Trajectory of COVID-19 Data in India: Investigation and Project Using Artificial Neural Network, Fuzzy Time Series and ARIMA models

Abstract

Due to the strength of Coronavirus (COVID-19) pandemic that exists today, all countries, national and international organizations are in a continuous effort to fall efficient and accurate statistical models for forecasting the future pattern of COVID infection. Accurate forecasting should help governments to take optimal decisions to master the pandemic spread. In this article, we explore the COVID-19 data base of India over the period (17-march to 1 July 2020), then we estimated two nonlinear time series models: Artificial Neural Network (ANN) and Fuzzy Time Series (FTS) by comparing them with ARIMA model. In terms of model adequacy, the FTS model outperforms the ANN for the new cases and new deaths time series in India. We observed a short-term virus spread trend according to three forecasting models.. Such findings help the more efficient preparation for the health system in India.

Key Words: Artificial Neural Network, ARIMA, Covid-19 Forecasting, Fuzzy time series, India.

Introduction

The COVID-19 pandemic is the overall global health of progressing time and the most extraordinary overall supportive go facing the world has looked since World War II. [1 and 2]. The COVID-19 cases initially observed in the Wuhan province of China are now immensely increasing around the world as a consequence of which the Government of India has call upon the powers under the Epidemic Diseases Act, 1897 to raise preparedness and containment of the virus and declared COVID-19 a 'notified disaster' under the Disaster Management Act 2005. On 24 March, the Government of India imposed nationwide lockdown for 21 days, which was further extended on 14 April and 01 May until 17 May, for preventive measure against the 2020 corona virus pandemic in India.

In the third phase of lockdown the Government divided the whole country into three different zones – Green, Red and Orange, with relaxations applicable accordingly. Government of India declared fourth lockdown from 18 May to 31 May. But after all the lockdowns cases are increasing day by day. Globally, 14,043,176 confirmed cases were recorded till July 19, 2020

along with 597583 fatalities owing to the infection [3].In India, after June 8, a phase of reopening of economy started with Unlock 1 to revive economic growth which increased the caseload as the spread of infection peaked due to unlock and manifested into 11,88223 confirmed cases till July 20, 2020. But India is still at lower trajectory in case of deaths owing to COVID 19 compared to any other country in the world due to recovery rate of 62.8 percent which points towards the fact that despite high population density and vulnerability towards community transmission, India has contained the transmission of virus to a great extent. Presently, total active cases in the country are 412404 with 28,712 deaths. As on date In terms of active cases, we have nearly same pattern over the provinces; Maharashtra, with 327031 recorded cases and recovery rate of 54 percent is highest infected state of the country followed by Delhi, having 125,096 total cases with recovery rate of 71 percent at 20 July 2020.

Considering the onset of infections in the country the Indian government quickly activated its health management system and issued necessary travel advisories which included screening of all travelers coming into the country from COVID affected nations as early as January 2020 and setup of institutional quarantine and isolation centers using government infrastructure which ranged from schools, community centers, hotels to rail couches. Simultaneously, with increasing infections, Indian administration was forced to hastily scale up its critical care infrastructure. According to health ministry of India, to fight against COVID-19 32000 ventilators have been installed before March 30, 2020.India like huge populating counties, the government had increased the spending to 2.5 percent of GDP as compared to 1.4 percent as public expenditure on healthcare.

Meanwhile, several non-medical equipment companies in the country have also risen to the occasion and transformed their manufacturing to make ventilators and other required equipments. According to the Association of Indian Medical Device Industry states had nominated 958 COVID hospitals across the country, as well as 2,313 COVID health centers for those who do not need too much medical support and 7,525 COVID care centers for patients with mild infections who are unable isolate themselves at home. Along with this, domestic manufacturers are providing enough medical devices, protective gear, diagnostics,

hospital equipments, and telemedicine services to efficiently overcome the adversities of COVID pandemic in the second largest populated nation.

Time series forecasting models have great scope in present era especially in case of epidemic diseases projection. Researcher considered temporal dynamic for projection of Covid-19 on China, Italy and French [4]. Another study is forecasted the Covid -19 outbreak in Canada using the LSTM network[5]. For Covid -19 in India forecasting using the SIR and logistic model [6]. Someone used Exponential smoothing method [7] and [8] ARIMA models for forecasting purpose. Today's scenario machine learning is useful technique for forecasting aim and researchers are using in different fields of science. Machin learning models compared différent forecasting model on Covid-19 related data and found machine learning performance is better than other models[9]. For few data SVP performance was better than machine learning. Also used machine learning techniques for projection of Covid-19 data in India[10]. This all projections help to make planning to fight against this epidemic disease.

Material and Methods

The main goal of building of mathematical models in time series data is forecasting the future pattern and trajectory. The classical models such: Exponential smoothing [11], Autoregressive Integration Moving Average (ARIMA) [12] ,Kalman filter [13] are the main used in application. We focus mainly on two approaches named: Artificial Neural Network models [14; 15) and Fuzzy Time Series models (FTS) [16] and [17] by comparing them with ARIMA technique. Both of ANN and FTS belong to the non linear time series models.

Auto Regressive Integrated Moving Average (ARIMA) :

Given a time series of data X_t, the ARMA model is a tool for understanding and, perhaps, predicting future values in this series[27]. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The model is usually then referred to as the **ARMA** (p,q) model where p is the order of the autoregressive part and q is the order of the moving average part (as defined below).

Autoregressive model

The notation AR (p) refers to the autoregressive model of order p. The AR(p) model is written

$$X_t = c + \sum_{i=1}^{P} \rho_i X_t + \varepsilon_t$$

where $\rho_1, \rho_2 ..... \rho_p$ are the parameters of the model, c is a constant and ε_t is white noise. Sometimes the constant term is avoided.

Moving Average model

The notation MA (q) refers to the moving average series of order q:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Where the $\theta_1, ..., \theta_q$ are the parameters of the model, μ is the expectation of X_t (often assumed to equal 0), and the $\varepsilon_t, \varepsilon_{t-1} \ldots$

A time series $\{X_t\}$ is stationary and if for every t,

$$X_t - \phi_1 X_{t-1} - ..... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}$$

where, $\{Z_t\} \sim WN(0,\sigma^2)$ and the polynomials

$(1 - \phi_1 Z - ..... - \phi_p Z^p)$ and $(1 + \theta_1 Z + ... + \theta_q Z^q)$ have no common factors.

where p and q are respectively the AR and MA terms.

Artificial Neural Network Models (ANN)

For the *Artificial Neural Network models*, [18] stated that these models provide a great deal of promise in forecasting. We found them applied in several fields, electricity prices predication [19], in Hydrology [20] in Biology [21] etc the theoretical background of ANN is depicted in Figure 1; the essential components that determine the ANNs are Architecture structure and learning algorithm. For the first component, we follow the Feed-forward back propagation network (as in Figure 1).

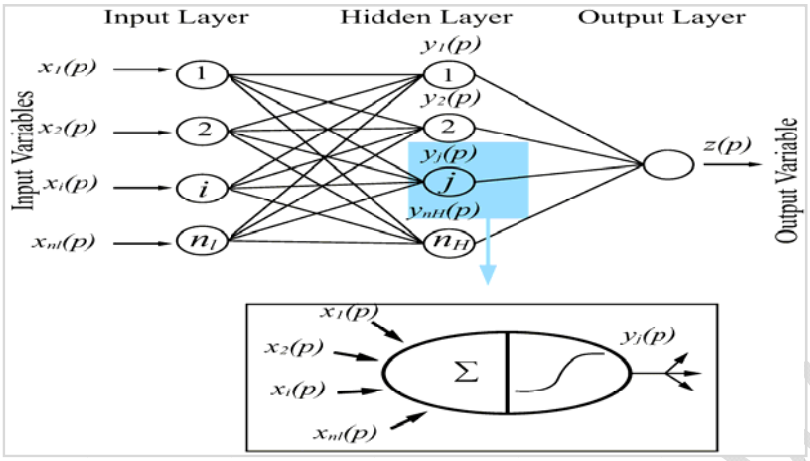

Figure1 Architecture structure of the Artificial Neural Network model

Fuzzy time Series

The FTS models are based on the fuzzy logic and fuzzy sets theory developed by [22], this method is considered as a support of decision making. Furthermore, the FTS doesn't require any prior assumptions for time series and model building, an advantage compared the classical methods (e.g. Box-Jenkins approach). Several statisticians have been contributed to develop this technique, among them: [23; 24'25;17 and 26], to not cite others.

The theoretical steps to construct a fuzzy time series model are:

(1) *Determination the universe of discourse*: which is the range of (or the interval) covering all data, the most used formula is: $\Omega = [Min(y_t) - d_1; Max(y_t) + d_2]$, $d_1$andd_2 are arbitrarily real numbers.

(2) *Definition of Fuzzy Sets:* they are the sub-intervals of the universe of discourse; we work on the equal lengths of each sub-interval; according to [23] there are no prior assumptions on determining how many linguistic variables to be fuzzy sets.

(3) *Fuzzification of the original data set: (convert the raw data (numbers) to linguistic form.*

(4) *Definition of fuzzy logic relationships* (stated the relation among the fuzzy sets), , where we define the fuzzy relation matrix.

(5) Estimated the forecasted output and interpret the results.

In the next section, these two approaches have been applied on the real data set of Covid-19 in India represented by the new number of deaths and the new confirmed cases.

Results & Discussions

The data source is from the World Health Organization (WHO) daily situation reports, (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports). We first, proceed to describe the data, Table 1, shows the summary statistics of the two time series, the number of observations covered the period (17-March to 01-July) is 107 for each. The rang values for the two time series are : [0, 507] and [0, 19906].

Table1 Summary statistics for the new death and new confirmed cases time series.

Variable	Obs(N)	Min	Max	Mean	S.D	C.V (%)	Skwnes	Kurtosis	JB statisic (**)
New_deaths	107	0	507	147	141	95.9	0.796	2.362	13.11
New_cases	107	0	19906	5471	5580	102.2	0.976	2.892	17.05

Notes. J.B: Jark-bera test for normality, (**) indicates statistically significant for p- value <0.001.

The coefficient of variation indicates that the dispersion is nearly the same for the new deaths and new confirmed cases time series over the study period. Rapid and onward changes in number of cases and number of deaths results such high CV percentage in both the parameters under consideration. According to the homogeneity test, we see that the change points for the two time series (New cases, new death) are the observations correspond on 2168 and 12564 new cases, and 28 and 247 new deaths. The positive value of skewness (0.796) which indiçâtes the probability of increasing in the new deaths.

The coefficient of variation indicates that the dispersion is nearly the same for the new deaths and new confirmed cases time series over the study period. Rapid and onward changes in number of cases and number of deaths results such high CV percentage in both the parameters under consideration. According to the homogeneity test, we see that the change points for the two time series (New cases, new death) are the observations correspond on 2168 and 12564 new cases, and 28 and 247 new deaths. The positive value of skewness (0.796) which indiçâtes the probability of increasing in the new deaths.

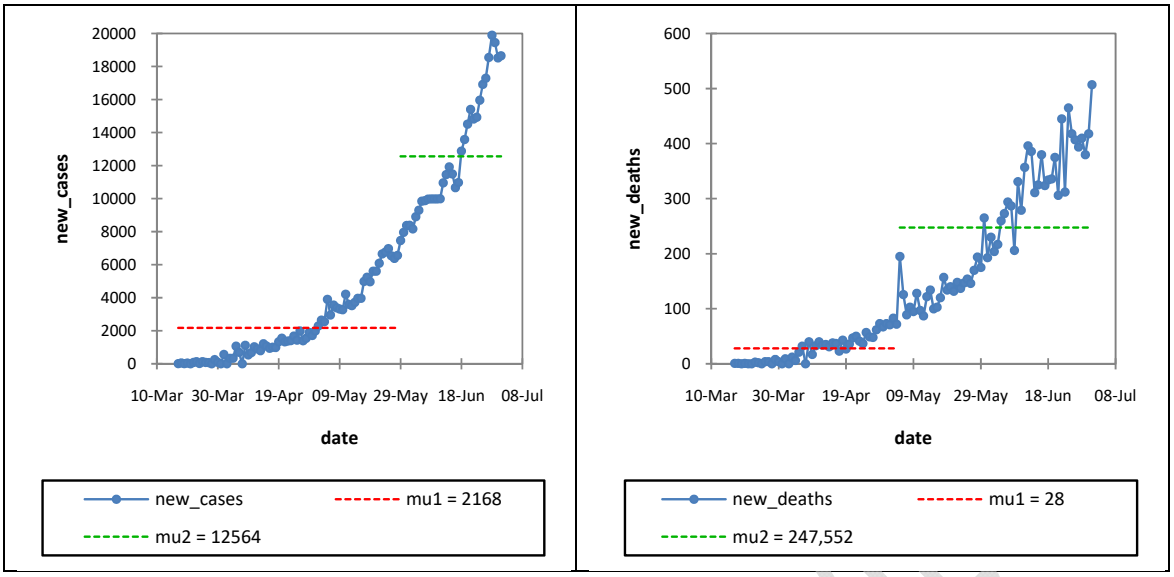

Figure2 Dynamics and structural changes of the new cases and new death time series; mu1 and mu2 are (respectively) the average (mean) number of cases (deaths or confirmed cases) for the first regime in red (or cycle) and second regime in green.

For normality assumption, the kurtosis is nearly than in a normal distribution, this assumption is an important one to fit the ARIMA models; in contrast the skewness measures (0.79 and 0.97) for both new cases and new deaths time series indicate data are skewed right in figure2.

What about the correlation between dynamics in confirmed new cases and specific new deaths?

As logical and biological facts, there is a relationship between the number of the confirmed new cases and numbers of death caused by a specific disease. Statistically, the functional form of this relationship varies from a disease (or a pandemic) to other. In case of COVID-19 data from India, we estimate the cross-correlation function (CCF), which is a generalization of simple correlation coefficient between these two variables. The results are shown below,

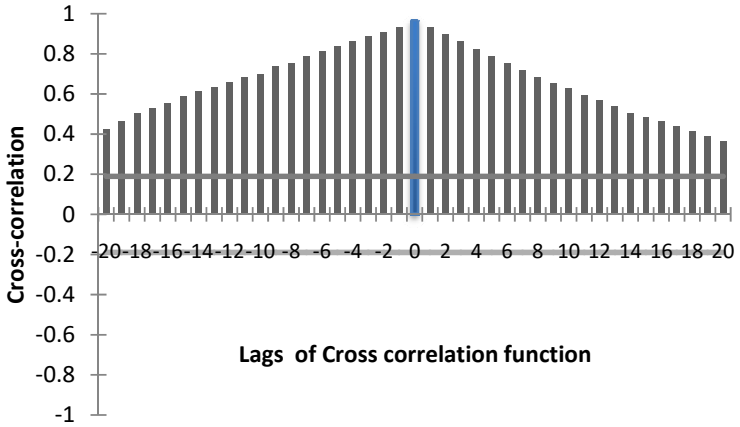

Figure3 . displays the cross correlations from both positive and negative lags. The value at lag 0 is the simple correlation between these two variables, it equals 0.9668.

The CCF depicts a positive and symmetric (compared to simple correlation at lag 0) correlation through the lags $[-20, -19, \dots, 0, \dots, 19, 20]$. When we jump to the causal inference, we run a simple linear regression model to estimate the effect of the new cases (NC) variable on the new number of deaths (ND), the estimation results are summarized in Table(2).

Table.2 Estimation **of Model parameters**

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	13.2921	4.9170	2.7033	0.0080	3.5425	23.0416
Newcase	0.0245	0.0006	38.7830	< 0,0001	0.0232	0.0257

Note: the coefficient of determination $R^2 = 0.937$, the *p-value* of the F-statistics is < 0.005, we accept this model.

The utility of such modeling reside in predictive the future death cases and the imputation of missing observations for the two time series; for more reliability and result accuracy, we can simply generalize the estimation of this relationship for other countries. It indicates that lag regression relationship between new total cases and new deaths.

Time Series modeling and Forecasting

According to the partial autocorrelation functions (PACF), which have considered as an important in data analysis and modeling, especially to identify the lag extending in Box-

Jenkins approach to indentify the stationary of dataset. From the (Figure-3.), the new cases time series on day (t) depends only and strongly with the past days (t-1) confirmed new cases. In contrast, the dependence structure of the new death time series is featured by a positive multi-lagged dependence (three days).

Table-3 Estimation accuracy measures for the three methods

Time Series	Method	ME	RMSE	MAE	MASE
New Cases	FTS	-641.1	917.6	773.9	1.723
	ARIMA(3,2,3)	62.41	436.55	34.97	0.843
	ANN	**19.22**	**38.22**	**23.12**	**0.564**
New Deaths	FTS	-14.55	66.86	39.8	1.891
	ARIMA(1,1,1)	-0.141	31.08	21.10	0.927
	ANN	7.23	614.66	145.33	4.239

Notes. The optimal ARIMA models have been selected according to the information criterion: BIC and AIC.

According to fitting adequacy for both new cases and new deaths time series, we see clearly from Table-3 that the ARIMA model fit better the data compared with Fuzzy time series models.

Table-4 Prediction of confirmed new cases and new deaths in India for the period (08-07 to 14-07 2020)

Time series	Date	FTS	ANN	ARIMA
New cases	08-07-2020	21085	22458	21849
	09-07-2020	20038	22664	22537
	10-07-2020	19245	22870	24224
	11-07-2020	18977	23076	26595
	12-07-2020	19583	23282	27840
	13-07-2020	20665	23488	27368
	14-07-2020	21743	23694	26405
New deaths	08-07-2020	505	480	457
	09-07-2020	534	505	456
	10-07-2020	558	492	462
	11-07-2020	580	496	466
	12-07-2020	599	501	470
	13-07-2020	618	508	474
	14-07-2020	636	511	478

Notes: the forecasts were estimated by the R program.

Table 4 and figure 4 depicts the forecast results for the nearest future of virus spread in India measured by the new cases and new deaths dynamics over the last 3 months. We stated that the three statistical methods provide us nearly the same expected trajectory of the virus in India. The common feature is a positive dynamics especially in term of daily new deaths

caused by this virus, where the ANN models seem to expect a speed dynamic compared with other two methods (FTS and ARIMA).

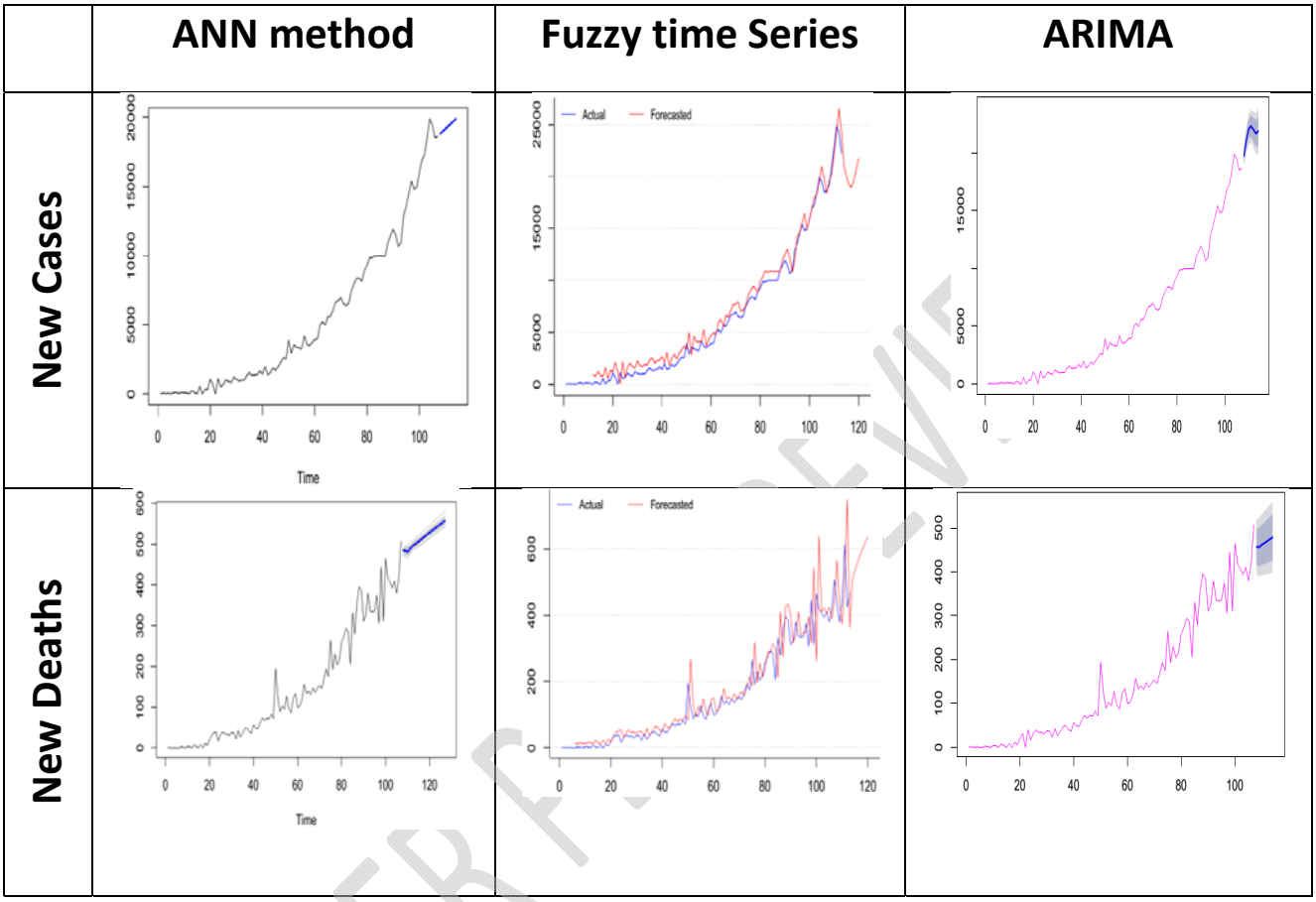

Figure-4 Plots forecasts results for new cases and new deaths time series by the three methods

Conclusion

This paper, deal with modeling and forecasting the short-term trajectory of the Covid-19 in India, where the Artificial Neural Network models, Fuzzy Time series and Box-jenkins approach on the 4 months data of the new cases and new deaths have been applied. Also in last four month, India has worked lot of medical facilities due to death rate decreased over the time. The results indicate and expect a positive dynamics of the virus spread, especially in terms of deaths over the next weeks. For models selections, the ARIMA and FTS are found more appropriate to forecast the virus trajectories.

References

[1] Chen, Y., Liu, Q., Guo, D.(2020). *Emerging coronaviruses: Genome structure, replication, and pathogenesis.* J Med Virol .

[2] Paules, C.I., Marston, H.D., Fauci, A.S.(2020) *Coronavirus infections-More than just the common cold.* JAMA - J Am Med Assoc .

[3]World Health Organistation (2020) : Coronavirus disease 2019 (COVID-19) Situation Report – 181.

[4] Fanelli, D., Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals*, *134*, 109761.

[5] Chimmula, V. K. R., Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, 109864.

[6] Malavika, B., Marimuthu, S., Joy, M., Nadaraj, A., Asirvatham, E. S., Jeyaseelan, L. (2020). Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models. *Clinical Epidemiology and Global Health*.

[7] Petropoulos, F., Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. *PloS one*, *15*(3), e0231236.

[8] Alzahrani, S. I., Aljamaan, I. A., Al-Fakih, E. A. (2020). Forecasting the Spread of the COVID-19 Pandemic in Saudi Arabia Using ARIMA Prediction Model Under Current Public Health Interventions. *Journal of Infection and Public Health*.

[9] Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B., Aslam, W., Choi, G. S. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access*

[10] Sujath, R., Chatterjee, J. M., Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, 1.

[11] Gardner Jr, E. S. (1985). Exponential smoothing: The state of the art. *Journal of forecasting*, *4*(1):1-28.

[12] Box, G ., Jenkins, Gwilym (1970). Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

[13] Xie, Y., Zhang, Y., Ye, Z. (2007). Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition. *Computer-Aided Civil and Infrastructure Engineering*, *22*(5):326-334.

[14] Zhang, G., Patuwo, B. E., Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, *14*(1):35-62.

[15] Khashei, M., Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with applications*, *37*(1): 479-489.

[16] Abbasov, A.M., Mamedova, M.H., (2003). Application of fuzzy time series to population forecasting, *Proceedings of 8th Symposium on Information Technology in Urban and Spatial Planning*, Vienna University of Technology, February 25-March1, : 545-552.

[17] Singh, S.R., (2008). A computational method of forecasting based on fuzzy time series. Mathematics and Computers in Simulation. 79: 539-554.

[18] Zhang, G., Patuwo, B. E., Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, *14*(1):35-62.

[19] Mandal, P., Senjyu, T., Funabashi, T. (2005, December). Neural network models to predict short-term electricity prices and loads. In *2005 IEEE International Conference on Industrial Technology* :164-169). IEEE.

[20] Jain, A., Kumar, A. M. (2007). Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing*, *7*(2):585-592.

[21] Price, R. K., Spitznagel, E. L., Downey, T. J., Meyer, D. J., Risk, N. K., El-Ghazzawy, O. G. (2000). Applying artificial neural network models to clinical decision making. *Psychological Assessment*, *12*(1): 40.

[22] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8: 338–353.

[23] Song.Q and Chissom. B. S.(1993). *Forecasting enrollments with fuzzy time series-part 1*. Fuzzy Sets and Systems, vol. 54:1-9.

[24] Chen, S.M. Hsu, C.C., (2004). A New method to forecast enrollments using fuzzy time series. *International Journal of Applied Science and Engineering*, 12: 234-244.

[25] Huarng, H., (2001). Huarng models of fuzzy time series for forecasting. *Fuzzy Sets and Systems*. 123: 369-386.

[26] Jiang, P Q. Dong, P. Li, Lian. L. (2017). A novel high-order weighted fuzzy time series model and its application in nonlinear time series prediction, *Appl. Soft Comput*. 55: 44–62

[27] Mishra, P., Fatih, C., Niranjan, H.K., Tiwari, S., Devi, M. and Dubey, A (2020). ,Modelling and Forecasting of Milk Production in Chhattisgarh and India. Indian Journal of Animal Research, 54 (7): 912-917