Short Communication
Advancements in Life Sciences – International Quarterly Journal of Biological Sciences

Open Access

Wavelet-based Statistical and Mathematical Analysis of Spread of COVID-19

Samreen Fatima*, Mehwish Shafi Khan, Yumna Sajid

Abstract

The outbreak of coronavirus-19 (NCoV-19) has developed a universal crisis due to high rate of infection and mortality. Therefore, the researchers are using various available methods to study the pattern of spread of COVID–19 which will help in planning to control the disease and to manage the health care resources. This study compares Autoregressive Integrated Moving Average (ARIMA) (statistical), Logistic, Gompertz (mathematical) and their hybrid using Wavelet–based Forecast (WBF) models to model and predict the number of confirmed cases of COVID–19. The study area includes the countries: Iran, Italy, Pakistan, Saudi Arabia, USA, UK and Canada. Moreover, root mean squares error (RMSE) is used to compare the performance of studied models. Empirical analysis shows that confirmed cases could be adequately modelled using ARIMA and ARIMA-WBF for all the countries under consideration. However, for future prediction significance of the models varies region to region.

ARTICLE INFO

Date Received:
07/05/2021;
Date Revised:
26/09/2021;
Date Published Online:
31/12/2021;

Authors' Affiliation:
Department of Statistics, University of Karachi - Pakistan

***Corresponding Author:**
Samreen Fatima
Email:
sf_quddus@hotmail.com

How to Cite:
Fatima S, Khan MS, Sajid Y (2021). Wavelet-based Statistical and Mathematical Analysis of Spread of COVID-19. Adv. Life Sci. 8(4): 333-338.

Keywords:
COVID–19; ARIMA; Logistic; Gompertz; Wavelet–based–forecast

Introduction

Rresearchers from all over the world from various fields are engaged in studying the pattern of the COVID–19 outbreak. Concerning the modeling and forecasting of COVID–19 cases, a large number of studies have been carried out employing the existing mathematical or statistical models during the last few months including other epidemic variables which may cause rapid increase in pandemic [1-6].

Moreover, statistical models including short-term/long-term, box Jenkins Autoregressive Integrated Moving Average (ARIMA) model, time series generalized linear mixed models (GLMM), Generalized logistic model have been applied to study the pattern and growth of COVID–19 [3,7-12]. Classical mathematical models: Susceptible Infectious and Recovered (SIR) and Susceptible, Exposed, Infectious and Recovered (SEIR) models [12-14] along with its modifications eSIR have been used to model the current outbreak of COVID–19 [15]. In addition to these, transmission risk and prediction of the peak time of COVID–19 has also been done via mathematical model based on the clinical evolution [16].

As the COVID–19 data was available in limited quantity due to its early stage therefore, advanced deep learning methods might over fit the training set [17]. However, these models have different approach subject to geographical location, dimension of variables and time period. Therefore, it is important to develop country specific model which can model COVID–19 outbreaks precisely.

This study utilizes mathematical and statistical techniques for confirmed COVID–19 cases in two parts. In the first part, statistical model ARIMA and mathematical models: Logistic and Gompertz are employed. Whereas their wavelet-based hybrid models are developed in the second part. The daily confirmed cases data are selected for the countries: Pakistan, Iran, Italy, Saudi Arabia, UK, USA and Canada, which belong to different geographical location and different economic background. Among all considered countries, Canada, Italy, UK and USA are highly developed regions containing good health facilities and belong to highest income group of countries. However, Pakistan and Iran falling in the category of developing countries holding poor health facilities. Furthermore, to assess the performance of the studied models RMSE (root mean squares error) criterion is used.

Methods

Autoregressive Integrated Moving Average

ARIMA (p, d, q) model is the linear combination of Autoregressive (AR(p)) and Moving Average (MA(q)) process. Where, p and q represent the order of AR and MA processes and d is the number of times that the integrated process must be differenced to make a time-series stationary. ARIMA (p, d, q) model is described as follows:

$$x_t = \theta_0 + \sum_{q=1}^{n} \theta_q e_{t-q} + \sum_{p=1}^{m} \varphi_p x_{t-p} + e_t$$

Where, 'x_t' is time series, x_{t-p}'s are their lag values and e_t is error having mean zero and constant variance satisfy the condition of i.i.d. Moreover, φ_i and θ_j are the parameters of the model which are determined either by the least square method or method of maximum likelihood.

Logistic Growth model

Verhulst, a mathematician improved the Malthusian population growth model proposed in 1977 and named it as Logistic function [18]. It is widely used in various fields from biology to computer science. In this study three parameters based logistic model is used to fit the COVID–19 data of the total confirmed cases described as follows:

$$f(t) = \frac{a}{(1+\exp(b-c(t-t_0)))} + \varepsilon_t \ldots\ldots (1)$$

Where, $f(t)$ is the number of confirmed cases at time 't', constant 'c' is the continuous growth rate determining the rate of infection spread, 'a' is the maximum number of confirmed cases b is the fitting coefficients and t_o is the time when the first case occurred.

Gompertz model

The Gompertz function or Gompertz model was originally proposed by Gompertz for modeling of an infection spread [19]. It belongs to the family of sigmoid function which is categorized by a slow growth rate curve at the beginning and at the end of a given time period whereas the function grows sharply in the middle of the period. Mathematically, Gompertz function can be expressed as:

$$f(t) = \exp\,(a - b * \exp\,(c\,(t - t_o))) + \varepsilon_t \ldots \ldots (2)$$

Where, 't' is time and t_o is the time when the first case occurred, 'a' is the predicted maximum of confirmed cases and 'b' and 'c' are slopes.

The parameters of the Logistic and Gompertz models are determined by nonlinear least square method (NLSM) which minimizes the sum of squares of error.

Wavelet Based Forecast

WBF is used to extract information via signal decomposition in the domain of time and frequency [20]. Discrete Wavelet Transform (DWT) [21], with initial resolution of one signal as K can be defined as $x(t) = \sum_{j\epsilon\mathbb{Z}} r_{K,n}\ \Omega_{K,n}(t) + \sum_{i=K}^{\infty} \sum_{j\epsilon\mathbb{Z}} s_{i,n}\omega_{j,n}\ (t)$, where, $\Omega(t)$ is scaling function and $\omega(t)$ is the family of wavelets (see [22] for more detail). In this study, maximal overlap discrete wavelet transform (MODWT) is applied instead of DWT due to its limitations: requirement of sample size of multiple of 2 and output of DWT is highly dependent of origin of the signal that would be affected upon a slight shift in origin [23,24]. The MODWT pyramid algorithm [22] generates wavelet coefficients ($r_{j,n}$) and the scaling coefficients ($s_{j,n}$) from ($s_{j-1,n}$) defined as:

$$r_{j,n} = \sum_{i=0}^{L-1} \tilde{a}_{j,i}\, x_{n-i \bmod N} \quad , s_{j,n} = \sum_{i=0}^{L-1} \tilde{b}_{j,i}\, x_{n-i \bmod N}$$

Above expression is for periodic filter operations of the actual series (x_n) with filters: $\tilde{a}_{j,i} = a_{j,i}/2^{j/2}$ and $\tilde{b}_{j,i} = b_{j,i}/2^{j/2}$. Where, N is length of time series and n varies from 0 to N–1, a_j and b_j are coefficients of scaling and wavelet filters, respectively. Moreover, the original signals can be recovered by applying inverse pyramid algorithm [22]:

$$s_{j-1,n} = \sum_{i=0}^{L-1} \tilde{a}_i\, r_{j,n+2^j\, i \bmod N} \quad + \sum_{i=0}^{L-1} \tilde{b}_i\, s_{j,n+2^j\, i \bmod N}$$

Wavelet Based Hybrid Models

Here ARIMA-WBF is explained, and a similar approach is applied on all hybrid models (Logistic–WBF and Gompertz–WBF) as discussed here for ARIMA-WBF. A hybridization of ARIMA and WBF model has been opted to minimize the biases of the individual models [25]. In contrast to ARIMA, WBF are more likely to deal with non-stationary time series data. Furthermore, non-stationarity of pandemic data motivates the selection of WBF for modelling the data [26,27]. In this study, ARIMA–WBF has been applied steps-wise as explained by Gosh [25]. Firstly, ARIMA model is applied to model the time series of length n and m numbers of forecast are done on the basis of model obtained. Secondly, the non-stationary residual series of length n generated during previous step is further passed to WBF model. Moreover, WBF acquire in-sample residual predictions and provide out-sample forecast as well using decomposition levels as W_L = log (n) and boundary as 'periodic'. Finally, forecasts obtained from ARIMA and WBF are combined for both in-sample and out-sample data to achieve hybridized prediction.

Data Analysis

Univariate daily time-series data of COVID–19 for the time period 1st January 2020–4th May 2021 are collected from https://ourworldindata.org/coronavirus-source-data. The total number of confirmed COVID–19 cases per million and total number of deaths per million populations (now onwards referred as confirmed cases and death cases respectively) of seven countries from different geographical locations including: Canada, Iran, Italy, Pakistan, Saudi Arabia, USA and UK are considered in this study.

From basic statistics (graph\standard deviation), it is noted that increase in number of confirmed cases in a particular month leads to increase in mortality rate in the following month. Therefore, to study the relationship between the numbers of confirmed cases and death cases correlation and descriptive statistics are calculated for the countries under consideration (Table1). However, only confirmed cases are analyzed and modeled for future prediction via statistical and mathematical models. Moreover, the data is divided into model building from 1st January to 24th February, 2021 whereas validation period consists of 25th February, 2021 to 4th May 2021.

Among all the countries under analysis, average number of confirmed cases are high in US followed by UK, Italy, Canada, Iran, Saudi Arabia and Pakistan over the considered time period. Moreover, the average number of death cases are high in UK followed by Italy, US, Iran, Canada, Saudi Arabia and Pakistan. The Standard Deviation (Std. Dev) of confirmed cases is high in US followed by UK, Italy, Canada, Iran, Saudi Arabia and Pakistan. The Std. Dev of death cases is high in UK followed by Italy, US, Iran, Canada, Saudi Arabia and Pakistan. Overall, Pakistan and Saudi Arabia are less affected by COVID–19 among all considered countries on the basis of confirmed and death cases. However, the correlations between the confirmed cases and death cases are high in all the countries, (Table 2). To fit the logistic and Gompertz functions on confirmed cases, Mathematica routine using the minimum least squares method is compiled.

Results

For the statistical model ARIMA, ADF test is used to check the stationarity of the data. Moreover, *auto.arima*() is applied for acquiring suitable ARIMA model based on minimum AIC and BIC criterion including the condition of significance of parameter at 5% level under R environment. In the next step hybrid models: ARIMA–WBF, Logistic–WBF(GL–WBF) and Gompertz–WBF(GUP–WBF) are applied on the confirmed cases to enhance the prediction performance. Based on RMSE, ARIMA is found at the top whereas ARIMA–WBF is on the second number for in sample forecast for all the countries. However, in case of out sample both ARIMA and ARIMA-WBF found suitable for Pakistan, Saudi-Arabia, Italy and US. Whereas for the Iran and Canada GL-WBF and for US has logistic curve give minimum out sample forecast.

Moreover, Fig. (1 to 7) show the plots of actual and fitted data along with the forecasted one. Furthermore, during model building period it is observed that actual values are overlapped by most of the modelled fitted values except for ARIMA for Saudi Arabia and Logistic and Gompertz–WBF in case of Iran, Italy, UK and USA, respectively. However, during out-sample forecasting each model either underestimate or overestimate the prediction which has already been discussed via RMSE.

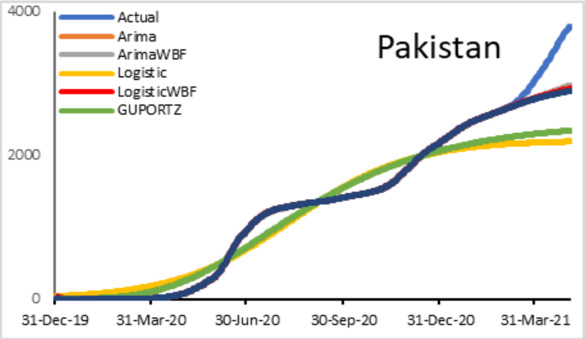

Figure 1: Actual and predicted data of confirmed cases of Pakistan

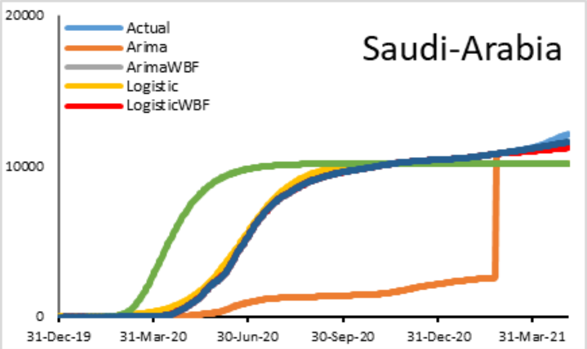

Figure 2: Actual and predicted data of the confirmed cases of Saudi Arabia

In the above fig 1, it is observed that Gompertz and Logistic models poorly estimate the in-sample and out-sample forecast. Fig 2 shows that in-sample and out-sample forecast is fairly mapped by Gompertz-WBF, Logistic and Logistic-WBF. Whereas ARIMA and Gompertz poorly captured the in-sample as well as out-sample forecast. Fig 3 shows that the in-sample forecast fairly captured by almost all models except Gompertz - WBF. Whereas all the models over-predict in out-sample forecast. Fig 4 shows that the in-sample forecast fairly captured by almost all models except Gompertz. Whereas out-sample forecast over-predict by Gompertz-WBF, Logistic and Gompertz.

Cases	Death Cases			Confirmed Cases		
Country	Mean	Median	Standard Deviation	Mean	Median	Standard Deviation
Pakistan	33.39314	29.743	21.96006	1320.541	1340.694	1091.516
Saudi Arabia	118.6821	146.12	70.47095	6545.627	9070.294	4563.837
Italy	862.5729	592.688	552.9806	17904.55	4452.628	21282.25
Canada	318.2692	254.5695	184.9327	8549.225	3416.551	9650.527
Iran	373.0677	302.335	278.6882	7977.795	4467.184	8292.351
UK	891.8178	625.166	573.0443	20082.08	4947.603	24204.41
US	763.5508	628.635	541.3997	32051.99	18220.44	34065.37

Table 1: Descriptive statistics of total deaths and confirmed COVID-19 cases.

Pakistan	Canada	Saudi Arabia	Italy	Iran	UK	US
0.99882	0.95757	0.965464	0.973682	0.962133	0.970445	0.987663

Table 2: Correlation between confirmed and death cases

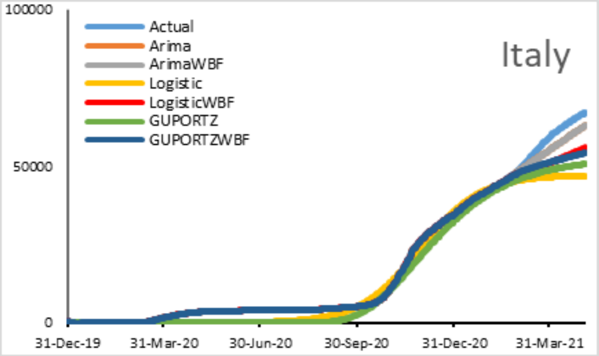

Figure 3: Actual and predicted data of confirmed cases of Italy

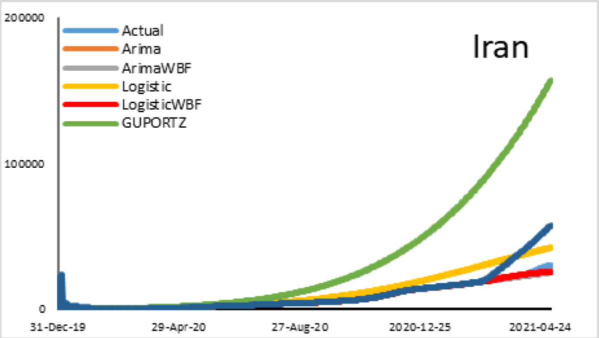

Figure 4: Actual and predicted data of the confirmed cases of Iran.

Figure 5 below shows that in-sample forecast poorly captured by Logistic and Gompertz. Whereas out-sample forecast is poorly captured by all models except Logistic-WBF. Fig 6 shows that the in-sample forecast captured by all models are good fitted and out-sample forecast are over predicted by Gompertz-WBF and Logistic model. Graph of Canada shows that the in-sample forecast decently captured by ARIMA, Arima-WBF, Logistic-WBF and Gompertz-WBF. Whereas out-sample forecast captured by Arima-WBF, Logistic and Gompertz are over predicted.

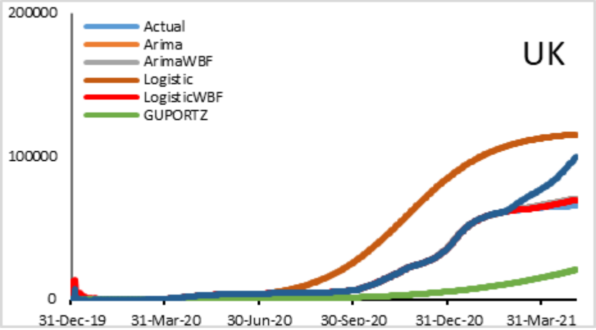

Figure 5: Actual and predicted data of confirmed cases of UK.

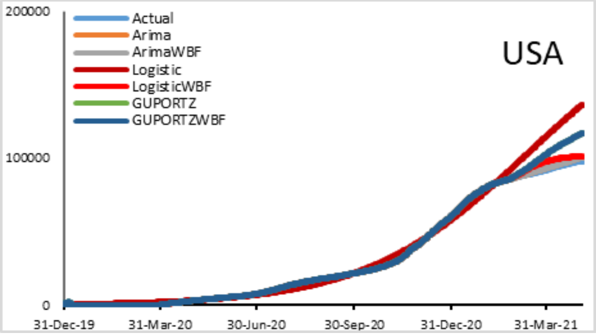

Figure 6: Actual and predicted data of the confirmed cases of USA.

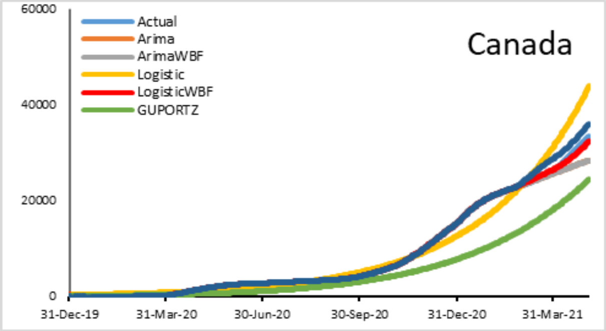

Figure 7: Actual and predicted data of confirmed cases of Canada

Discussion

In this paper, statistical and mathematical models are applied with their hybrid using WBF including: ARIMA, Logistic, Gompertz, ARIMA–WBF, Logistic–WBF and Gompertz–WBF on COVID 19 confirmed cases. COVID–19 data is considered from 31st December, 2019 to 4th May, 2021 for Pakistan, Saudi Arabia, Iran, Italy, United Kingdom, Canada and United States of America. Moreover, above models being data driven behave in different manner in different regions. Therefore, no specific model could be ranked as best or worst however for a particular region and period the model could be chosen smartly on the basis of statistical and graphical analysis. From both visual and statistical analysis, it is found that ARIMA was able to capture in-sample forecast adequately in all countries. In contrast, Wavelet based hybrids of Logistic and Gompertz failed to capture in-sample pattern for all the countries except for Canada. Whereas, for out-sample forecast ARIMA–WBF are adequate except Iran, UK and Canada. However, LG-WBF is found appropriate for Iran and Canada while Logistic is suitable for UK in out sample forecast. Over all, in most of the cases ARIMA and hybrid ARIMA–WBF models performed well. As the COVID–19 data is assumed as an exponential growth curve nevertheless this curve consists of piecewise linear pattern therefore, ARIMA model and hybrid ARIMA–WBF could capture the pattern effectively.

Conflict of Interest Statement

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Contributions

All authors contributed equally to this study.

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine, (2020).
2. Silverstein WK, Stroud L, Cleghorn GE, Leis JA. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020; 395: 689–97—In this Article. Lancet, (2020); 395689-697.
3. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief, (2020); 105340.
4. Dehesh T, Mardani-Fard HA, Dehesh P. Forecasting of covid-19 confirmed cases in different countries with arima models. medRxiv, (2020).
5. Zeb A, Alzahrani E, Erturk VS, Zaman G. Mathematical Model for Coronavirus Disease 2019 (COVID-19) Containing Isolation Class. BioMed research international, (2020); 2020.
6. Martinez EZ, Aragon DC, Nunes AA. Short-term forecasting of daily COVID-19 cases in Brazil by using the Holt's model. Revista da Sociedade Brasileira de Medicina Tropical, (2020); 53.
7. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The lancet infectious diseases, (2020).
8. Zhuang Z, Zhao S, Lin Q, Cao P, Lou Y, et al. Preliminary estimation of the novel coronavirus disease (COVID-19) cases in Iran: A modelling analysis based on overseas cases and air travel data. International Journal of Infectious Diseases, (2020); 9429-31.
9. Tandon H, Ranjan P, Chakraborty T, Suhag V. Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. arXiv preprint arXiv:200407859, (2020).
10. Duan X, Zhang X. ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data. Data in Brief, (2020); 105779.
11. Saez M, Tobias A, Varga D, Barceló MA. Effectiveness of the measures to flatten the epidemic curve of COVID-19. The case of Spain. Science of the Total Environment, (2020); 138761.
12. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, (2020); 5256-263.

13. Anderson RM, Anderson B, May RM Infectious diseases of humans: dynamics and control. Chapter: Book Name. 1992 of publication; Oxford university press.
14. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PloS one, (2020); 15(3): e0231236.
15. Wangping J, Ke H, Yang S, Wenzhe C, Shengshu W, et al. Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China. Frontiers in medicine, (2020); 7169.
16. Tang B, Wang X, Li Q, Bragazzi NL, Tang S, et al. Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. Journal of clinical medicine, (2020); 9(2): 462.
17. Hastie T, Tibshirani R, Friedman J The elements of statistical learning: data mining, inference, and prediction. Chapter: Book Name. 2009 of publication; Springer Science & Business Media.
18. Verhulst P-F (1977) A note on the law of population growth. Mathematical Demography: Springer. pp. 333-339.
19. Gompertz B. XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c. Philosophical transactions of the Royal Society of London, (1825); (115): 513-583.
20. Percival DB, Walden AT Wavelet methods for time series analysis. Chapter: Book Name. 2000 of publication; 4; Cambridge university press.
21. Mallat S A wavelet tour of signal processing. Chapter: Book Name. 1999 of publication; Elsevier.
22. Alarcon-Aquino V, Barria J. Change detection in time series using the maximal overlap discrete wavelet transform. Latin American applied research, (2009); 39(2): 145-152.
23. Jothimani D, Shankar R, Yadav SS. Discrete wavelet transform-based prediction of stock index: A study on national stock exchange fifty index. arXiv preprint arXiv:160507278, (2016).
24. Al Wadi S, Ababneh F, Alwadi H, Ismail MT. Maximum overlap discrete wavelet methods in modeling banking data. Far East Journal of Applied Mathematics, (2013); 84(1): 1.
25. Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos, Solitons & Fractals, (2020); 109850.
26. Aminghafari M, Poggi J-M. Forecasting time series using wavelets. International Journal of Wavelets, Multiresolution and Information Processing, (2007); 5(05): 709-724.
27. Benaouda D, Murtagh F, Starck J-L, Renaud O. Wavelet-based nonlinear multiscale decomposition model for electricity load forecasting. Neurocomputing, (2006); 70(1-3): 139-154.