A Perspective Study of COVID-19 - A Case from Indian Subcontinent

K. Kumaraswamy

***Abstract* — The novel coronavirus disease (COVID - 19) now has a large global impact. The relationship between variables such as confirmed number of positive cases and number of deceased people is investigated in this paper. The purpose of a time-varying correlational study is to determine how one variable influences the other. The most vulnerable states on the variables studied were identified using statistical process control methods. To discover the association shift between the variables, the Exponential Moving Average Method covariance was performed. Time series models such as Autoregressive Integrated Moving Average (ARIMA) and GARCH models are examined in an attempt to estimate on future proven cases models. The performance measures of Mean Absolute per Error and Akaike Information Criteria are used to compare the predicted time series models.**

***Keywords* —ARIMA, correlation coefficient, COVID-19, GARCH, statistical process control.**

I. Introduction

COVID - 19 is a novel coronavirus disease that originated in Wuhan, China, in mid-December 2019 and has since spread around the world. The world economy was harmed by the coronavirus. Interventions are being made in order to stabilize the economies and existence of the countries. Millions of people fell into extreme poverty as a result of this epidemic, which had a negative impact on agriculture and other parts of daily life. The coronavirus outbreak is wreaking havoc in India. With a population of over 130 million people and a total size of 32.87 lakhs square kilometers, India's land is characterized by a significant deal of diversity in its physical qualities. India has the larger population density in the world. The first case of COVID - 19 was discovered in Kerala in the last week of January 2020. Over a short period of time, the number of instances has risen to an alarming level. The government took the initiative to confront the consequences of COVID-19 by establishing a rigorous lockdown across the country. With an unified spirit, India introduced the India's indigenous Covaxin vaccine developed by Bharat Biotech and granted the green light to Oxford - Astra Zeneca (Covishield) under manufacturing license to Serum Institute of India.

On January 16, 2021, India launched COVID - 19 vaccination drives for all health care and frontline workers who were qualified to get the vaccine. On March 1, 2021, anyone over 60 years old and over 45 years old with co-morbidities became eligible for the following phase, and on April 1, 2021, this was extended to everyone over 45 years old. On May 1, 2021, the immunization campaign against COVID - 19 was extended to the full adult population over the age of 18. On October 21, 2021, a total of one billion vaccination doses will have been provided to eligible beneficiaries in India over a ten-month period.

Lockdown 1.0 was in effect from March 15 to April 14, Lockdown 2.0 was in effect from April 15 to May 3, Lockdown 3.0 was in effect from May 4 to May 17, and Lockdown 4.0 was in effect from May 18 to May 31 of 2020. India has seen a negative impact on its economy, agriculture, education, work, and daily life activities as a result of tight adherence to the imposed Lockdown. The Indian government intervened and a number of actions were performed to stabilize the performance of all allied sectors in order to strengthen the economy and other sectors. Unlock 1.0 sessions were held from June 1 to June 30, Unlock 2.0 sessions were held from July 1 to July 31, Unlock 3.0 sessions were held from August 1 to August 31, Unlock 4.0 between September 1 and September 30, and Unlock 5.0 between October 1 and October 15, 2020. With the country's unlocking process in the second week of September, the number of confirmed positive cases skyrocketed.

The focus of this research is on the structural behavior of dependence between the variables under investigation. To grasp the relationship between the variables within the study period, statistical approaches are used. These tools, including as the correlation coefficient, statistical process control, and time series models, are widely used across all disciplines. The paper is divided into four sections: section 1 is a brief introduction to COVID-19 and the Indian subcontinent, section 2 is a description of the various approaches utilized for the study variables, section 3 is a summary of the findings, and section 4 is the conclusion.

II. Methodology

Some statistical tools are used to comprehend and identify the pattern of the variables under examination. The following is a list of the various approaches that were employed in is paper.

Published on February 18, 2022.

K. Kumaraswamy, Kaloji Narayana Rao University of Health Sciences, Warangal, India.
(corresponding e-mail: kumaraswami.kandukuri@gmail.com)

DOI: http://dx.doi.org/10.24018/ejmath.2022.3.1.91

A. Correlation Coefficient

The Pearson's Correlation coefficient was used to quantify the monotonic connection between these two study variables. The correlation coefficient is one of the easiest methods for calculating the association between any two variables under consideration. From -1.00 to +1.00, the greatest notable strength or weakness of monotonic association can be found. Weak positive correlation (0, 0.3), moderate positive correlation (0.3, 0.7), and strong positive correlation (0.3, 0.7) are the threshold levels for different types of segregation. Similarly, weak negative correlation (-0.3, 0), moderate negative correlation (-0.7, -0.3), and strong negative correlation (-1.0, -0.7). The Pearson's Correlation Coefficient *r* was obtained using the relation

$$r = \frac{Cov\,(x,y)}{\sigma_x.\sigma_y} = \frac{E(xy) - E(x)*E(y)}{\sigma_x.\sigma_y} = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)}.\sqrt{\left(n\sum y^2 - \left(\sum y\right)^2\right)}} \tag{1}$$

where n = number of pairs, $\sum x$ =sum of x observations, $\sum y$ =sum of y observations, $\sum xy$ =sum of product of observations, $\sum x^2$ =sum of squared x observations and $\sum y^2$ =sum of squared y observations.

B. Time Varying Correlation Coefficient

When the data behaves volatile, this method of study is employed to detect tiny shifts. To find time varying shifts between the variables under examination, the Exponential Weighted Moving Average correlation approach is of significant relevance. For each observation at each time period, EWMA estimates the volatile covariance. This method use the Maximum Likelihood methodology to estimate the parameter with a weighted mean sequence by assigning higher weights to recent data, which diminishes geometrically with time. *EWMA V$_n$* is defined as

$$EWMA\,V_n = \lambda * Cov\,V_{n-1} + (1-\lambda) * r_n \tag{2}$$

where, λ is a weighted constant such that $\lambda \in [0, 1]$.

C. Statistical Process Control

For a certain length of time, we created Shewartz control charts to identify the most vulnerable states under investigation. The control chart comprises three lines that are used to determine whether or not the process is under control. The preceding's mean behavior is represented by the central line (CL), while action lines are 3 times standard deviations from the central line and warning limits are 2 times standard deviations from the central line [5].

The control limits are presented as per Individual chart for better understanding (I - Chart)

$$s\tan dard\ \ deviation\ \ (\sigma) = \frac{\sqrt{\sum_{i=1}^{N}(x_i-\mu)^2}}{N} \tag{3}$$

where, x_i = the individual sample value, μ = sample mean and N = size of the sample, Upper Control (Action) limit = $\mu + 3*\sigma$, Lower Control Limit (Action) limit = $\mu - 3*\sigma$, Lower Warning limit = $\mu + 2*\sigma$, Upper Warning limit = $\mu - 2*\sigma$ and Central limit = μ .While plotting the distinct results on the I-Chart, the "out of control" situations are examined.

D. Time Series Models

A time series is a representation of a stochastic process, which is a collection of random variables that are ordered across time. Mean Absolute %age Error (MAPE) is an efficiency metric that is used to measure prediction performance.

1) ARIMA Model

Box and Jenkins (1976) established the ARIMA (*p, d, q*) time series model $\{y_t; t = 0,1,2,...\}$ is still very popular and is given as:

$$\varphi(B)\Delta^d y_t = \theta(B)\varepsilon_t \tag{4}$$

where y_t , ε_t and *B* represents number of confirmed cases, random error terms and backward shift operator at time *t* respectively. *The* $\varphi(B)$ and $\theta(B)$ are order of *p* and *q* defined as:

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \\ \theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \end{aligned} \tag{5}$$

where $\phi_1, \phi_2, \phi_3, \ldots, \phi_p$ are the autoregressive coefficients and $\theta_1, \theta_2, \theta_3, \ldots \theta_q$ are described as moving average coefficients that attempt to anticipate a system's output based on prior results.

2) GARCH Model

The ARIMA (*p, d, q*) model does not capture the heteroskedastic effects of a time series process, which are typically recognized as excessive kurtosis or clustering of volatilities, as well as the leverage effect. The variance equation for the GARCH (*p, q*) model is as follows:

$$y_t = E_{t-1}(y_t) + \varepsilon_t \tag{6}$$

$$\varepsilon_t = Z_t \sigma_t \tag{7}$$

$$Z_t \sim \psi_T(0,1) \tag{8}$$

where $E_{t-1}(.)$ reflects expectation conditional on knowledge accessible at time *t*-1, and $Z_t$is a sequence of i.i.d. random variables with mean zero and unit variance.

III. Findings

A thorough investigation was conducted. In the first week of November 2021, India registered almost 3.45 crore confirmed positive cases, with 4.60 lakh people dead. As a whole, the recovered probability is 0.9823, whereas the deceased probability is 0.0134. The growth rate of confirmed cases and the number of deceased persons are depicted in Fig.1a, indicating that there is a strong direct proportional relationship, i.e., as the confirmed cases rise, the deceased ratio grows, and vice versa. The confirmed cases and deceased cases have both climbed significantly in the weeks of April 6, 2020, with a 286.914% increase and April 13, 2020, with a 276.92% increase, respectively. Both variables under investigation have a dynamic relationship. The variables have a high positive dependence relationship, with a Pearson's correlation coefficient of 0.8635 between them. The time changing correlations between the variables were evaluated using the EWMA approach in Fig.1b, and the variables have strong positive associations ranging from 0.891 to +1.00.

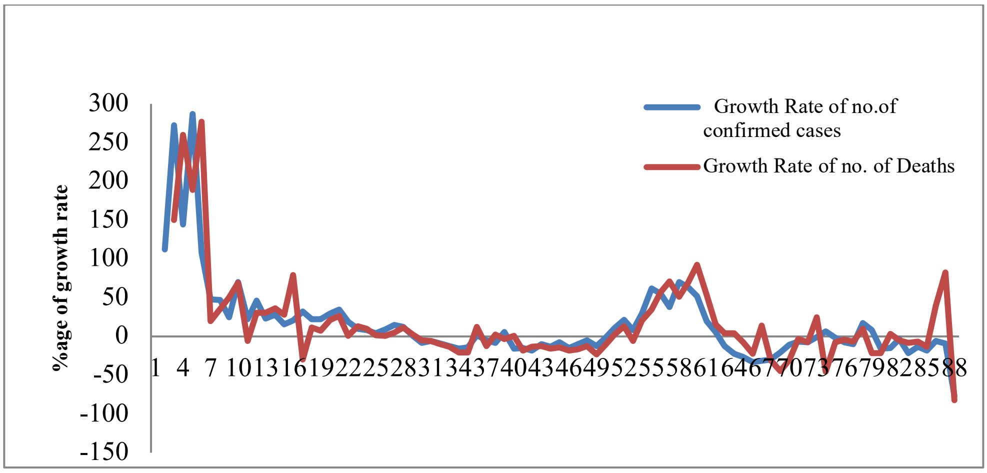

Fig.1(a). Growth rate of confirmed cases and deceased people.

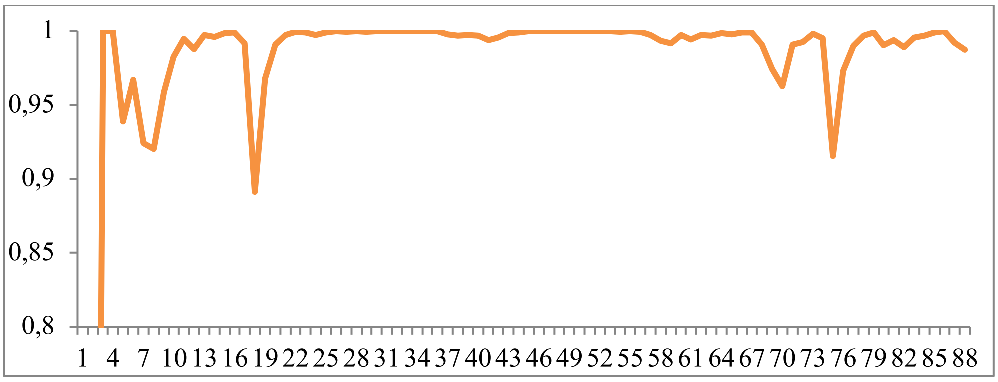

Fig. 1(b). Time varying correlation between the confirmed cases and deceased people.

The statistical process control technique, also known as the Control chart, is used to research and identify performance indicators (variables under investigation) of states that pose a greater risk to public health. Fig. 2a shows that the states with the highest number of confirmed cases are listed in ascending order, with Maharastra (19.2676 %), Kerala (14.5491 %), Karnataka (8.7065 %), Tamil Nadu (7.8829 %), and Andhra Pradesh (6.022 %) accounting for more than half (56.43 %) of the total number of confirmed cases in India. Fig.2b shows that the death rate is highest in Maharastra (30.5181 %), Karnataka (8.2837 %), Tamilnadu (7.8697 %), Kerala (7.118 %), Delhi (5.456 %), and so on, while it is lowest in Punjab (2.75 %), Nagaland (2.154 %), Uttarakhand (2.152 %), Maharastra (2.12 %), and Goa (2.12 %) (1.888 %). Only one UT, Dadra and Nagar Haveli, and Daman Diu have so far registered the lowest deceased ratio compared to all 36 states / UT's.

The weekly confirmed cases are modeled and predicted using time series forecasting techniques. There are 87 weeks of data used, with 95 % (83 weeks) being used to train the model and the remaining 4 weeks being used for testing. The ARIMA (3, 2, 3) model was chosen based on the AIC values to forecast the mean behavior pattern of confirmed cases after a series of models were run. The residuals of the ARIMA (3, 2, 3) model are volatile, hence a couple of GARCH models were performed to overcome this flaw. The AIC value was used to select the ARIMA - GARCH (1, 1) model, which also performed well on test data. The performance measure statistics, which are shown in Table II(B), confirm this.

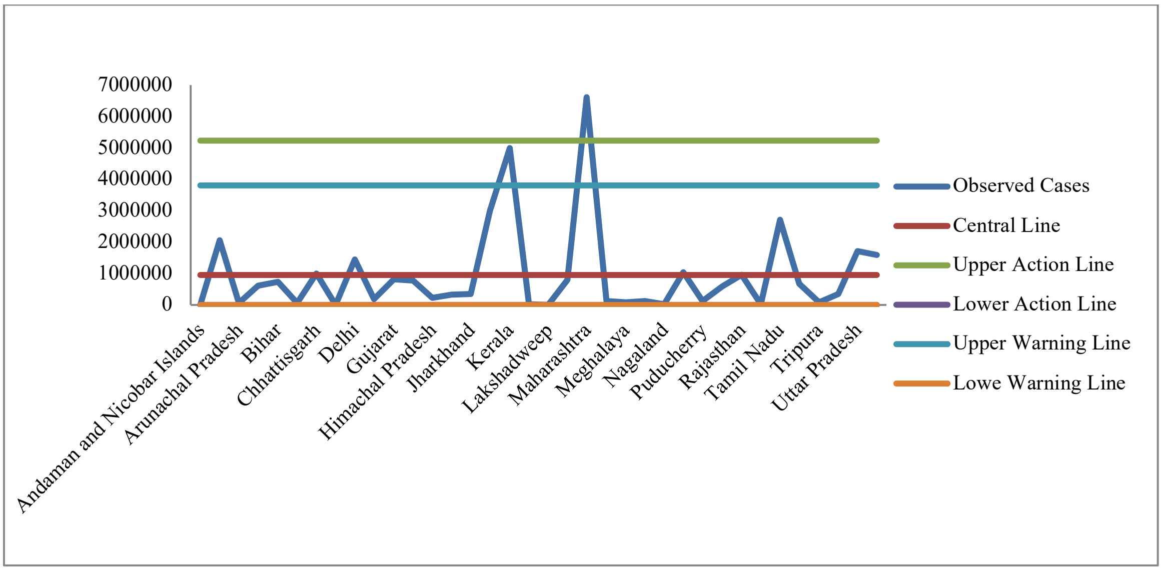

Fig.2(a). Control chart for confirmed cases.

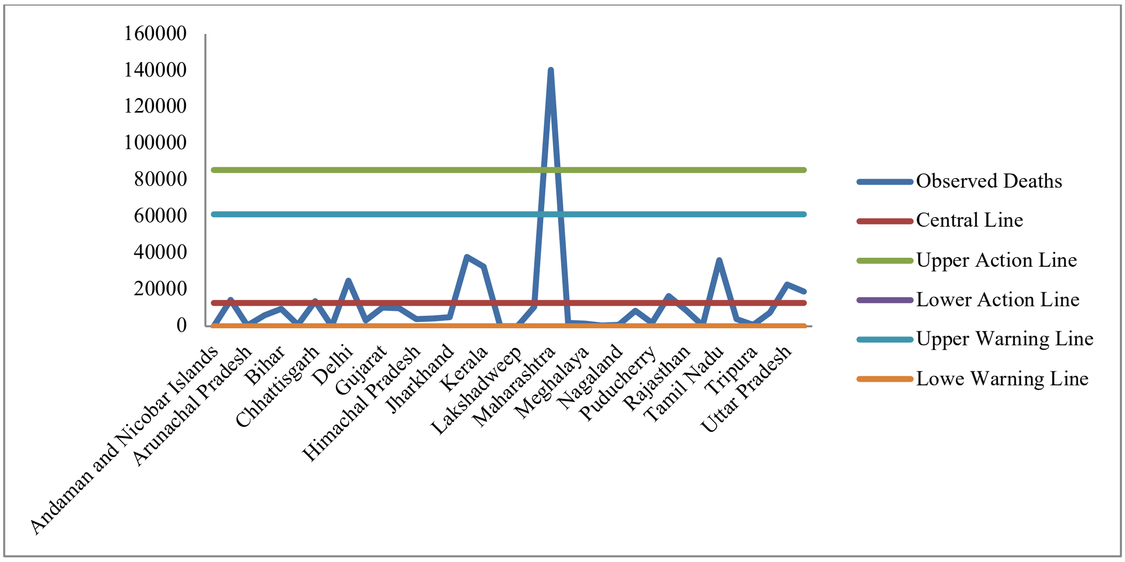

Fig. 2(b). Control chart for number of deceased people.

TABLE I: COVID-19 STATE WISE STATUS AS ON 05 NOVEMBER 2021 [4]

S.No.	State/UT	Confirm Cases	Active	Discharge	Deaths	Active Ratio	Discharge Ratio	Death Ratio
1	Andaman and Nicobar Islands	7659	10	7520	129	0.130565	98.18514	1.684293
2	Andhra Pradesh	2067558	3830	2049338	14390	0.185243	99.11876	0.695990
3	Arunachal Pradesh	55174	68	54826	280	0.123246	99.36926	0.507485
4	Assam	611656	3430	602207	6019	0.560773	98.45517	0.984049
5	Bihar	726120	44	716415	9661	0.00606	98.66344	1.330496
6	Chandigarh	65357	32	64505	820	0.048962	98.69639	1.254647
7	Chhattisgarh	1006129	279	992267	13583	0.02773	98.62224	1.350025
8	Dadra and Nagar Haveli and Daman and Diu	10682	1	10677	4	0.009362	99.95319	0.037446
9	Delhi	1440003	303	1414609	25091	0.021042	98.23653	1.742426
10	Goa	178245	337	174542	3366	0.189066	97.92252	1.888412
11	Gujarat	826680	220	816370	10090	0.026612	98.75284	1.220544
12	Haryana	771287	110	761127	10050	0.014262	98.68272	1.303016
13	Himachal Pradesh	224619	1646	219205	3768	0.732796	97.58969	1.677507
14	Jammu and Kashmir	332651	981	327232	4438	0.294904	98.37096	1.334131
15	Jharkhand	348828	127	343563	5138	0.036408	98.49066	1.472932
16	Karnataka	2989275	8296	2942884	38095	0.277525	98.44808	1.274389
17	Kerala	4995255	75171	4887350	32734	1.504848	97.83985	0.655301
18	Ladakh	21005	98	20698	209	0.466556	98.53844	0.995001
19	Lakshadweep	10365	0	10314	51	0	99.50795	0.492040
20	Madhya Pradesh	792888	114	782250	10524	0.014378	98.65832	1.327299
21	Maharashtra	6615299	18691	6456263	140345	0.282542	97.59593	2.121521
22	Manipur	123957	704	121326	1927	0.567939	97.87749	1.554571
23	Meghalaya	83763	391	81916	1456	0.466793	97.79496	1.738237
24	Mizoram	124026	6141	117445	440	4.951381	94.69385	0.354764
25	Nagaland	31894	195	31012	687	0.6114	97.23459	2.154010
26	Odisha	1042773	3537	1030889	8347	0.339192	98.86034	0.800461
27	Puducherry	128134	350	125924	1860	0.273152	98.27524	1.451605
28	Punjab	602466	240	585664	16562	0.039836	97.21112	2.749034
29	Rajasthan	954450	48	945448	8954	0.005029	99.05683	0.938131
30	Sikkim	32019	174	31447	398	0.543427	98.21356	1.243012
31	Tamil Nadu	2706493	10895	2659407	36191	0.40255	98.26025	1.337191
32	Telangana	672052	3879	664212	3961	0.577187	98.83342	0.589388
33	Tripura	84557	146	83595	816	0.172665	98.86230	0.965029
34	Uttarakhand	343924	146	336377	7401	0.042451	97.80562	2.151928
35	Uttar Pradesh	1710181	95	1687184	22902	0.005555	98.65528	1.339156
36	West Bengal	1596332	8193	1568951	19188	0.513239	98.28475	1.202005

TABLE II (A): ARIMA - GARCH FORECASTS

Week	84	85	86	87	88	89	90	91	92	93
Obs.	139572	114244	107749	97832	82236	81771	73106	62110	60732	57255
Arima (3, 2, 3)Predict	102621	12319	58869	100872	112862	104294	86508	68896	56830	51777
Arima - Garch (1, 1) Predict	130222	108930	98866	98340	104744	115559	128772	142956	157198	170982

TABLE II (B): PERFORMANCE STATISTICS OF THE MODELS

Model	MAPE
ARIMA (3, 2, 3)	0.2742
ARIMA - GARCH (1,1)	0.6526

IV. Conclusion

There is a limitation to the study because it was performed over such a short period of time. Between the variables, a causative relationship characteristic is extracted. 11 states had confirmed cases that were higher than the national average, while 13 states had died people that were higher than the national average. The most densely populated states/UTs are more likely to see an increase in instances. To alert states, statistical process control charts are utilized to determine what actions should be made to reduce the number of confirmed cases and deaths. Understanding the behavioral trend of the variables is aided by time varying correlations. To anticipate future extreme confirmed cases, a credible time series model is built.

Acknowledgement

The author is grateful to his Ph.D. supervisor for his endless encouragement to focus on research.

Funding

Not applicable.

Conflict of Interest

Not applicable.

References

[1] Douglas C. Montgomery. *Introduction to Statistical Quality Control*, 6th ed., John Wiley & Sons, 2008.
[2] Kumaraswamy K, Jayalakshmi C. Literacy Level in Andhra Pradesh and Telangana States – A Statistical Study. *The International Journal of Engineering and Science.* 2017; 6(6): 70–77.
[3] *India: WHOCoronavirus Disease (COVID-19) Dashboard* [Internet]. 2021. [updated 2021 Nov 1]. Available from: https://Covid19.who.int/region/searo/country/in.
[4] MoHFWGoI, *Covid-19: Statewise Status* [Internet]. 2021. [updated 2021 November 5, cited 2021 November 5]. Available from: https://my.gov.in/Covid-19.
[5] Oakland, JS. *Statistical Process Control*. 2007.

K. Kumaraswamy completed his PhD in Statistics from Osmania University, Hyderabad, India in 2021. His major area of interests study are Stochastic Process, Distribution Theory and Statistical Modelling.

He previously served as Mandal Planning and Statistical Officer / Assistant Statistical Officer in Directorate of Economics and Statistics, Telangana State and currently working as Statistical Officer in Kaloji Narayana Rao University of Health Scicences, Warangal, Telangana State, India.

Dr. Kumaraswamy is a life member of the Indian Society for Probability and Statistics as well as the Society for Development of Statistics, Hyderabad, and published about 5 research papers in Interantional journals. He was a fellow of UGC - BSR (RFSMS).