997.9065	ARIMA(2,2,3)	Inf		
ARIMA(2,2,1)	1037.541	ARIMA(3,2,1)	988.5688	ARIMA(1,2,1)	966.6439		
ARIMA(4,2,1)	1039.886	ARIMA(3,2,0)	994.1668	ARIMA(1,2,3)	932.3201		
ARIMA(3,2,2)	Inf	ARIMA(3,2,2)	990.9985	ARIMA(3,2,1)	943.4252		
ARIMA(4,2,2)	Inf	*Best model*	*ARIMA(2,2,1)*	ARIMA(3,2,3)	Inf		
Best model	*ARIMA(3,2,1)*			*Best model*	*ARIMA(2,2,2)*		

OPEN ACCESS

Table 2. 10-days forecasts of confirmed cases of COVID-19 in Italy, Spain, Germany and France.

Date	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Italy					
3/29/2020	5974	5265.75	6682.25	4890.82	7057.18
3/30/2020	5974	4972.38	6975.62	4442.16	7505.84
3/31/2020	5999	4747.27	7200.73	4097.88	7850.12
4/1/2020	6034	4557.5	7390.5	3807.65	8140.35
4/2/2020	6079	4390.3	7557.7	3551.95	8396.06
4/3/2020	6094	4239.15	7708.85	3320.77	8627.23
4/4/2020	6164	4100.15	7847.86	3108.19	8839.81
4/5/2020	6265	3970.77	7977.24	2910.32	9037.68
4/6/2020	6352	3849.25	8098.75	2724.47	9223.53
4/7/2020	6464	3734.32	8213.69	2548.7	9399.3
Spain					
3/29/2020	9706.62	8760.08	10653.2	8259.02	11154.2
3/30/2020	9433.22	8425.01	10441.4	7891.3	10975.1
3/31/2020	10269.1	9050.61	11487.6	8405.58	12132.6
4/1/2020	10276.3	8952.38	11600.2	8251.53	12301.1
4/2/2020	11336.4	9730.46	12942.3	8880.34	13792.4
4/3/2020	11478.4	9715.09	13241.7	8781.65	14175.1
4/4/2020	12254.7	10239.7	14269.6	9173.1	15336.3
4/5/2020	12512.6	10304.6	14720.5	9135.79	15889.4
4/6/2020	13255.4	10780.7	15730	9470.77	17039.9
4/7/2020	13587.3	10888.5	16286.2	9459.78	17714.9
Germany					
3/29/2020	7685.61	7037.77	8333.45	6694.82	8676.4
3/30/2020	8249.06	7448.25	9049.86	7024.33	9473.78
3/31/2020	8623.36	7703.93	9542.8	7217.21	10029.5
4/1/2020	9183.31	8064.88	10301.7	7472.82	10893.8
4/2/2020	9722.18	8415.17	11029.2	7723.29	11721.1
4/3/2020	10208.5	8719.38	11697.7	7931.07	12486
4/4/2020	10725.8	9035.03	12416.6	8139.99	13311.6
4/5/2020	11246.6	9347	13146.2	8341.42	14151.8
4/6/2020	11755.4	9642.88	13868	8524.56	14986.3
4/7/2020	12268.5	9934.15	14602.9	8698.4	15838.7
France					
3/29/2020	4873.91	4468.09	5279.73	4253.26	5494.56
3/30/2020	5306.67	4900.61	5712.73	4685.66	5927.69
3/31/2020	5873.52	5443.21	6303.83	5215.41	6531.63
4/1/2020	6270.24	5727.57	6812.9	5440.3	7100.17
4/2/2020	6700.45	6090.97	7309.94	5768.33	7632.58
4/3/2020	7180.28	6485.41	7875.16	6117.56	8243
4/4/2020	7621.95	6805.24	8438.65	6372.91	8870.99
4/5/2020	8063.49	7125.19	9001.79	6628.48	9498.5
4/6/2020	8520.02	7451.54	9588.49	6885.93	10154.1
4/7/2020	8969.03	7754.96	10183.1	7112.28	10825.8

Parameter estimation and model selection

For parameter estimations, the "auto.arima" function was used in R package "forecast".[32,33] The purpose of using this package is to fit best the ARIMA model to univariate time series and returns best ARIMA model according to either Akaike Information Criterion (AIC), or its small-sample equivalent (AICc) or Bayesian Information Criterion (BIC) value.[34,35] The function conducts a search over possible model[36] within the order constraints provided.[1] In Table 1, the details of the model with corresponding AIC values are documented. On the basis of AIC, the best model of Italy, Germany, France and Spain are highlighted.

Results

After model selection, the best fit models are used to forecast the growth of COVID-19 confirmed cases in all four countries. Based on confirmed COVID-19 cases, predictions are made for the next 10 days for the top four European countries, namely Italy, Spain, Germany, and France. Table 2 details the forecasts of next ten days for the four countries under consideration based on 80% and 95% confidence interval (CI). The minimum and maximum

[1] The detailed documentation is available at: https://www.rdocumentation.org/packages/forecast/versions/8.11/topics/auto.arima

values for both the confidence intervals are also presented in the table. For instance, it is predicted that in Spain there would be a trend of increasing additional number of cases in the coming 10 days, with an average addition of 11,410 cases. In the case of Spain, the 95% confidence interval shows an increase of number of cases would be between a minimum of 8770 to a maximum of 10975. Likewise, in Italy there would be additional 6190 cases on average ranging from 3540 (lower bound) to 8407 (upper bound) in the next ten days, by the end of first week of April 2020. Similar increasing trend can be observed in case of Germany and France from 29th March 2020 to 4th April 2020. In the case of Germany, an average increase of 9966 confirmed cases would be experienced in the next ten days, ranging from a minimum of 7776 to a maximum of 12,156 per day, statistically significant at 0.05 level. As compared to Germany, the addition in France is little low. We can predict an average addition of 6937 cases per day, with a minimum of 5848 to a maximum of 8027 cases per day, statistically significant at 0.05 level. The forecast of additional number of cases is presented in Figure 1. The blue line shows the forecast value, dark gray shows the 95% confidence interval, while the light grey area shows the 80% lower and upper bounds. The ACF and PACF plots in Figures 2 and 3 shows no significant autocorrelations indicating that the residuals are behaving like white noise. To test the overall randomness based on a number of lags, a portmanteau test is applied to the residuals of all fitted ARIMA models. The significant p-values of Box-Pierce test also suggesting that the residuals are white noise.
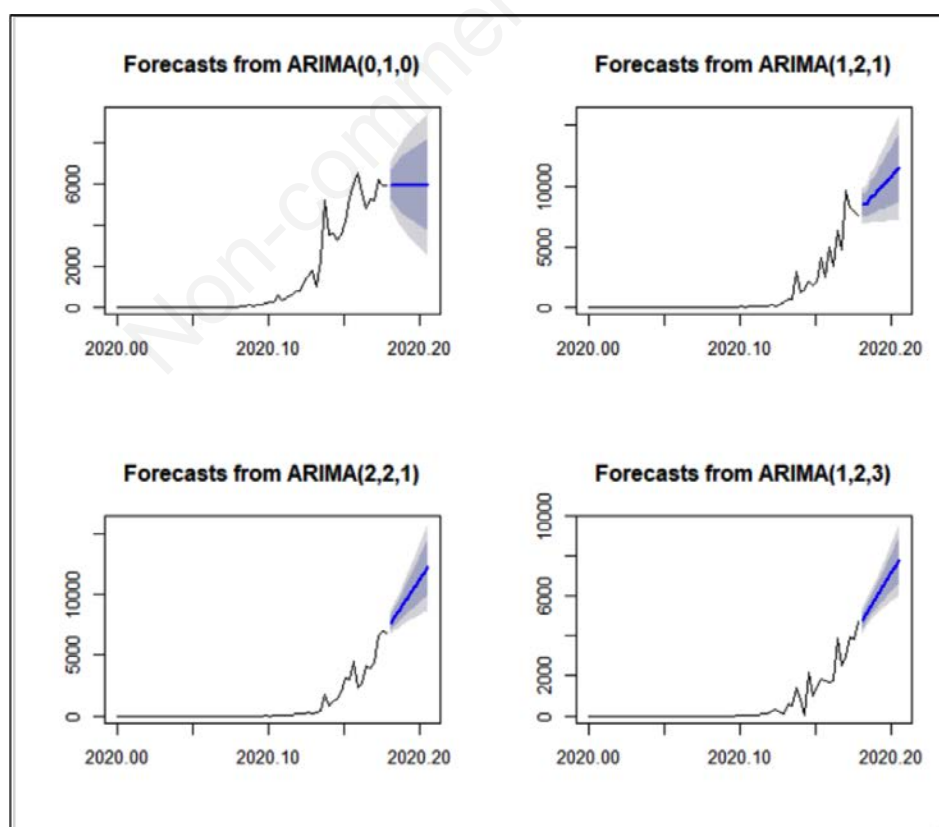

Figure 1. 10-days daily forecast of confirmed cases for Italy (top left), Spain (top right), Germany (bottom left) and France (bottom right).
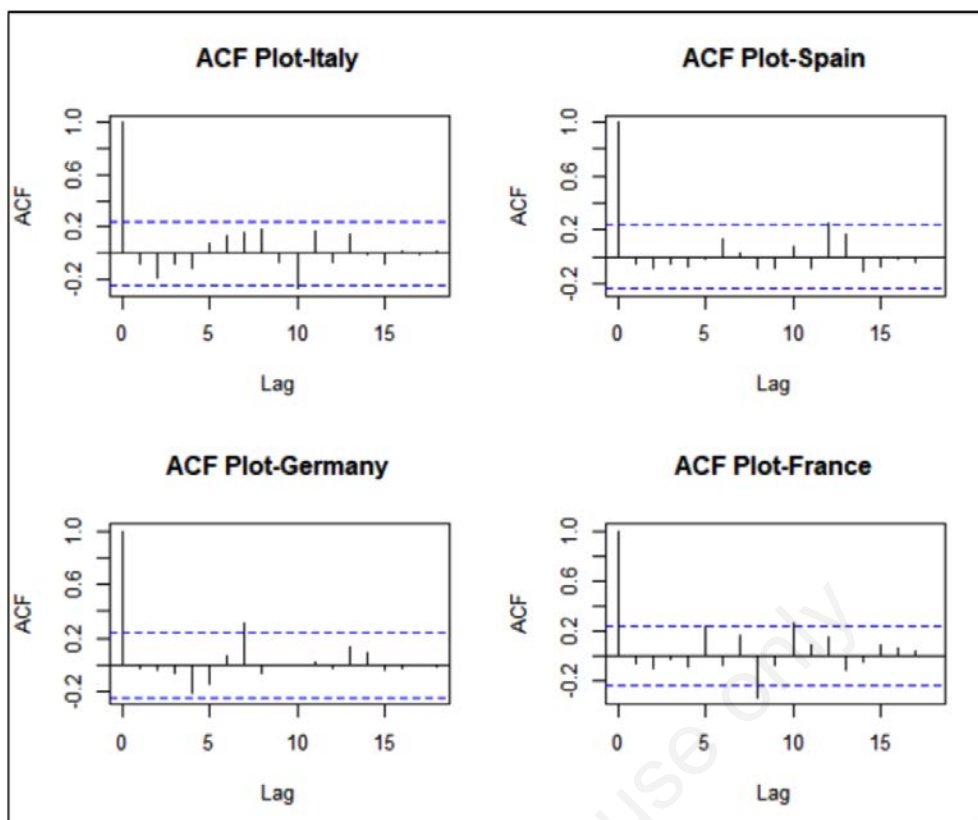
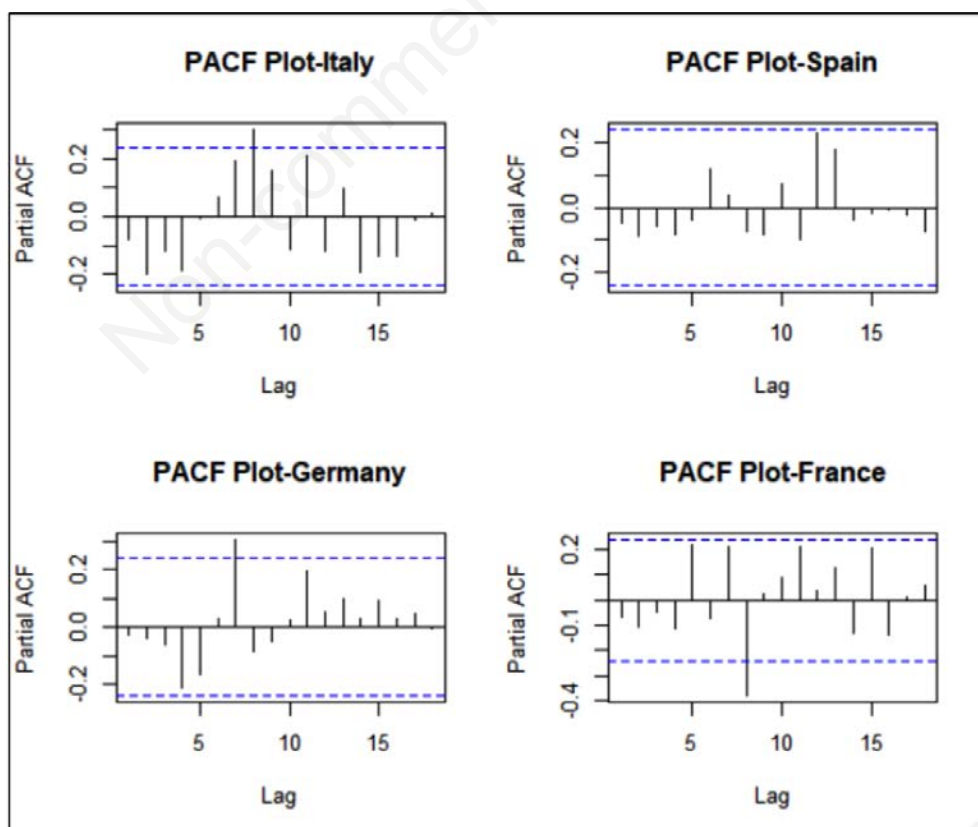Figure 2. ACF Plots of Italy (top left), Spain (top right), Germany (bottom left) and France (bottom right).

Figure 3. PACF Plots of Italy (top left), Spain (top right), Germany (bottom left) and France (bottom right).

Discussion

The purpose of this study was to predict the upcoming confirmed cases of COVID-19 in the top 4 countries (where till date the confirmed cases are highest in number). These countries are Italy, Spain, Germany and France. It is a needed study as through the estimates for next ten days, governments can have an idea whether the cases will be increasing or decreasing. Also, they can make their strategies accordingly and medical facilities can be managed accordingly. The ten days prediction of these four countries showed that there is an increasing trend and there will be more destruction in these countries in the coming days. The confirmed cases for the four countries show an increasing trend for the next ten days with Spain having an average of 11,410 additional cases in next ten days, Italy on average will have 6190 additional confirmed cases, Germany will probably have 9966 new cases and in France 6937 new cases will possibly emerge in next ten days of this deadly pandemic. Hospitals need to prepare more isolation wards and medical supplies are to be ensured for the upcoming cases. Furthermore, more investments in health are needed and primary prevention is needed for this pandemic burden.

Correspondence: Dr. Tahir Mumtaz Awan, Assistant Professor, Department of Management Sciences, COMSATS University Islamabad, Park Road, 45550 Islamabad, Pakistan.
Tel. +92.317.5777-577.
E-mail: tahir_mumtaz@comsats.edu.pk

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020.
2. The Strait Times [Internet]. How early signs of the coronavirus were spotted, spread and throttled in China. 26 February 2020. Available from: https://www.straitstimes.com/asia/east-asia/how-early-signs-of-the-coronavirus-were-spotted-spread-and-throttled-in-china
3. BBC News [Internet]. Coronavirus disease named Covid-19. Channel News Asia, China. 11 February 2020. Available from: https://www.bbc.com/news/world-asia-china-51466362
4. BBC News [Internet].WHO names novel coronavirus as 'COVID-19'. Channel News Asia. 11 February 2020. Available from: https://www.bbc.com/news/world-asia-china-51466362
5. NPR [Internet]. Coronavirus: COVID-19 is now officially a pandemic, WHO says. 11 March 2020. Available from: https://www.npr.org/sections/goatsandsoda/2020/03/11/814474930/coronavirus-covid-19-is-now-officially-a-pandemic-who-says.
6. BBC News [Internet]. Coronavirus confirmed as pandemic by World Health Organization. 11 March 2020. Available from: https://www.bbc.com/news/world-51839944
7. Business Insider [Internet]. A comprehensive timeline of the new coronavirus pandemic, from China's first COVID-19 case to the present. 11 March 2020. Available from: https://www.businessinsider.com/coronavirus-pandemic-time-line-history-major-events-2020-3?IR=T
8. BBC News [Internet]. Coronavirus: Europe now epicentre of the pandemic, says WHO. 13 March 2020. Available from: https://www.bbc.com/news/world-europe-51876784
9. WHO. Infectious disease outbreaks reported in the Eastern Mediterranean Region in 2018. 2018. Available from: https://www.emro.who.int/pandemic-epidemic-diseases/news/infectious-disease-outbreaks-reported-in-the-eastern-mediterranean-region-in-2018.html
10. WHO. Middle East respiratory syndrome coronavirus (MERS-CoV) - Republic of Korea. 2015. Available from: https://www.who.int/csr/don/25-october-2015-mers-korea/en/
11. WHO. Case−control study to assess potential risk factors related to human illness caused by Middle East Respiratory Syndrome Coronavirus (MERS−CoV). Protocol Update. 2014. Available from: https://www.who.int/csr/disease/coronavirus_infections/MERSCoVCaseControlStudyofExposures28March2014.pdf?ua=1
12. Smith RD. Responding to global infectious disease outbreaks: lessons from SARS on the role of risk perception, communication and management. Social Sci Med 2006;63:3113-23.
13. European Centre for Disease Prevention and Control. Covid-19 situation update for the EU/EEA and the UK. Available from: https://www.ecdc.europa.eu/en/cases-2019-ncov-eueea
14. Wu N, Green B, Ben X, O'Banion S. Deep transformer models for time series forecasting: The influenza prevalence case. arXiv: 2001.08317.
15. Contreras J, Espinola R, Nogales FJ, Conejo AJ. ARIMA models to predict next-day electricity prices. IEEE Trans Power Syst 2003;18:1014-20.
16. Ediger VŞ, Akar S. ARIMA forecasting of primary energy demand by fuel in Turkey. Energy Policy 2007;35:1701-8.
17. Ariyo AA, Adewumi AO, Ayo CK. Stock price prediction using the ARIMA model. Proceedings UK Sim-AMSS 16th Int Conf on Computer Modelling and Simulation, Cambridge, 2014. IEEE Xplore. doi: 10.1109/UKSim.2014.67.
18. Faruk DÖ. A hybrid neural network and ARIMA model for water quality time series prediction. Engin Appl Artific Intell 2010;23:586-94.
19. Van Der Voort M, Dougherty M, Watson S. Combining Kohonen maps with ARIMA time series models to forecast traffic flow. Transp Rese Part C: Emerg Technol 1996;4:307-18.
20. Cong J, Ren M, Xie S, Wang P. Predicting Seasonal influenza based on SARIMA model, in mainland China from 2005 to 2018. Int J Environ Res Public Health 2019;16:4760.
21. Su K, Xu L, Li G, et al. Forecasting influenza activity using self-adaptive AI model and multi-source data in Chongqing, China. EBioMed 2019;47:284-92.
22. Tapak L, Hamidi O, Fathian M, Karami M. Comparative evaluation of time series models for predicting influenza outbreaks: application of influenza-like illness data from sentinel sites of healthcare centers in Iran. BMC Res Notes

2019;12:353.
23. Stoeldraijer L. Mortality forecasting in the context of non-linear past mortality trends: an evaluation. PhD Thesis, University of Groningen; 2019.
24. Kim TH, Hong KJ, Do Shin S, et al. Forecasting respiratory infectious outbreaks using ED-based syndromic surveillance for febrile ED visits in a metropolitan city. Am J Emerg Med 2019;37:183-8.
25. Wang L, Chen J, Marathe M. DEFSI: Deep learning based epidemic forecasting with synthetic information. Proceedings AAAI Conference on Artificial Intelligence; 2019.
26. Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data Brief 2020;29:105340.
27. Roosa K, Lee Y, Luo R, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect Dis Modell 2020;5:256-63.
28. Box GE, Jenkins GM. Time series analysis: Forecasting and control. San Francisco: Holden-Day; 1976.
29. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003;50:159-75.
30. Brockwell PJ, Davis RA. Introduction to time series and forecasting. Springer Nature; 2016.
31. Hipel KW, McLeod AI, Lennox WC. Advances in Box−Jenkins modeling: 1. Model construction. Water Resour Res 1977;13:567-75.
32. Hyndman RJ, Athanasopoulos G, Bergmeir C, et al. Package 'forecast'. Available from: https://cran.r-project.org/web/packages/forecast/forecast.pdf
33. Wang X, Smith K, Hyndman R. Characteristic-based clustering for time series data. Data Min Knowl Disc 2006;13:335-64.
34. Jain G, Mallick B. A study of time series models ARIMA and ETS. Available at SSRN 2898968. 2017.
35. Dhamo E, Puka L, editors. Using the R-package to forecast time series: ARIMA models and Application. INTERNATIONAL CONFERENCE Economic & Social Challenges and Problems 2010 Facing Impact of Global Crisis; 2010.
36. Pravilović S, Appice A. The intelligent forecasting model of time series. Automation, Control and Intelligent Systems. 2014;1:90.