The Prediction Analysis of COVID-19 Cases Using ARIMA and KALMAN Filter Models: A Case of Comparative Study

Murali Krishna Iyyanki and Jayanthi Prisilla

Abstract The time series technique in machine learning is one of the important spaces for analysis and prediction. It includes many approaches to predict that involves time component. In the chapter, two approaches, i.e., autoregressive integrated moving average (ARIMA) and KALMAN filter models were demonstrated on the corona data (India) that was obtained from Ministry of Health and Family Welfare Web site. On modeling, it was found that ARIMA model gave better performance model over the KALMAN filter model. ARIMA (1, 1, 0) gave the approximate value of 35,303 for May 1, 2020 with sigma equal to 199.32, whereas the state-space model and error model of KALMAN filter generated the value of 33,116 and variance equal to 1356.18. The key purpose of the study is to understand and estimate the number of hospital beds and nursing care beds for the COVID-19 (CV-19) patients and make the indispensable arrangement for the patient treatment and avoid delay in action. In recovery cases, the highest value of difference is observed as 1153 on April 27, 2020, whereas the increases in reported cases are 2082 on April 28, 2020. More number of cases are reported with the peak in Maharashtra of 9915 (confirmed) and 1593 (recovery) on April 30, 2020. COVID-19 data visualization was carried out geographical information system with red color referring to the danger or more number of COVID-19 affected areas/state. Green color refers to normal and blue color refers to safe zone with no or single digit cases reported.

M. K. Iyyanki (✉)
Defence Research Development Organization, Hyderabad, India
e-mail: iyyanki@gmail.com

J. Prisilla
Technical Support Engineering, The Airports Authority of India Ltd, Hyderabad, India
e-mail: prisillaj28@gmail.com

C. Chakraborty et al. (eds.), *Internet of Medical Things for Smart Healthcare*, Studies in Big Data 80, https://doi.org/10.1007/978-981-15-8097-0_7

1 Introduction

In machine learning, the analysis of time series is found to be very popular and standard that is performed using various models. The experimental data analysis was observed at various points in time leads to new and unique complications in statistical modeling and inference [1]. In this chapter, ARIMA and KALMAN filter models are discussed for predicting COVID-19 cases. The prediction approach of events through a time sequence is referred as time series forecasting. By analyzing the historical trends of the past, assumption is favored for future trends. Time series (TS) are used in every field from medicine to finance, business, inventory planning, and dynamic system theory. The modern application of TS forecasting uses computer technologies that include machine learning, artificial neural networks, support vector machines, and so on. It is well-quoted by a data scientist that "time series forecasting is something of a dark horse in data science." On the other hand, according to Tealab [2] time series is a general problem solution of great practical interest in various disciplines. TS have evidence about the predictor variables of any system which determines dynamically. It is a sequence of values over the time of a system $y(t)$ which registers a sequence of experimental values given as $y\ (t_1)$, $y\ (t_2)$, $y\ (t_3)$,..., $y\ (t_n)$ for certain interval $t = n$ where $t_0 < t_1 < \ldots < t_n$. The aim of the study is to have the count of hospital beds and nursing beds made available on the prediction made to avoid delays and rushing. This would help the healthcare centers to arrange and be vigilant.

2 Predictive Modeling

Predictive modeling (PM) is a practice that uses data and mathematics to predict outcomes with data models. On the other hand, machine learning (ML) algorithms build the mathematical model based on the training data for prediction; ML algorithms uses statistical techniques to allow a computer to construct PMs. Predictive model stirs relations between ML, pattern recognition, and data mining. PM includes much more than the tools and techniques for unveiling patterns within data. PM training defines the development of a model process in a way that can understand and quantify the model's prediction accuracy on future, yet-to-be-seen data. The prime aim of PM is to produce accurate predictions and next is to interpret the model and understand how it works. But unfortunate reality/certainty is that as the model is pushed toward higher accuracy, models become more complex and their interpretability becomes more difficult [3]. PM performs curve and surface fitting, TS regression, or/and ML methods. One such example of TS regression; where the key convention of regression methods is that the patterns in the past data will be repeated in the future [4]. In this work, time series approach is carried out using ARIMA and KALMAN filter approach, the predictive results of CV-19 were analyzed to find that

the ARIMA model gave the nearest results of the confirmed cases in India. The objective of this prediction study is to understand the need of hospital beds and nursing care beds for CV-19 patients. This study helps to make the necessary arrangements for number of patient in-advance and to be cared for.

3 Time Series Using COVID-19 Datasets

A time series (TS) is a set of series of data points listed in the time order. A sequence that is successive equal spread out in points with time. The analysis encompasses methods for analyzing TS data to extract meaningful statistics and other data characteristics. The forecasting model of TS uses future values based on previously observed values for prediction. The time series data components are trend, seasonal variation, cyclical variation, and other irregular fluctuations.

Elmousalami [5] in their case study of CV-19 of analysis and modeling performed single exponential smoothing (SES) on the datasets of international confirmed cases. Figure 1 shows the graph of SES obtained and the Eq. 1 of SES is given as

$$F_{t+1} = (1 - \alpha) F_t + \alpha D_t \tag{1}$$

The results in Table 1 show that SES has the most accurate model for forecasting recovered cases of CV-19 with 517.54, 523335.16, 723.42, and 16.38% for mean absolute deviation (MAD), mean square error (MSE), root mean square

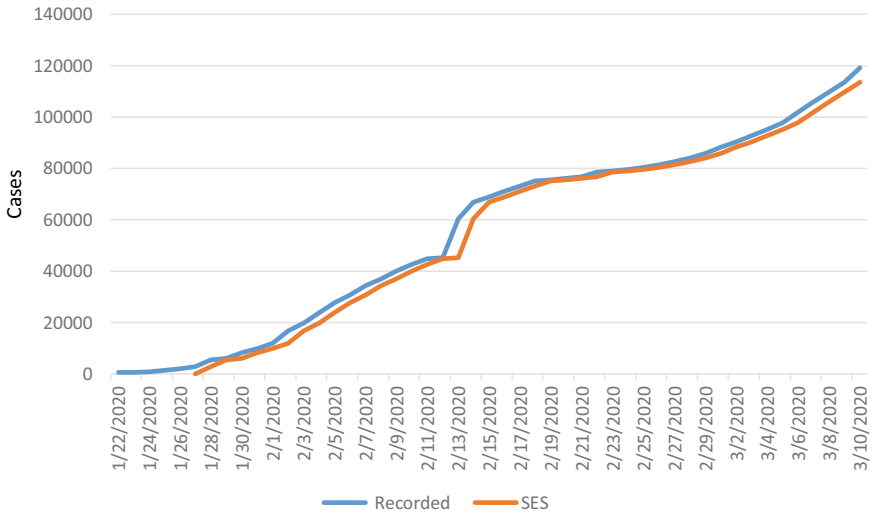

Fig. 1 SES for predicting the confirmed cases (international) [5]

Table 1 Forecasting models for international confirmed cases [5]

Model/Criteria	MAD	MSE	RMSE	MAPE
MA				
5 weeks ahead	7602.60	77470882.22	8801.75	21.42%
10 weeks ahead	18039.43	242692414.14	15578.59	28.18%
WMA	4614.96	32413677.85	5693.30	11.16%
SES	3385.65	20050014.56	4477.72	9.68%

error (RMSE), mean absolute percentage error (MAPE), respectively, against moving average (MA) and weighted moving average WMA.

Siedner [6] in their study of CV-19 in USA suggests that the due to social distancing, there is a lot of reduction in mean daily growth rate of CV-19 cases. The study involved a cumulative epidemic size of 4,171 cases (USA) where the reduction in growth rate estimated corresponds with a reduction in total cases from 26,356 to 23,266 at 7 days, and from 156,360 to 88,105 at 14 days after implementation. In brief, the uninterrupted TS model suggests that social distancing reduced the total number of CV-19 cases by nearly about 3,090 cases in 7 days after implementation and by 68,255 cases in 14 days. Table 2 displays the outcome of regression model for the growth rate daily wise after the social distancing was implemented.

In this study of CV-19 with dataset of different states of India, TS graph was implemented to understand the visualization of reported and recovery cases at a time. The data was obtained from https://www.mohfw.gov.in/ and the analysis is carried out on STATA-12 software.

The graph displays the different states confirmed (Fig. 2) and recovery (Fig. 3), in both the graphs Maharashtra is at the peak with 9915 (confirmed) and 1593 recovery cases on April 30, 2020.

Table 2 Linear regress for daily growth rate before versus after implementation of the first state-wide social distancing measure and state-wide restrictions on the internal movement [6]

	First state-wide social distancing measure			State-wide restriction on internal movement		
	Coef.	95% CI	P-value	Coef.	95% CI	P-value
Const	0.306	0.286 to 0.327	<0.001	0.209	0.190 to 0.229	<0.001
Time	−0.002	−0.007 to 0.004	0.53	−0.009	−0.011 to −0.006	<0.001
Post-intervention period	0.002	−0.032 to 0.035	0.92	−0.039	−0.138 to 0.06	0.44
Time x post-intervention period	−0.008	−0.014 to −0.002	0.008	0.003	−0.014 to 0.020	0.72

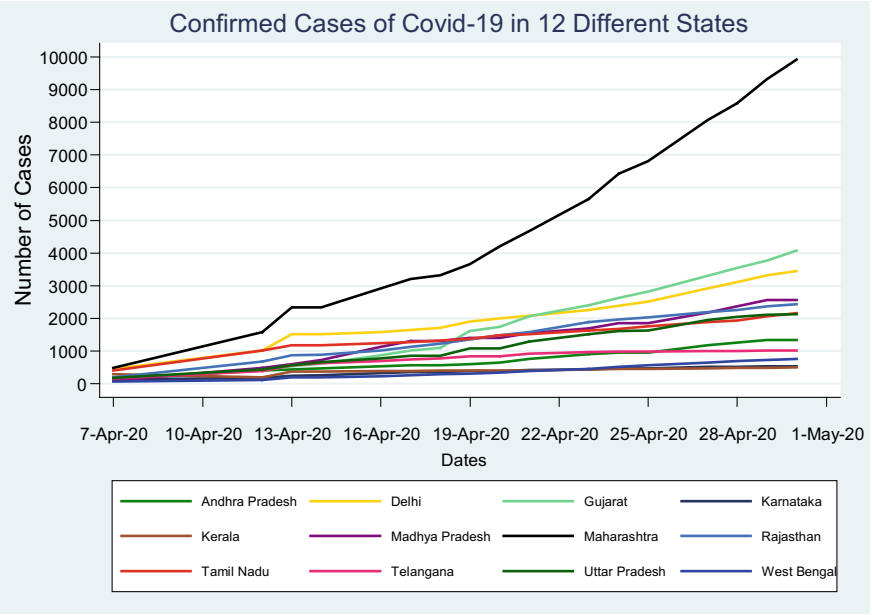

Fig. 2 Reported cases of different Indian states affected due to COVID-19

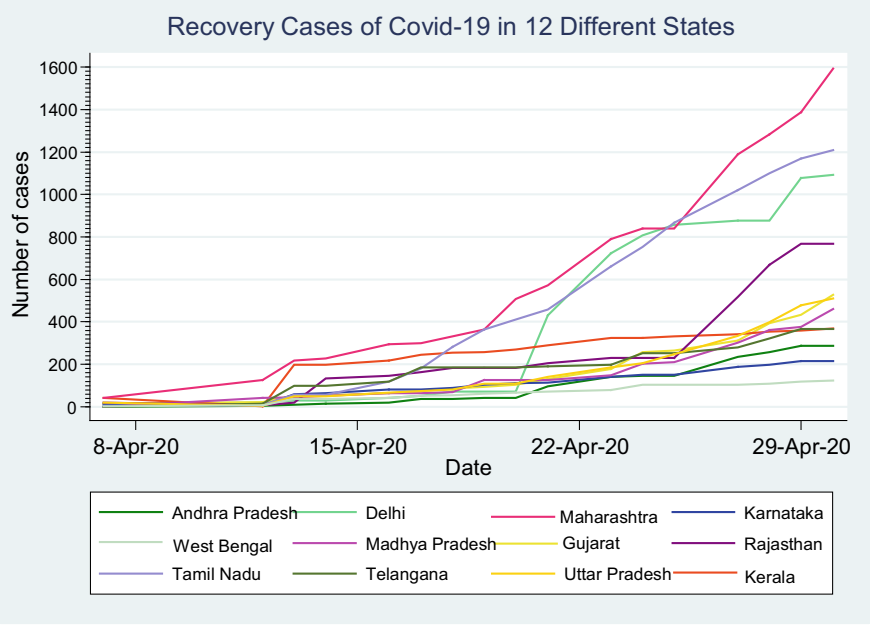

Fig. 3 Recovery cases in different states of India

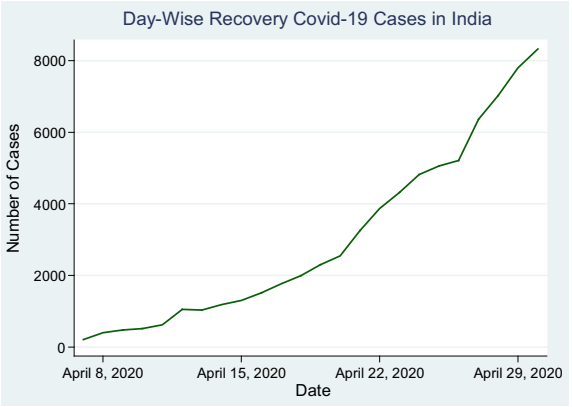

Fig. 4 Day-wise recovery cases graph

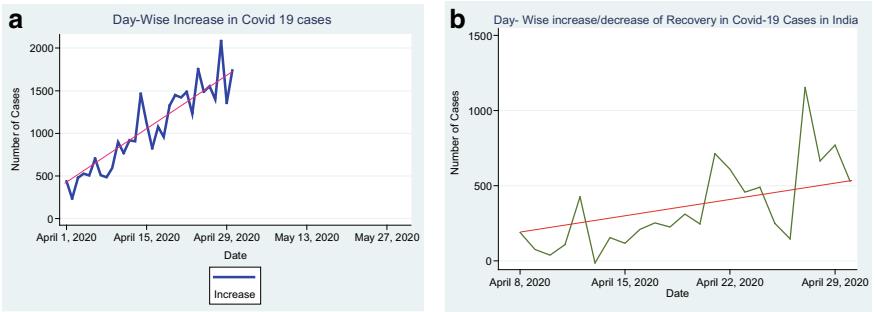

Fig. 5 **a** Day-wise increase in confirm cases, **b** increase/decrease graph for recovery cases

Figure 4 shows the recovery cases of CV-19 in India from March 14, 2020 to April 30, 2020. The graph represents slight decrease on April 13, 2020 giving April 14 on subtracting from April 12, 2020 recovery data. Figure 5 shows the comparison of two—reported cases (Fig. 5a) and recovery cases (Fig. 5b) with increase/decrease in the number of cases. The peak value in recovery difference is 1153 on April 27, 2020, whereas the highest increase in confirmed case is 2082 on April 28, 2020.

4 ARIMA

In TS exploration, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. ARIMA (ARM) models are the best models for the statistical models for analyzing and forecasting TS data. An ARM model is a filter that separates the data from the noise, and the

data is then extrapolated to obtain forecasts. The forecasting equation of ARM for stationary TS is a linear (regress) equation in which the predictors include of lags of the response variable and/or lags of the forecast errors. Predicted value (Y) is calculated with a constant and/or a weighted sum of one or more current values of Y and/or a weighted sum of one or more current values of the errors.

ARM models complex data pattern; and uses the export modeler for outlier detection, and produces for the drive of eXtended Markup Language (XML) files for prediction modeling of future data.

4.1 The Notation of ARIMA (P, D, Q)

The ARM model consists of autoregressive (AR), moving average (MA), and seasonal autoregressive integrated moving average (SARIMA) models [7]. The autoregressive terms are lags of the stationaries series in the forecasting equation; moving average are lags of the forecast errors, and a TS which needs to be differenced to be made stationary is integrated of a stationary series.

A non-seasonal ARM model is written as an ARIMA (p, d, q) model where p is the sum of autoregressive terms, d is the sum of integrated differences order, and q is the sum of moving average (lagged forecast errors) in the prediction equation.

Hence, the forecasting Eqs. 2, 3, and 4 is built as follows [8, 9].

$$\left(1 - \Phi_1 B - \ldots - \Phi_p B^p\right)\left(1 - \Phi_1 B^s - \ldots - \Phi_p B^{\mathrm{ps}}\right)(1 - B)^d \tag{2}$$

$$\left(1 - B^s\right)^D y_t = \left(1 + \theta_1 B + \ldots + \theta_q B^q\right)\left(1 + \ominus_1 B^s + \ldots + \ominus_Q B^{QS}\right) \in_t \tag{3}$$

where B represents the backshift operator that is defined by the following operation:

$$B^m y_t = y_t - m \tag{4}$$

whenever the parameter has a value of 0 is used; it represents that not to use that element of the model.

4.2 ARIMA in COVID-19 Cases—Datasets

In a case study of ARIMA model, [10] used the model for predicting the electricity prices. The two ARIMA models −1 and 2 were used predicted hourly prices in the electricity markets of Spain and California. The model of Spanish requires 5 h to predict future prices, as opposed to the 2 h needed by the Californian model. The

spot markets and long-term contracts, price forecasts are essential for developing bid strategies or negotiation skills.

Model 1 is given as

$$\begin{aligned}&\left(1-\Phi_1 B^1-\Phi_2 B^2-\Phi_3 B^3-\Phi_4 B^4-\Phi_5 B^5\right)\\&\times\left(1-\Phi_{23} B^{23}-\Phi_{24} B^{24}-\Phi_{47} B^{47}\right.\\&\left.-\Phi_{48} B^{48}-\Phi_{72} B^{72}-\Phi_{96} B^{96}-\Phi_{120} B^{120}-\Phi_{144} B^{144}\right)\\&\times\left(1-\Phi_{168} B^{168}-\Phi_{336} B^{336}-\Phi_{504} B^{504}\right)\log p_t=c\\&+\left(1-\theta_1 B^1-\theta_2 B^2\right)\left(1-\theta_{24} B^{24}\right)\\&\times\left(1-\theta_{168} B^{168}-\theta_{336} B^{336}-\theta_{504} B^{504}\right)\varepsilon_t\end{aligned}\tag{5}$$

Model 2 is given as

$$\begin{aligned}&\left(1-\Phi_1 B^1-\Phi_2 B^2\right)\times\left(1-\Phi_{23} B^{23}-\Phi_{24} B^{24}-\Phi_{47} B^{47}-\Phi_{48} B^{48}\right.\\&\left.-\Phi_{72} B^{72}-\Phi_{96} B^{96}-\Phi_{120} B^{120}-\Phi_{144} B^{144}\right)\\&\times\left(1-\Phi_{167} B^{167}-\Phi_{169} B^{169}-\Phi_{192} B^{192}\right)\times(1-B)\left(1-B^{24}\right)\left(1-B^{168}\right)\\&\log p_t=c+\left(1-\theta_1 B^1-\theta_2 B^2\right)\left(1-\theta_{24} B^{24}-\theta_{48} B^{48}-\theta_{72} B^{72}-\theta_{96} B^{96}\right)\\&\times\left(1-\theta_{144} B^{144}\right)\times\left(1-\theta_{168} B^{168}-\theta_{336} B^{336}-\theta_{504} B^{504}\right)\varepsilon_t\end{aligned}\tag{6}$$

Tables 3 and 4 are the statistical values of forecast mean square of errors (FMSE) was obtained on application of model 1 and 2. Table 5 displays the estimated and parameter values of two countries models.

Table 3 Statistical without explanatory variables [10]

	MWE (%)	$\bar{s}_R$	$\sqrt{FMSE}$
January (Spain)	12.06	0.106	71.98
February (Spain)	8.05	0.106	36.77
March (Spain)	11.28	0.104	71.75
April (Spain)	19.37	0.104	61.51
May (Spain)	4.99	0.083	19.91
June (Spain)	9.97	0.061	81.14
July (Spain)	9.39	0.067	42.59
August (Spain)	8.17	0.092	48.13
September (Spain)	12.01	0.097	70.82
October (Spain)	13.63	0.097	80.33
November (Spain)	7.32	0.098	47.51
April (California)	5.01	0.060	21.19
August (California)	15.65	0.121	469.85
November (California)	13.6	0.074	393.23

Table 4 Statistical with explanatory variables [10]

	MWE (%)	$\bar{s}_R$	$\sqrt{FMSE}$
January (Spain)	9.97	0.106	64.72
February (Spain)	8.13	0.107	45.10
March (Spain)	10.5	0.105	71.57
April (Spain)	14.68	0.102	45.24
May (Spain)	7.75	0.082	33.35
June (Spain)	10.8	0.061	80.99
July (Spain)	8.83	0.066	41.80
August (Spain)	9.39	0.092	49.35
September (Spain)	10.72	0.097	65.50
October (Spain)	13.69	0.094	77.57
November (Spain)	9.88	0.098	73.73
April (California)	5.21	0.060	21.82
August (California)	21.03	0.123	674.58
November (California)	13.68	0.074	397.27

Noureen [9] in a case study of ARIMA in forecasting is a small-scale agricultural load. For the TS data, ARIMA method was applied on the stationary TS data. As seasonal variations make a TS nonstationary, this study presented an analyses on testing stationarity and transforming non-stationarity into stationarity. The model was developed with a specific order selection for autoregressive terms, moving average terms, differencing and seasonality and the forecasting performance has been tested and compared with the actual value. After the plotting of ACF and PACF, augmented Dickey fuller (ADF) test is performed for hypothesis testing to confirm stationarity of TS. ADF is also known as unit root test. The model for the ADF test is shown in Eq. (7):

$$\partial Y_t = \mu + \beta t + \rho Y_{t-1} + \partial Y_{t-1} + \ldots + \partial_p Y_{t-p} + e_t \tag{7}$$

The seasonal ARIMA model is implemented to forecast the agricultural loads for the last one year of the three-year data. The mean absolute error (MAE) of our forecast is calculated to be 13.23%.

Benvenuto [7] implemented ARIMA on a dataset consisting of 22 number determinations. The overall prevalence of CV-19 presented an increasing trend that reached the epidemic plateau as shown in Fig. 6 and Table 6 gives the predicted values for the two days. The difference between cases of a day and cases of the previous day Δ(Xn-Xn-1) showed a non-constant increase in the number of confirmed cases. Figure 7 displays the correlogram and ARIMA forecast graph for the 2019-nCoV incidence.

Table 5 Estimated parameter values of the Spanish and Californian ARIMA models [10]

Spanish market		Californian market	
Parameters	Estimate	Parameters	Estimate
c	−0.0052	c	−0.00015
Φ_1	0.5432	Φ_1	0.447
Φ_2	0.8373	Φ_2	0.206
Φ_3	−0.4174	Φ_{23}	0.0625
Φ_4	−0.0271	Φ_{24}	−0.0193
Φ_5	0.0243	Φ_{47}	−0.0095
Φ_{23}	0.0384	Φ_{48}	−0.3013
Φ_{24}	0.4165	Φ_{72}	−0.021
Φ_{47}	0.0532	Φ_{96}	−0.0581
Φ_{48}	0.004	Φ_{120}	−0.0698
Φ_{72}	0.0196	Φ_{144}	−0.2739
Φ_{96}	0.019	Φ_{167}	0.0222
Φ_{120}	−0.003	Φ_{168}	−0.5922
Φ_{144}	0.1474	Φ_{169}	0.0595
Φ_{168}	0.3342	Φ_{192}	−0.0681
Φ_{336}	0.2873	θ_1	0.9326
Φ_{504}	0.2661	θ_2	0.0291
θ_1	−0.0941	θ_{24}	0.7752
θ_2	0.6998	θ_{48}	−0.2376
θ_{24}	0.1607	θ_{72}	0.2386
θ_{168}	0.2304	θ_{96}	−0.0053
θ_{336}	0.1726	θ_{144}	−0.2713
θ_{504}	0.2232	θ_{168}	0.0248
		θ_{336}	0.5082
		θ_{504}	0.0007

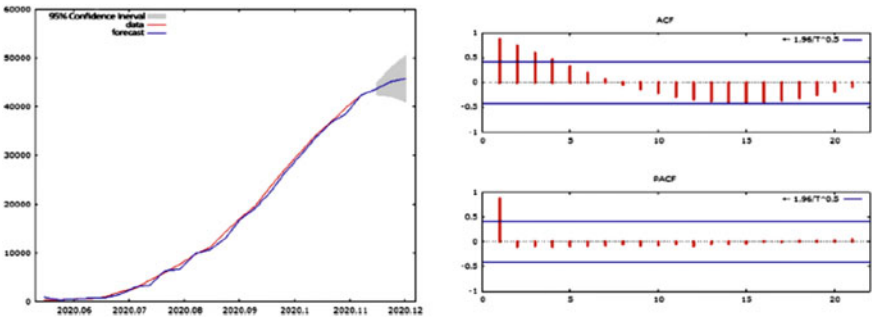

Fig. 6 Correlogram and ARIMA forecast graph for the 2019-nCoV prevalence [7]

Table 6 Forecast value for two days after the analysis for the prevalence and for the incidence of the CV-19 [7]

	Date	Forecast	95% CI
Prevalence	February 11, 20	43599.71	42347.53–44851.9
	February 12, 20	45151.45	42084.88–48218.02
Incidence	February 11, 20	2070.66	1305.23–2836.09
	February 12, 20	2418.47	1534.43–3302.51

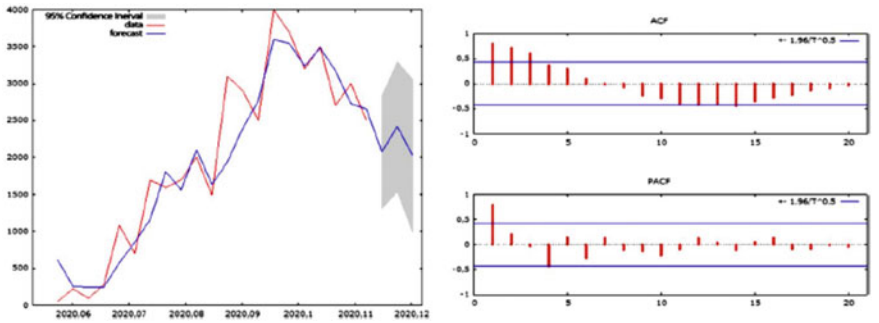

Fig. 7 Correlogram and ARIMA forecast graph for the 2019-nCoV incidence [7]

4.3 ARIMA Model on COVID-19—India Dataset

In this case study of COVID-19 (India), ARIMA (ARM) model was built using STATA software. Here, the comparative study on the prediction results obtained from the two models state that ARIMA (1,1,0) model gives much better accurate results over the KF predicted values. The number of cases reported is shown in Table 7 and ARIMA (1,1,0) was modeled to obtain Table 8 with log likelihood value of −316.86; and the predicted value of both the models is shown in Table 15.

Using ARM model, when the parameters of the model were given as $p = 1$, $d = 1$, and $q = 0$; then the p-value = 0 and the predicted values were to the nearest data values. The z-test statistic for the predictor (ConfirmCases) is 823.3/554.8 = 1.48. Coefficient of ARMA(ar) = 0.96; wald chi2(1) is wald chi-square statistic. It is mainly used for hypothesis test where at least one of the predictors' regression coefficients is not equal to zero. Here, in this case, 1 refers to the number of degrees of freedom of the chi-square distribution used to test the wald chi-square statistic and is distinct by the number of predictors (1)./sigma is the estimated standard error of the ARM regression with 199.32 value.

Correlograms/autocorrelation function (ACF) and partial correlograms/partial autocorrelation function (PACF) are shown in Fig. 8, with confidence interval (CI) of −0.9–0.9 in ACF and in PACF, CI is −0.03–0.03. The x-axis denotes the lag and y-axis represents the first-order differential of cases. The blue dot represents the autocorrelation between the lag variable and unlag variable of cases in this study. The

Table 7 Number of cases reported

Date	Number of cases reported
March 14, 2020	84
March 15, 2020	110
March 16, 2020	114
March 17, 2020	137
March 18, 2020	151
March 19, 2020	173
March 20, 2020	223
March 21, 2020	315
March 22, 2020	360
March 23, 2020	468
March 24, 2020	519
March 25, 2020	606
March 26, 2020	694
March 27, 2020	834
March 28, 2020	918
March 29, 2020	1,024
March 30, 2020	1,251
March 31, 2020	1,397
April 1, 2020	1,834
April 2, 2020	2,069
April 3, 2020	2,547
April 4, 2020	3,072
April 5, 2020	3,577
April 6, 2020	4,281
April 7, 2020	4,789
April 8, 2020	5,274
April 9, 2020	5,865
April 10, 2020	6,761
April 11, 2020	7,529
April 12, 2020	8,447
April 13, 2020	9,352
April 14, 2020	10,815
April 15, 2020	11,933
April 16, 2020	12,759
April 17, 2020	13,835
April 18, 2020	14,792
April 19, 2020	16,116

(continued)

Table 7 (continued)

Date	Number of cases reported
April 20, 2020	17,565
April 21, 2020	18,985
April 22, 2020	20,471
April 23, 2020	21,700
April 24, 2020	23,452
April 25, 2020	24,942
April 26, 2020	26,496
April 27, 2020	28,380
April 28, 2020	29,974
April 29, 2020	31,787
April 30, 2020	33,610

dots which are well-outside the interval are known to be large and will be least equal

Table 8 ARIMA model—ARIMA(1,1,0)

Arima Confirm_Cases1, arima (1,1,0)					
ARIMA regress					
Sample: March 15, 2020–April 30, 2020			Num. of obs = 47		
			Wald chi2(1) = 256.04		
LL = − 316.86			Prob > chi2 = 0.0000		
D.ConfirmCases1	Coef.	OPG SE.	z	P > |z|	[95% CI]
ConfirmCases1_con	823.3	554.8	1.48	0.138	−264.07 to 1910.74
ARMA ar.L1	0.96	0.60	16.00	0.000	0.84 to 1.08
/sigma	199.32	19.02	10.48	0.000	162.04 to 236.60

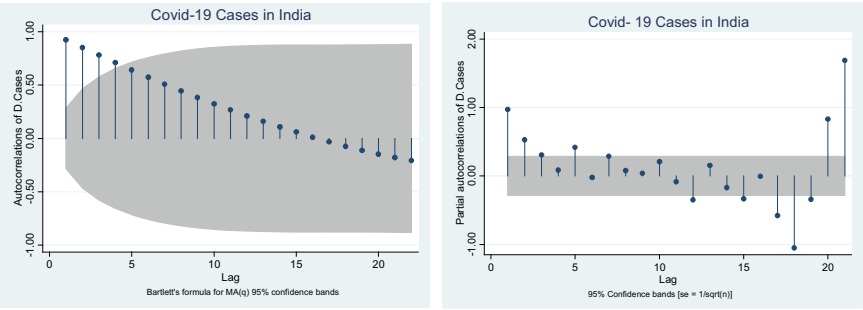

Fig. 8 ACF and PACF graph of COVID cases

Table 9 ACF and PACF values

Correlogram Confirm_Cases1, lags(20)					−1 0 1	−1 0 1
LAG	AC	PAC	Q	Prob > Q	[Autocorrelation]	[Partial autocor]
1	0.9176	1.0686	39.629	0.0000		
2	0.8364	−0.5011	73.34	0.0000		
3	0.7564	−0.3528	101.58	0.0000		
4	0.6801	−0.1634	124.99	0.0000		
5	0.6042	−0.1479	143.93	0.0000		
6	0.5309	−0.4203	158.95	0.0000		
7	0.4605	0.0746	170.54	0.0000		
8	0.3936	−0.3604	179.25	0.0000		
9	0.3298	−0.2024	185.54	0.0000		
10	0.2670	−0.2052	189.79	0.0000		
11	0.2063	−0.3993	192.4	0.0000		
12	0.1462	−0.4029	193.75	0.0000		
13	0.0893	0.4218	194.27	0.0000		
14	0.0389	0.5808	194.37	0.0000		
15	−0.0088	−1.0014	194.38	0.0000		
16	−0.0533	0.9417	194.58	0.0000		
17	−0.0949	0.5621	195.26	0.0000		
18	−0.1325	2.0889	196.63	0.0000		
19	−0.1674	1.5300	198.89	0.0000		
20	−0.2005	2.5197	202.28	0.0000		

to 1, i.e., $p = 1$. Each spike that rises above or falls below the CI range is considered to be statistically significant. ACF and PACF table is mentioned in Table 9.

The analysis procedures include ACF and PACF that are used to calculate correlation in the data [11].

5 KALMAN Filter

KALMAN filter (KF) is widely known as an *optimal estimator*—i.e., infers factors of interest from indirect, inaccurate, and uncertain observations. The new measurements are processed by the recursive property of KF. KF minimizes the mean square error of the estimated parameters, if the noise is Gaussian; and it is a best linear estimator, given the mean and standard deviation of the noise. The technique of finding the best estimate from noisy data amounts to filter out the noise is referred as filtering; and this practice is carried out by KF (Kleeman). KF is a two-step process, namely

prediction and update steps. For the likelihood, one has to find f ($y_t|Y_{t-1}$) [12]. The two steps are given in Eqs. 8 and 9. (prediction) and Eqs. 10, 11, and 12 (update).

1. Prediction equation

$$\hat{\mathbf{x}}_k^- = A\,\hat{\mathbf{x}}_{k-1}^- + BU_k \tag{8}$$

$$\mathbf{P}_k^- = A\,\mathbf{P}_{k-1}^- + A^T + \mathbf{Q} \tag{9.}$$

2. Updating equation

$$\mathbf{K}_\mathrm{k} = \mathbf{P}_\mathrm{k}\mathbf{C}^T\left(\mathbf{C}\mathbf{P}_\mathbf{k}^-\mathbf{C}^\mathrm{T} + \mathbf{R}\right)^{-1} \tag{10}$$

$$\hat{\mathbf{x}}_\mathbf{k} = \hat{\mathbf{x}}_\mathbf{k}^- + \mathbf{K}_\mathbf{k}\left(\mathbf{Y}_\mathbf{k} - \mathbf{C}\hat{\mathbf{x}}_\mathbf{k}^-\right) \tag{11}$$

$$P_k = (1 - K_k C)P_k^- \tag{12}$$

The state-space model consists of covariance and error forms; both the forms follow two equations first one is state Eq. 13 and observation Eq. 14. The notation of a state-space model is as follows:

$$y_t = Z_t\alpha_t + S_t\xi_t \tag{13}$$

$$\alpha_t = T_t\alpha_{t-1} + R_t\eta_t \tag{14}$$

with $\begin{pmatrix}\eta_t\\ \xi_t\end{pmatrix} \sim$ iid N $\left(0, \begin{bmatrix} Q & 0\\ 0 & H\end{bmatrix}\right)$ and the initial observation is given as $y_1 \sim N(y_{1|0}, F_1)$.

5.1 KALMAN Filter—for Prediction in Different Studies

Rhudy [13] in their work of KF using MATLAB gives an illustration of a simple object in freefall presuming that there is no air resistance. The purpose of filter is to determine the location of the object based on uncertain information about the starting location of the object as well as measurements of the location provided by a laser rangefinder. The acceleration of the given object will be the same to the acceleration due to gravity. In their study, the measurement system has a standard deviation of error of 2 m, and variance of 4 m^2. In the measurement noise, uncertainty in the initial state is considered. The starting point is known to be 105 m before the ball is dropped, while the actual starting point is 100 m as shown in Fig. 9. The initial guess

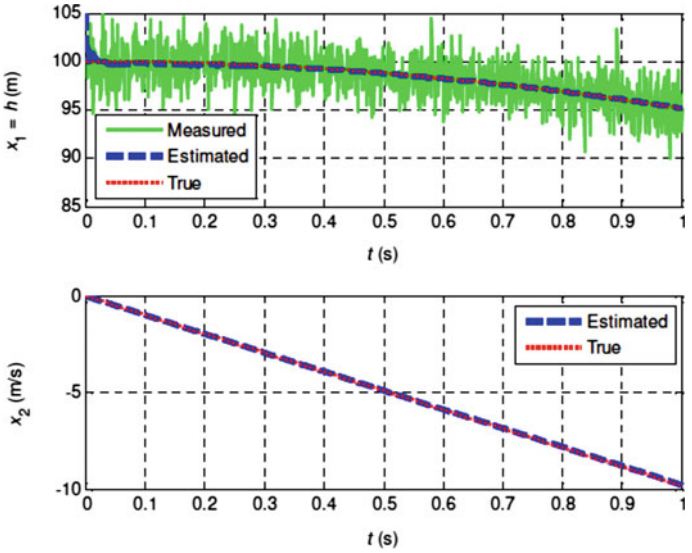

Fig. 9 Example of KF estimated and true states [13]

was roughly determined and has a relatively high corresponding initial covariance. The error of 10 m^2 for the initial position is assumed as the object starts from rest; a smaller uncertainty value of 0.01 m^2/s^2 is obtained as shown in Fig. 10.

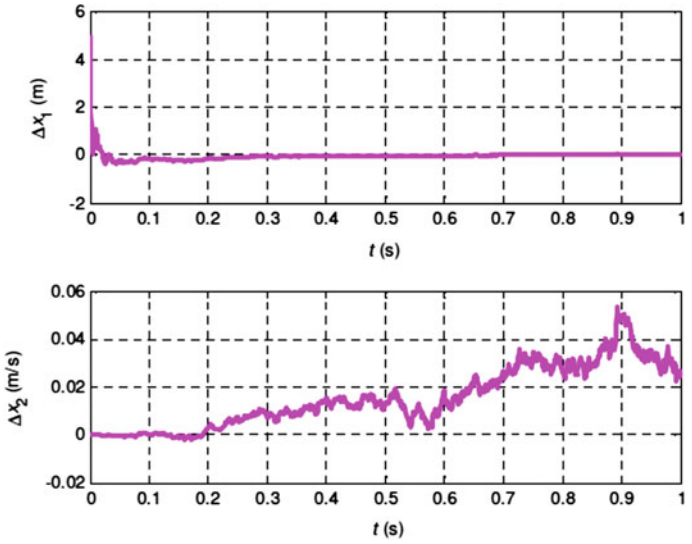

Fig. 10 Example of KF using estimation errors [13]

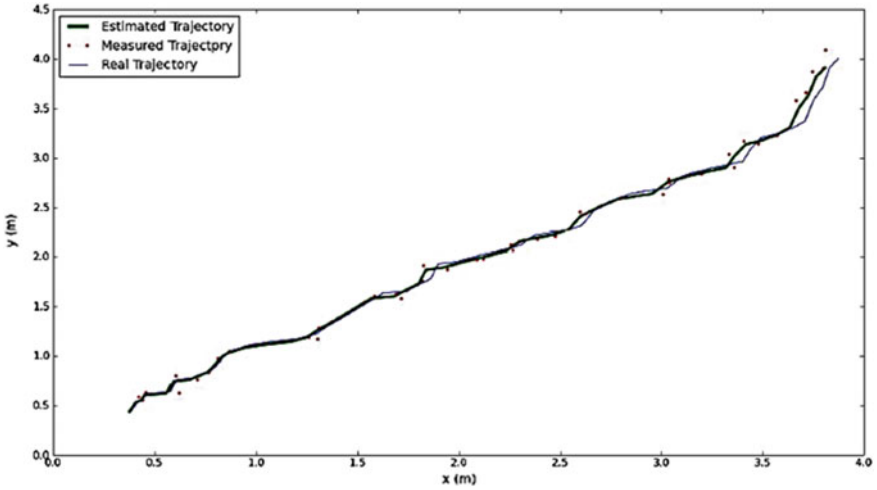

Fig. 11 KF applied to ToA-based localization [14]

Laaraiedh [14] in a case study of KF in telecommunications used on the mobile tracking user connected to a wireless network. A simple tracking algorithm was implemented using Python language by a mobile user who is moving in a room and connected to at least three wireless antennas. The estimated position of the mobile using a trilateration algorithm is indicated by matrix of measurement Y with at least three values of time of arrival (ToA) at time step k as shown in Fig. 11. The values are computed using ranging procedures between the mobile and the three antennas. Initialization of different matrices and using the updated matrices for each step and iteration; estimated, and the real trajectory of the mobile user, and the measurements are performed by the least square-based trilateration. KF enhances the tracking accuracy compared to the static least square-based estimation.

Rankin [15] in their case study of KF for the market price application was based on yearly, quarterly, monthly, weekly, and daily prices. A study was also carried out on open, high, low, and close prices. The use of averages (e.g., weekly or monthly) or stock indexes may alter the results of a study. Table 10 shows the comparison of expenses for the consumer strategy. The first sample (DJT#1) consisted of 1036 hourly readings from February 22, 1985 to September 23, 1985. The second (DJT#2) was of 896 hourly readings from January to June, 1984. The third sample (DJT#3) was gathered from July through December, 1983 and consisted of 896 samples from the 128 day period. KF program produced N-step ahead forecasts for TS. MSE of the forecast errors are calculated to measure model accuracy.

Malleswari [16] in a case study of KF in the error modeling (like ionospheric delays, atmospheric delays, tropospheric delays, and so on) affecting the GPS signals as they travel from satellite to the user who is on earth. In this methodology, it showed that the variations in the signal related to WGS—84 data can be smoothened using KF with the studies made and the analysis yielded better accuracies as shown in

Table 10 Expense comparison for consumer strategy [15]

Strategy	DJT#1	DJT#2	DJT#3
1-Day Period Open Close KF	$51,772 $51,785 $51,697	$57,576 $57,459 $57,429	$65,559 $65,585 $65,506
2-Day Period Open Close KF	$26,220 $26,230 $26,128	$28,832 $28,718 $27,670	$32,747 $32,782 $32,699
3-Day Period Open Close KF	$18,141 $18,145 $18,047	$19,788 $19,665 $19,642	$22,401 $22,433 $22,342
4-Day Period Open Close KF	$13,439 $13,443 $13,542	$14,754 $14,622 $14,607	$16,642 $16,679 $16,587
5-Day Period Open Close KF	$10,748 $10,741 $10,678	$12,231 $12,088 $12,061	$13,765 $13,806 $13,705

*DJT = Dow Jones Transportation

Tables 11 and 12 that Φkf—the latitude in degrees on KF application is 0.004221766 for Gandipet and 0.00003667424 for Hussain Sagar. Similarly, λkf—longitude in degrees on KF application KF is 0.03084715 for Gandipet and 0.0006331302 for Hussain Sagar.

Table 11 Comparison of longitude for Gandipet (left) and Hussain Sagar (right) [16]

Longitude	Variance	Longitude	Variance
λ**rx**—Longitude in degrees (Receiver)	27.0337204	λ**rx**—Longitude in degrees (Receiver)	13.39910013
λ_{prg}—Longitude in degrees before applying KALMAN filter	27.0337194	λ_{prg}—Longitude in degrees before applying KALMAN filter	0.001264126
λ_{kf}—Longitude in degrees after applying KALMAN filter	0.03084715	λ_{kf}—Longitude in degrees after applying KALMAN filter	0.0006331302
$\lambda_{s/w}$—Longitude in degrees after applying Web soft	6.904284042	$\lambda_{s/w}$—Longitude in degrees after applying Web soft	0.00355667

Table 12 Comparison of latitude for Gandipet (left) and Hussain Sagar (right) [16]

Latitude	Variance	Latitude	Variance
Φrx—Latitude in degrees (Receiver)	6.89938593	Φrx—Latitude in degrees (Receiver)	0.020659696
Φ_{prg}—Latitude in degrees before applying KALMAN filter	7.3488	Φ_{prg}—Latitude in degrees before applying KALMAN filter	0.000292464
Φ_{kf}—Latitude in degrees after applying KALMAN filter	0.004221766	Φ_{kf}—Latitude in degrees after applying KALMAN filter	0.00003667424
$\Phi_{s/w}$—Latitude in degrees after applying Web soft	0.011725697	$\Phi_{s/w}$—Latitude in degrees after applying Web soft	0.000198005

5.2 *KALMAN Filter—for COVID-19 Prediction—India Dataset*

There are two forms in state-space model, namely covariance and error form models. There are shown in Tables 13 and 14, respectively.

Table 13 State-space model—covariance

Space (CC L4.Datee, state) (Confirm_Cases1 L4.Datee) in 34/44					
State-space model					
Sample: April 16, 2020—April 26, 2020			Num. of obs = 11		
			Wald chi2(1) = 2923.86		
LL = − 77.19			Prob > chi2 = 0.0000		
ConfirmCases1	Coef.	OIM SE.	z	P > |z|	[95% CI]
CC Date L4. _cons	−0.468 0.301	– –	– –	– –	– –
Confirm_Cases1 Date L4. _cons	1392.43 −3.06e + 07	25.75 567088.4	54.07 −54.04	0.000 0.000	1341.96 to 1442.89 −3.18e + 07 to −2.95e + 07
Var. CC Confirm_Cases1	1356.18 73018.58	31135.21	2.35	0.010	11994.69 to 134042.5

*LL = Log Likelihood, I0 to I8 = Iteration 0 to 9, SE = standard error

Table 14 State-space model—error form

Space (CCE L4.Datee, state) (Confirm_Cases1 L4.Datee) in 34/44					
State-space model					
Sample: April 16, 2020—April 26, 2020				Num of obs = 11	
				Wald chi2(1) = 2923.86	
LL = − 77.19				Prob > chi2 = 0.0000	
ConfirmCases1	Coef.	OIM SE.	z	P > |z|	[95% CI]
CCE Date L4. _cons	−0.468 0.301	– –	– –	– –	– –
Confirm_Cases1 Date L4. _cons	1392.43 −3.06e + 07	25.75 567088.4	54.07 −54.04	0.000 0.000	1341.96–1442.89 −3.18e + 07 to − 2.95e + 07
Var CCE Confirm_Cases1	1356.18 73018.58	−31135.21	– 2.35	– 0.010	– 11994.69–134042.5

5.2.1 Covariance

See Table 13.

5.2.2 Error Form

In Fig. 12, the dotted lines show the prediction of CV-19 data obtained using ARIMA and KF model. The solid blue line indicates cases of training data. The results of KF model—covariance and error model are shown in Tables 13 and 14, respectively. The log likelihood in refine estimates is −77.19, wald chi2(1) is 2923.86. Z-test

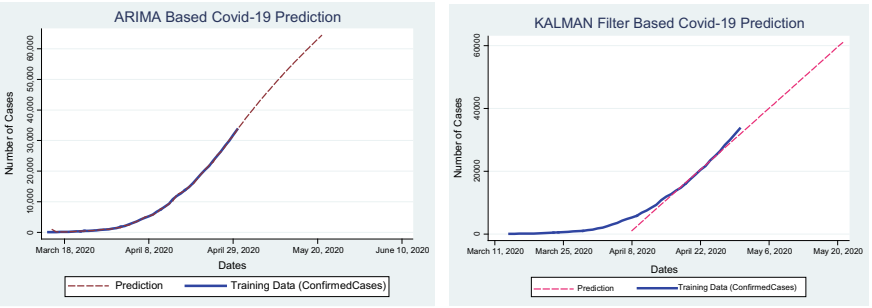

Fig. 12 Comparison of ARIMA and KALMAN filter prediction graphs

of predictor (Confirm_Cases1) is 73018.58/31135.21 = 2.35 and z-test of date is 1392.43/25.75 = 54.07. The variance is given as 1356.18. Two models are best fit model; as all of the *p*-values are very significant with $p < 0.001$ and $p < 0.05$.

Table 15 shows the ARM and KF predicted value from May 1, 2020 and the data. The prediction was calculated from May 1, 2020 up to May 20, 2020. The predicted values are thus compared with the data to check if it lies within the nearest range. Figure 13 describes the prediction of ARM and KF along with data in a single graph and on comparison one can see the predictive curve of ARM increases accurately with the dates, whereas the KF would not give accurate predicted values.

6 Geographic Information Systems—Visualization and Prediction—COVID-19 Datasets

Geographic information systems (GIS) are a computer-based tool that examines spatial relationships, patterns, and trends. This through connecting geography with data, GIS better understands data using a geographic context. It stores, analyze, and visualize data for geographic positions on earth's surface. The four main characteristics of GIS *create* geographic data, *manage* it in a database, *analyze* and find patterns, and *visualize* it on a global map. In viewing and analyzing data on maps gives better understanding of data, and one can make better decisions. It helps in understanding what is where. Spatial-temporal GIS, or 4D GIS, has become necessary in areas where GIS is needed for predicting dimensions across time. GIS is increasingly needed with a real-time platform that offers not just monitoring of events but can take input and predict what could happen as a type of forecasting tool. Figure 14 shows the GIS visualization of CV-19 in different states of India. The color red in Maharashtra indicates that the numbers of reported cases (confirmed cases) are more in number and is known as red zone. The less brightness of red indicates the little less than Maharashtra state. The green color indicates normal with less number of CV-19 cases. The color blue refers to safe zone where no or single digit confirmed cases are reported. In red and green zone states, one find lockdown implemented to overcome the increase in the number of cases. If no lockdown was implemented in India, one would find more number of cases as in aboard along with death cases.

7 Conclusions

In a comparative study of two predictive models of TS, ARIMA, and KALMAN filter in this chapter predicts day-wise cases in COVID-19 of India. Both the models are used on the stationary TS datasets. But, it was found that ARIMA model gave better results over KALMAN filter model for the COVID-19 dataset.

Table 15 Predicted values using ARIMA and KALMAN filter from April 27, 2020 to May 20, 2020

Date	ARIMA predicted values	KALMAN filter projected values	Number of cases reported	Remarks
March 14, 2020			84	
March 15, 2020	907		110	
March 16, 2020	166		114	
March 17, 2020	149		137	
March 18, 2020	190		151	
March 19, 2020	195		173	
March 20, 2020	225		223	
March 21, 2020	302		315	
March 22, 2020	434		360	
March 23, 2020	434		468	
March 24, 2020	603		519	
March 25, 2020	599		606	
March 26, 2020	721		694	
March 27, 2020	810		834	
March 28, 2020	1,000		918	
March 29, 2020	1,030		1,024	
March 30, 2020	1,157		1,251	
March 31, 2020	1,500		1,397	
April 1, 2020	1,568		1,834	
April 2, 2020	2,286		2,069	
April 3, 2020	2,326		2,547	
April 4, 2020	3,038		3,072	
April 5, 2020	3,608		3,577	
April 6, 2020	4,094		4,281	
April 7, 2020	4,989		4,789	
April 8, 2020	5,309	1,091	5,274	
April 9, 2020	5,772	2,483	5,865	
April 10, 2020	6,465	3,875	6,761	
April 11, 2020	7,654	5,268	7,529	
April 12, 2020	8,299	6,660	8,447	
April 13, 2020	9,361	8,053	9,352	
April 14, 2020	10,254	9,445	10,815	
April 15, 2020	12,254	10,838	11,933	
April 16, 2020	13,040	12,230	12,759	
April 17, 2020	13,585	13,622	13,835	

(continued)

Table 15 (continued)

Date	ARIMA predicted values	KALMAN filter projected values	Number of cases reported	Remarks
April 18, 2020	14,902	15,015	14,792	
April 19, 2020	15,744	16,407	16,116	
April 20, 2020	17,421	17,800	17,565	
April 21, 2020	18,991	19,192	18,985	
April 22, 2020	20,383	20,585	20,471	
April 23, 2020	21,932	21,977	21,700	
April 24, 2020	22,914	23,369	23,452	
April 25, 2020	25,169	24,762	24,942	
April 26, 2020	26,407	26,154	26,496	
April 27, 2020	28,023	27,547	28,380	
April 28, 2020	30,224	28,939	29,974	
April 29, 2020	31,539	30,332	31,787	
April 30, 2020	33,563	31,724	33,610	
May 1, 2020	35,303	33,116		
May 2, 2020	37,009	34,509		
May 3, 2020	38,681	35,901		
May 4, 2020	40,322	37,294		
May 5, 2020	41,932	38,686		
May 6, 2020	43,512	40,079		
May 7, 2020	45,064	41,471		
May 8, 2020	46,589	42,863		
May 9, 2020	48,087	44,256		
May 10, 2020	49,560	45,648		
May 11, 2020	51,008	47,041		
May 12, 2020	52,433	48,433		
May 13, 2020	53,836	49,825		
May 14, 2020	55,217	51,218		
May 15, 2020	56,576	52,610		
May 16, 2020	57,916	54,003		
May 17, 2020	59,236	55,395		
May 18, 2020	60,538	56,788		
May 19, 2020	61,822	58,180		
May 20, 2020	63,088	59,572		

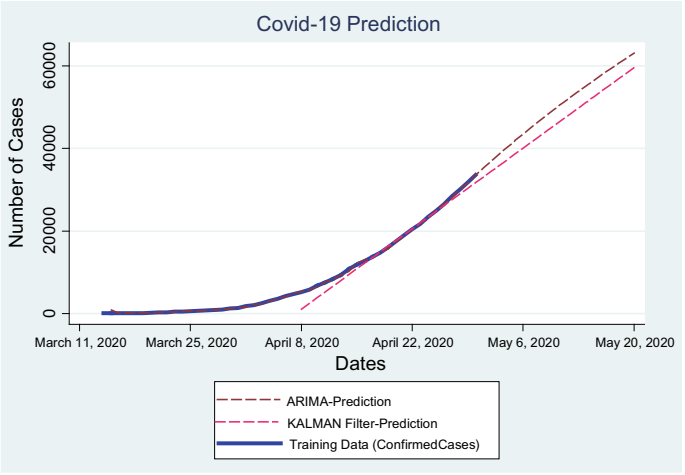

Fig. 13 Two models prediction graph of COVID-19 India cases

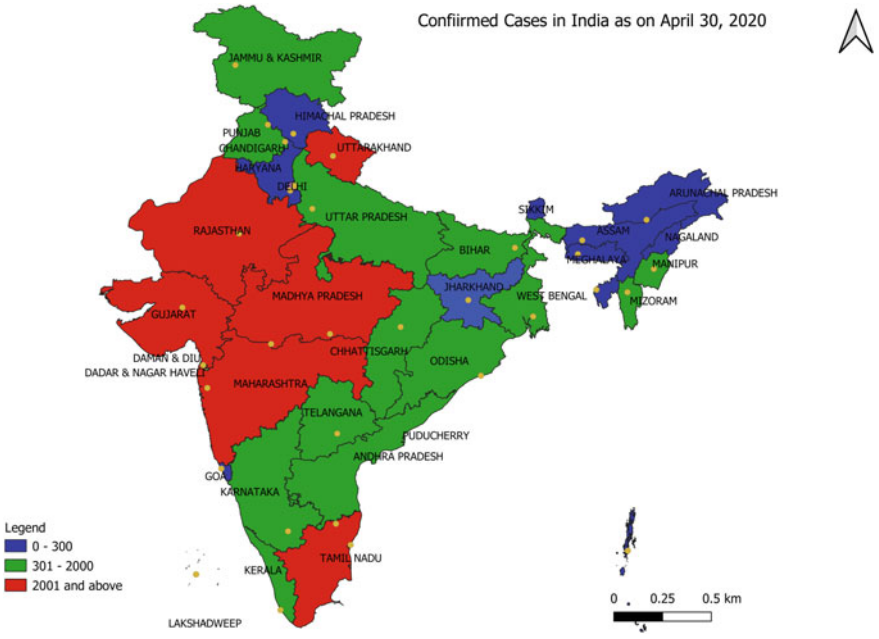

Fig. 14 GIS visualization on COVID-19 in different states on India

Conflict of Interest The authors declare that they have no known conflict of interests.

References

1. Shumway, R. H., & Stoffer, D. S. (2006) Characteristics of Time Series. In: Time Series Analysis and Its Applications. Springer Texts in Statistics. Springer, New York, NY
2. Tealab, A.: Time series forecasting using artificial neural networks methodologies: a systematic review. Faculty of Computers and Information Technology, Future University in Egypt. Elsevier B. V. Future Computing and Informatics Journal. **3**, 334e340 (2018). https://doi.org/10.1016/j.fcij.2018.10.003. 2314-7288/
3. Kuhn, M., Johnson, K.: Measuring performance in classification models. Appl. Predictive Model. doi 10.1007/978-1-4614-6849-3 11,© Springer Science and Business Media, New York, 2013, p. 247
4. Pavlyshenko, B.M.: Machine-learning models for sales timeseries forecasting. Data. **4**, 15 (2019). https://doi.org/10.3390/data4010015 www.mdpi.com/journal/data
5. Elmousalami, H.H., Hassanien, A.E.: Day level forecasting for coronavirus disease (COVID-19) spread: analysis, modeling and recommendations. (2020). arXiv:2003.07778
6. Siedner, M.J., Harling, G., Reynolds, Z., Gilbert, R., Venkataramani, A.S., Tsai, A.C.: Social distancing to slow the U.S. COVID-19 epidemic: interrupted time-series analysis. (2020). medRxiv preprint doi:https://doi.org/10.1101/2020.04.03.20052373
7. Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., Ciccozzi, M.: Application of the ARIMA model on the COVID-2019 epidemic dataset, 2352–3409/© 2020 The Authors. Published by Elsevier Inc. (2019). https://doi.org/10.1016/j.dib.2020.105340
8. Cleophas, T.J., Zwinderman, A.H.: Autoregressive models for longitudinal data (120 mean monthly population records). Machine Learning in Medicine- a Complete Overview. (2015). ISBN 978-3-319-15194-6, https://doi.org/10.1007/978-3-319-15195-3
9. Noureen, S., Atique, S., Roy, V., Bayne, S.: Analysis and application of seasonal ARIMA model in energy demand forecasting: a case study of small scale agricultural load. (2019). 978-1-7281-2788-0/19/$31.00 ©2019 IEEE
10. Contreras, J., Espínola, R., Nogales, F.J., Conejo, A.J.: ARIMA models to predict next-day electricity prices. IEEE Trans. Power Syst. **18**(3), (Aug 2003)
11. Martin, R.D., Victor J.Y.: Influence functionals for time series. Ann. Stat. **14**(3), 781–818 (1986). Accessed September 8, 2020. http://www.jstor.org/stable/3035535
12. Mikusheva, A.: Filtering. State space models. Kalman Filter.course materials for 14.384 TimeSeries Analysis. MITOpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology. (2007)
13. Rhudy, M.B, Salguero, R.A., Holappa, K.: A kalman filtering tutorial for undergraduate students. Int. J. Comput. Sci. Eng. Surv. (IJCSES). **8**(1), (Feb 2017). https://doi.org/10.5121/ijcses.2017.8101
14. Laaraiedh, M.: Implementation of Kalman filter with Python language, (2012). https://arxiv.org/pdf/1204.0375
15. Rankin, J.M.: Kalman filtering approach to market price forecasting. Retrospective Theses and Dissertations. 8291, (1986). https://lib.dr.iastate.edu/rtd/8291
16. Malleswari, B.L. MuraliKrishna, I.V., Lalkishore, K., Seetha, M., Hegde, N.P.: The role of kalman filter in the modelling of GPS errors. J. Theor. Appl. Inform. Technol. (2009). www.jatit.org
17. Kleeman, L.: Understanding and applying kalman filtering, https://www.cs.cmu.edu/
18. Ramirez-Amaro, K., Chimal-Eguia, J.C.: Machine learning tools to time series forecasting. IEEE Xplore. (2007). https://doi.org/10.1109/micai.2007.42