

STATISTICAL ANALYSIS OF PREVALENCE AND MORTALITY OF CORONA VIRUS (COVID19) IN INDIA

Submitted in partial fulfilment of the requirements for the award of
Bachelor of Science degree in Mathematics (B.Sc. Mathematics)

By

B. KAMALESH (38060006)



DEPARTMENT OF MATHEMATICS
SCHOOL OF SCIENCE AND TECHNOLOGY

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

JAN 2021 – APRIL 2021



SATHYABAMA



INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with “A” grade by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600 119

DEPARTMENT OF MATHEMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of B. KAMALESH (38060006) who carried out the project entitled “STATISTICAL ANALYSIS OF PREVALENCE AND MORTALITY OF CORONA VIRUS (COVID19) IN INDIA” under our supervision from Jan 2021 to Apr 2021.

Internal Guide

External Guide

(Dr. M.NIRMALA)

Head of the Department

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I B. KAMALESH hereby declare that the Project Report entitled "STATISTICAL ANALYSIS OF PREVALENCE AND MORTALITY OF CORONA VIRUS (COVID 19) IN INDIA" done by me under the guidance of Dr. E. PRIYADARSHINI (Internal) and _____ (External) at Sathyabama Institute of Science and Technology is submitted in partial fulfilment of the requirements for the award of Bachelor of Science degree in Mathematics.

DATE:

PLACE: SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of SATHYABAMA for their kind encouragement in doing this project and for completing it successfully. I am very grateful to them.

I convey my thanks to Dr. M. Nirmala, Head of the Department, Dept. Of Mathematics for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide Dr.E.PRIYADARSHINI for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Mathematics who were helpful in many ways for the completion of the project.

ABSTRACT

The novel corona virus (covid19) is a severe respiratory infection that officially occurred in Wuhan, China in December 2019. In late February, the disease began to spread quickly across the world, causing serious health, social, and economic emergencies in this paper the COVID data for India was collected from WHO website from Apr 2020 to Jan 2021 and Statistical analysis was conducted for the following method. 1.Run test, 2. Mann Kendell test, 3. Moving average, 4. Rescaled range analysis and the results were tabulated.Also the error measures for moving average was computed for the predicted values. Further more precise forecasting may be provided by continuously updating these data, adding interventions and other real aspects, and applying the model to other countries.

KEY WORDS:

Run test, Mann Kendell test, Moving average, Rescaled range analysis, trend, persistence

TITLE OF CONTENT

Contents Number	Contents	Pg. No
1	Introduction	2
1.1	Definition (COVID 19)	2
1.2.1	Virus classification	3
1.2.2	Etymology	3
1.2.3	History	4
1.3	Microbiology	6
1.4	Classification	9
1.5	Origin	9
1.6	Infection in humans	11
1.6.1	Infection in animals	12
1.6.2	Symptoms	12
1.6.3	Mathematical and statistical methods for analysis	12
1.6.4	Types of analysis	12
1.6.5	Time series analysis	12
1.6.6	Uses of time series analysis	13
1.6.7	Components of time series analysis	13
2	Literature survey	15
3	Aim and scope	20
4.	Methodology	22
4.1.	Run test	22
4.2	Mann Kendall test	23
4.3	Moving average	24
4.4	Range scale analysis	25

5	Result and discussion	28
5.1	Run test	28
5.2	Mann Kendall test	29
5.3	Moving average	30
5.4	Rescaled range analysis	33
6	Results and Summary	38
7	References	39

Table Number	Table name	Pg. No
5.1 -	One-sample runs test / Two-tailed test Recovered	28
5.2 -	One-sample runs test / Two-tailed test Confirmed	28
5.3	One-sample runs test / Two-tailed test - Death	29
5.4	Mann-Kendall trend test/two –tailed test - Recovered	29
5.5	Mann-Kendall trend test/two –tailed test -Confirmed	29
5.6	Mann-Kendall trend test/two –tailed test-Death	30
5.7	Error Measures of Moving Average	33

Figure Number	FIGURE	Pg.no
1.1	Illustration of a coronavirus	2
1.2	Transmission electron micrograph of organ cultured Coronavirus OC43	5
1.3	Transmission electron micrograph of organ cultured Coronavirus CO43	7
1.4	After binding of the ACE2 receptor, SARS-CoV spike is Activated and cleaved at the S1 S2 level	8
1.5	Seasonal distribution of HCoV-NL63 in Germany shows A preferential detection from November to March	11
5.1	Moving Average of Recoveries	30
5.2	Residual plot of Recoveries	31
5.3	Moving Average of Confirmed Cases	31
5.4	Residual plot of Confirmed cases	32
5.5	Moving Average of Death Cases	32
5.6	Residual plot of Death cases	33
5.7	R/S Chart for confirmed cases	33
5.8	V Chart for confirmed cases	34
5.9	R/S Chart for recovered cases	34
5.10	V Chart for recovery cases	35
5.11	R/S Chart for death cases	35
5.12	V Chart for death cases	35

CHAPTER 1

1. INTRODUCTION:

1.1. DEFINITION:

Coronaviruses are a group of related RNA viruses that cause diseases in mammals and birds. In humans and birds, they cause respiratory tract infections that can range from mild to lethal. Mild illnesses in humans include some case of the common cold, while more lethal varieties can cause **SARS**, **MERS**, and **COVID-19**. In cows and pigs, they cause diarrhea, while in mice they cause hepatitis and encephalomyelitis.

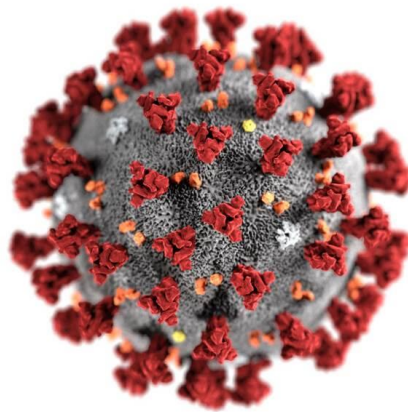


Figure 1.1

Illustration of a coronavirus

- Red: **spike proteins (S)**
- Grey: **lipid bilayer envelope**
- Yellow: **envelope proteins (E)**
- Orange: **membrane proteins (M)**

1.2. VIRUS CLASSIFICATION

(Unranked):	Virus
Realm:	Riboviria
Kingdom:	Orthornavirae
Phylum:	Pisuviricota
Class:	Pisoniviricetes
Order:	Nidovirales
Family:	Coronaviridae
Subfamily:	Orthocoronavirinae

Genera

- Alphacoronavirus
- Betacoronavirus
- Gammacoronavirus
- Deltacoronavirus

Coronaviruses constitute the subfamily **Orthocoronavirinae**, in the family Coronaviridae, order Nidovirales, and realm Riboviria. They are enveloped viruses with a positive-sense single-stranded RNA genome and a nucleocapsid of helical symmetry. The genome size of coronaviruses ranges from approximately 26 to 32 kilobases, one of the largest among RNA viruses. They have characteristic club-shaped spikes that project from their surface, which in electron micrographs create an image reminiscent of the solar corona, from which their name derives.

1.2.1. Etymology:

The name "coronavirus" is derived from Latin corona, meaning "crown" or "wreath", itself a borrowing from Greek κορώνη korónē, "garland, and wreath". The name was coined by June Almeida and David Tyrrell who first observed and studied human coronaviruses.¹ The word was first used in print in 1968 by an

informal group of virologists in the journal *Nature* to designate the new family of viruses. The name refers to the characteristic appearance of virions (the infective form of the virus) by electron microscopy, which have a fringe of large, bulbous surface projections creating an image reminiscent of the solar corona or halo. This morphology is created by the viral spike peplomers, which are proteins on the surface of the virus.

The scientific name *Coronavirus* was accepted as a genus name by the International Committee for the Nomenclature of Viruses (later renamed International Committee on Taxonomy of Viruses) in 1971. As the number of new species increased, the genus was split into four genera, namely *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, and *Gammacoronavirus* in 2009.¹ The common name coronavirus is used to refer to any member of the subfamily *Orthocoronavirinae*.¹ As of 2020, 45 species are officially recognised.

1.2.2 History:

The earliest reports of a coronavirus infection in animals occurred in the late 1920s, when an acute respiratory infection of domesticated chickens emerged in North America. Arthur Schalk and M.C. Hawn in 1931 made the first detailed report which described a new respiratory infection of chickens in North Dakota. The infection of new-born chicks was characterized by gasping and listlessness with high mortality rates of 40–90%. Leland David Bushnell and Carl Alfred Brandy isolated the virus that caused the infection in 1933. The virus was then known as infectious bronchitis virus (IBV). Charles D. Hudson and Fred Robert Beaudette cultivated the virus for the first time in 1937. The specimen came to be known as the Beaudette strain. In the late 1940s, two more animal coronaviruses, JHM that causes brain disease (murine encephalitis) and mouse hepatitis virus (MHV) that causes hepatitis in mice were discovered. It was not realized at the time that these three different viruses were related.

Human coronaviruses were discovered in the 1960s using two different methods in the United Kingdom and the United States.¹ E.C. Kendall, Malcolm Bynoe, and David Tyrrell working at the Common Cold Unit of the British Medical Research Council collected a unique common cold virus designated B814 in 1961. The virus could not be cultivated using standard techniques which had successfully cultivated rhinoviruses, adenoviruses and other known common cold viruses. In 1965, Tyrrell and Bynoe successfully cultivated the novel virus by serially passing it through organ culture of human embryonic trachea. The new cultivating method was introduced to the lab by Bertil Hoorn. The isolated virus when intranasally inoculated into volunteers caused a cold and was inactivated by ether which indicated it had a lipid envelope. Dorothy Hamre and John Procknow at the University of Chicago isolated a novel cold from medical students in 1962. They isolated and grew the virus in kidney tissue culture, designating it 229E. The novel virus caused a cold in volunteers and, like B814, was inactivated by ether.

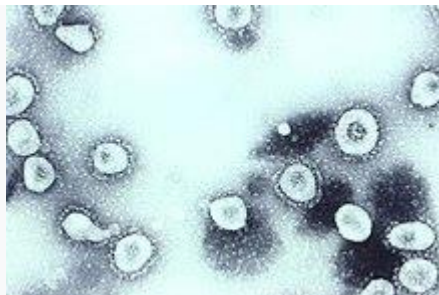


Figure 1.2. Transmission electron micrograph of organ cultured coronavirus OC43

Scottish virologist June Almeida at St. Thomas Hospital in London, collaborating with Tyrrell, compared the structures of IBV, B814 and 229E in 1967. Using electron microscopy the three viruses were shown to be morphologically related by their general shape and distinctive club-like spikes. A research group at the National Institute of Health the same year was able to isolate another member of this new group of viruses using organ culture and named one of the samples OC43 (OC for organ culture). Like B814, 229E, and IBV, the novel cold virus OC43 had distinctive club-like spikes when observed with the electron microscope.

The IBV-like novel cold viruses were soon shown to be also morphologically related to the mouse hepatitis virus. This new group of viruses were named coronaviruses after their distinctive morphological appearance. Human coronavirus

229E and human coronavirus OC43 continued to be studied in subsequent decades. The coronavirus strain B814 was lost. It is not known which present human coronavirus it was.^[41] Other human coronaviruses have since been identified, including SARS-CoV in 2003, HCoV NL63 in 2003, HCoV HKU1 in 2004, MERS-CoV in 2013, and SARS-CoV-2 in 2019. There have also been a large number of animal coronaviruses identified since the 1960s.

1.2.3. Microbiology:

Coronaviruses are large, roughly spherical particles with unique surface projections. Their size is highly variable with average diameters of 80 to 120 nm. Extreme sizes are known from 50 to 200 nm in diameter. The total molecular weight is on average 40,000 kDa. They are enclosed in an envelope embedded with a number of protein molecules. The lipid bilayer envelope, membrane proteins, and nucleocapsid protect the virus when it is outside the host cell.

The viral envelope is made up of a lipid bilayer in which the membrane (M), envelope (E) and spike (S) structural proteins are anchored. The molar ratio of E: S: M in the lipid bilayer is approximately 1:20:300. The E and M protein are the structural proteins that combined with the lipid bilayer to shape the viral envelope and maintain its size. S proteins are needed for interaction with the host cells. But human coronavirus NL63 is peculiar in that its M protein has the binding site for the host cell, and not its S protein. The diameter of the envelope is 85 nm. The envelope of the virus in electron micrographs appears as a distinct pair of electron-dense shells (shells that are relatively opaque to the electron beam used to scan the virus particle).

The M protein is the main structural protein of the envelope that provides the overall shape and is a type III membrane protein. It consists of 218 to 263 amino acid residues and forms a layer 7.8 nm thick. It has three domains, a short N-terminal ectodomain, a triple-spanning transmembrane domain, and a C-terminal endodomain. The C-terminal domain forms a matrix-like lattice that adds to the extra-thickness of the envelope. Different species can have either N- or O-linked glycans in their protein amino-terminal domain. The M protein is crucial during

the assembly, budding, envelope formation, and pathogenesis stages of the virus lifecycle.

The E proteins are minor structural proteins and highly variable in different species. There are only about 20 copies of the E protein molecule in a coronavirus particle. They are 8.4 to 12 kDa in size and are composed of 76 to 109 amino acids. They are integral proteins (i.e. embedded in the lipid layer) and have two domains namely a transmembrane domain and an extramembrane C-terminal domain. They are almost fully α -helical, with a single α -helical transmembrane domain, and form pentameric (five-molecular) ion channels in the lipid bilayer. They are responsible for virion assembly, intracellular trafficking and morphogenesis (budding).^[46]

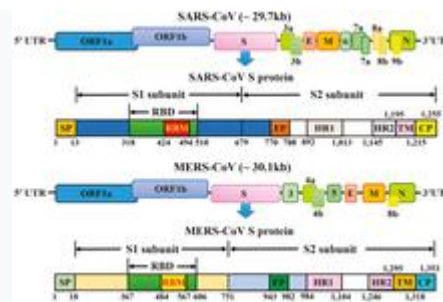


Figure1.3. Transmission electron micrograph of organ cultured coronavirus OC43

The spikes are the most distinguishing feature of coronaviruses and are responsible for the corona- or halo-like surface. On average a coronavirus particle has 74 surface spikes. Each spike is about 20 nm long and is composed of a trimer of the S protein. The S protein is in turn composed of an S1 and S2 subunit. The homotrimeric S protein is a class I fusion protein which mediates the receptor binding and membrane fusion between the virus and host cell. The S1 subunit forms the head of the spike and has the receptor-binding domain (RBD). The S2 subunit forms the stem which anchors the spike in the viral envelope and on protease activation enables fusion. The two subunits remain non covalently linked as they are exposed on the viral surface until they attach to the host cell membrane. In a functionally active state, three S1 are attached to two S2 subunits. The subunit complex is split into individual subunits when the virus binds and fuses with the host cell under the action of proteases such as cathepsin family and transmembrane protease serine 2 (TMPRSS2) of the host cell.

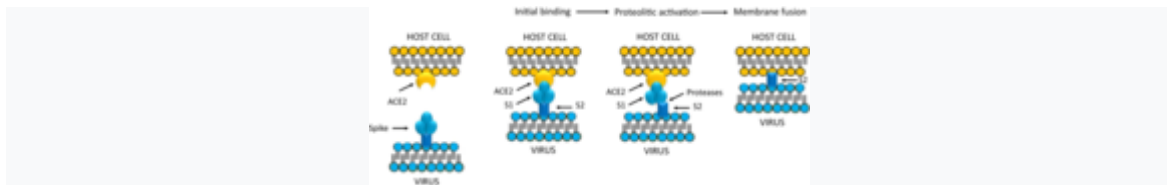


Figure 1.4. After binding of the ACE2 receptor, SARS-CoV spike is activated and cleaved at the S1/S2 level

S1 proteins are the most critical components in terms of infection. They are also the most variable components as they are responsible for host cell specificity. They possess two major domains named N-terminal domain (S1-NTD) and C-terminal domain (S1-CTD), both of which serve as the receptor-binding domains. The NTDs recognize and bind sugars on the surface of the host cell. An exception is the MHV NTD that binds to a protein receptor carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1). S1-CTDs are responsible for recognizing different protein receptors such as angiotensin-converting enzyme 2 (ACE2), aminopeptidase N (APN), and dipeptidyl peptidase 4 (DPP4).

A subset of coronaviruses (specifically the members of betacoronavirus subgroup A) also has a shorter spike-like surface protein called hemagglutinin esterase (HE). The HE proteins occur as homodimers composed of about 400 amino acid residues and are 40 to 50 kDa in size. They appear as tiny surface projections of 5 to 7 nm long embedded in between the spikes. They help in the attachment to and detachment from the host cell.

Inside the envelope, there is the nucleocapsid, which is formed from multiple copies of the nucleocapsid (N) protein, which are bound to the positive-sense single-stranded RNA genome in a continuous beads-on-a-string type conformation. N protein is a phosphoprotein of 43 to 50 kDa in size, and is divided into three conserved domains. The majority of the protein is made up of domains 1 and 2, which are typically rich in arginines and lysines. Domain 3 has a short carboxy terminal end and has a net negative charge due to excess of acidic over basic amino acid residues.

1.4. Classification:

Coronaviruses form the subfamily Orthocoronavirinae, which is one of two sub-families in the family Coronaviridae, order Nidovirales, and realm Riboviria they are divided into the four

genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus. Alphacoronaviruses and betacoronaviruses infect mammals, while gammacoronaviruses and deltacoronaviruses primarily infect birds.

- Genus: **Alphacoronavirus**; type species: Alphacoronavirus 1
 - Species: Alphacoronavirus 1 (TGEV, Feline coronavirus, Canine coronavirus), Human coronavirus 229E, Human coronavirus NL63, Miniopterus bat coronavirus 1, Miniopterus bat coronavirus HKU8, Porcine epidemic diarrhea virus, Rhinolophus bat coronavirus HKU2, Scotophilus bat coronavirus 512
- Genus **Betacoronavirus**;¹ type species: Murine coronavirus (MHV)
 - Species: Betacoronavirus 1 (Bovine Coronavirus, Human coronavirus OC43), Hedgehog coronavirus 1, Human coronavirus HKU1, Middle East respiratory syndrome-related coronavirus, Murine coronavirus, Pipistrellus bat coronavirus HKU5, Rousettus bat coronavirus HKU9, Severe acute respiratory syndrome-related coronavirus (SARS-CoV, SARS-CoV-2), Tylonycteris bat coronavirus HKU4
- Genus **Gammacoronavirus**; type species: Avian coronavirus (IBV)
 - Species: Avian coronavirus, Beluga whale coronavirus SW1
- Genus **Deltacoronavirus**; type species: Bulbul coronavirus HKU11
 - Species: Bulbul coronavirus HKU11, Porcine coronavirus HKU15

1.5. Origin:

The most recent common ancestor (MRCA) of all coronaviruses is estimated to have existed as recently as 8000 BCE, although some models place the common ancestor as far back as 55 million years or more, implying long term coevolution with bat and avian species.¹ The most recent common ancestor of the alphacoronavirus line has been placed at about 2400 BCE, of the betacoronavirus line at 3300 BCE, of

the gammacoronavirus line at 2800 BCE, and of the deltacoronavirus line at about 3000 BCE. Bats and birds, as warm-blooded flying vertebrates, are an ideal natural reservoir for the coronavirus gene pool (with bats the reservoir for alphacoronaviruses and betacoronavirus – and birds the reservoir for gammacoronaviruses and deltacoronaviruses). The large number and global range of bat and avian species that host viruses has enabled extensive evolution and dissemination of coronaviruses.

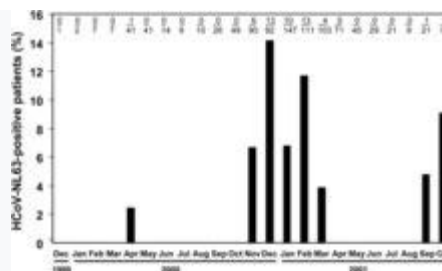
Many human coronaviruses have their origin in bats. The human coronavirus NL63 shared a common ancestor with a bat coronavirus (ARCoV.2) between 1190 and 1449 CE. The human coronavirus 229E shared a common ancestor with a bat coronavirus (GhanaGrp1 BT CoV) between 1686 and 1800 CE. More recently, alpaca coronavirus and human coronavirus 229E diverged sometime before 1960. MERS-CoV emerged in humans from bats through the intermediate host of camels. MERS-CoV, although related to several bat coronavirus species, appears to have diverged from these several centuries ago. The most closely related bat coronavirus and SARS-CoV diverged in 1986. The ancestors of SARS-CoV first infected leaf-nose bats of the genus *Hipposideridae*; subsequently, they spread to horseshoe bats in the species *Rhinolophidae*, then to Asian palm civets, and finally to humans.

Unlike other betacoronaviruses, bovine coronavirus of the species *Betacoronavirus 1* and subgenus *Embecovirus* is thought to have originated in rodents and not in bats. In the 1790s, equine coronavirus diverged from the bovine coronavirus after a cross-species jump. Later in the 1890s, human coronavirus OC43 diverged from bovine coronavirus after other cross-species spillover event. It is speculated that the flu pandemic of 1890 may have been caused by this spillover event, and not by the influenza virus, because of the related timing, neurological symptoms, and unknown causative agent of the pandemic. Besides causing respiratory infections, human coronavirus OC43 is also suspected of playing a role in neurological diseases. In the 1950s, the human coronavirus OC43 began to diverge into its present genotypes.¹ Phylogenetically, mouse hepatitis virus (*Murine coronavirus*), which infects the mouse's liver and central nervous system, is related to human coronavirus OC43 and bovine coronavirus. Human coronavirus HKU1, like the aforementioned viruses, also has its origins in rodents.

1.6. Infection in humans:

Coronaviruses vary significantly in risk factor. Some can kill more than 30% of those infected, such as MERS-CoV, and some are relatively harmless, such as the common cold. Coronaviruses can cause colds with major symptoms, such as fever, and a sore throat from swollen adenoids. Coronaviruses can cause pneumonia (either direct viral pneumonia or secondary bacterial pneumonia) and bronchitis (either direct viral bronchitis or secondary bacterial bronchitis). The human coronavirus discovered in 2003, SARS-CoV, which causes severe acute respiratory syndrome (SARS), has a unique pathogenesis because it causes both upper and lower respiratory tract infections.

Six species of human coronaviruses are known, with one species subdivided into two different strains, making seven strains of human coronaviruses altogether.



1.5. Seasonal distribution of HCoV-NL63 in Germany shows a preferential detection from November to March

Four human coronaviruses produce symptoms that are generally mild, even though it is contended they might have been more aggressive in the past:

1. Human coronavirus OC43 (HCoV-OC43), β -CoV
2. Human coronavirus HKU1 (HCoV-HKU1), β -CoV
3. Human coronavirus 229E (HCoV-229E), α -CoV
4. Human coronavirus NL63 (HCoV-NL63), α -CoV

Three human coronaviruses produce symptoms that are potentially severe:

1. Middle East respiratory syndrome-related coronavirus (MERS-CoV), β -CoV
2. Severe acute respiratory syndrome coronavirus (SARS-CoV), β -CoV
3. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), β -CoV

1.6.1. Infection in animals:

Coronaviruses have been recognized as causing pathological conditions in veterinary medicine since the 1930s. They infect a range of animals including swine, cattle, horses, camels, cats, dogs, rodents, birds and bats. The majority of animal related coronaviruses infect the intestinal tract and are transmitted by a fecal-oral route. Significant research efforts have been focused on elucidating the viral pathogenesis of these animal coronaviruses, especially by virologists interested in veterinary and zoonotic diseases.

1.6.2. Symptoms:

1. Common cold
2. Severe acute respiratory syndrome (SARS)
3. Middle East respiratory syndrome

1.6.3. Mathematical and statistical method for data analysis:

Mathematical and statistical science have much to contribute towards data mining management and analysis. In fact, most data mining techniques are statistical data analysis tools. Some of methods and techniques are well known and effective.

1.6.4. Types of analysis:

1. Descriptive analysis
2. Regression analysis
3. Factor analysis
4. Dispersion analysis
5. Discriminant analysis
6. Time series analysis

1.6.5. Time series analysis:

A time series analysis is a set of statistical observation arranged in chronological order. The procedure by which the time related factors that influence the values observed in the time series are identified and segregated is called time series analysis.

1.6.6. Uses of time series analysis:

1. To understand past behaviour
2. To plan future operation
3. To evaluate current accomplishments
4. To compare

1.6.7. Component of time series:

1. Trend-T
2. Cyclic fluctuation-C
3. Seasonal variation-S
4. Irregular variation-I

CHAPTER 2

Literature survey

Gaetano Perone [2] had forecasted the short to medium-term incidence of COVID-19 epidemic through the medium of an autoregressive integrated moving average (ARIMA) model, applied to Italy, Russia, and the USA the analysis is carried out on the number of new daily confirmed COVID-19 cases, collected by World meter website.

Naresh Kumar[3] in his paper had commented that as Millions of people have been infected and lakhs of people have lost their lives due to the worldwide ongoing novel coronavirus (COVID-19) pandemic. It is of utmost importance to identity the future infected cases and the virus spread rate for advanced preparation in the healthcare services to avoid deaths. Therefore, we use day level information of COVID-19 spread for cumulative cases from whole world and 10 mostly affected countries. We utilize the temporal data of coronavirus spread from Jan 22, 2020 to May 20, 2020. we model the evaluation of the COVID-19 outbreak, and perform prediction using ARIMA and prophet time series forecasting models. Effectiveness of the modes are evaluated based on the mean absolute error, root mean square error, root relative squared error, and mean absolute percentage error. Our analysis can help in understanding the trends of the disease outbreak, and provide epidemiological stage information of adopted countries. Our investigations shows that ARIMA model is more effective for forecasting COVID-19 prevalence. The forecasting results have potential to assist governments to plan policies to contain the spread of the virus.

Amirhoshang Hoseinpour Dehkordi[4] et.al had done his research and concluded that the 2019 novel-coronavirus (COVID-19) has affected 181 countries with approximately 1197405 confirmed cases (by 5th April). Understanding the transmission dynamics of the infection in each country which got affected on a daily basis and evaluating the effectiveness of control policies are critical for our further actions. To date, the statistics of COVID-19 reported cases show that more than 80% of infected are mild cases of disease, around 14% of infected have severe complications, and about 5% are categorized as critical disease victims.

Today's report (5th April 2020; daily updates in the prepared website) shows that the confirmed cases of COVID-19 in the United States, Spain, Italy, and Germany are 308850, 126168, 124632, and 96092, respectively. Calculating the total case fatality rate (CFR) of Italy (4th April 2020), about 13.3% of confirmed cases have passed away. Compared with South Korea's rate of 1.8% (seven times lower than Italy) and China's 4% (69% lower than Italy), the CFR of Italy is too high. Some effective policies that yielded significant changes in the trend of cases were the lockdown policy in China, Italy, and Spain (the effect observed after some days), the shutdown of all nonessential companies in Hubei (the effect observed after 5 days), combined policy in South Korea, and reducing working hours in Iran.

Amirhoshang Hoseinpour Dehkordi et.al[5] had provided both a model by using emulative cases and cumulative death toll for SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) outbreak in four countries, China, Italy, South Korea, and Turkey, starting from the first diagnosis and to compare associated indicators. The most successful estimation was obtained from the cubic model with natural logarithm for China, Italy, South Korea, and Turkey. The success of the models was around 99%. However, differences began to emerge in China, Italy, and South Korea after the second week. Although the highest number of new cases per 1 million people in China was 9.8 on February 28, 2020; it was 108.4 on March 21, 2020, in Italy; and this was 16.6 on March 5, 2020, in South Korea. On the other hand, the number of new cases was 24.6 per 1 million people on March 27, 2020, in Turkey. The log-cubic model proposed in this study has been set forth to obtain successful results for aforementioned countries, as well as to estimate the course of the COVID-19 outbreak. Other factors such as climacteric factors and genetic differences, which may have an impact on viral spreading and transmission, would also have strengthened the model prediction capacity.

Soumya Ranjan Nayak [11] has analysed that SARS Cov-2, COVID-19 (Coronavirus) emerged in Wuhan in early December 2019 and then spread exponentially across the globe. Although, a series of prevention strategies (lockdown, social-distancing) have been enforced to control this pandemic. In this study, she had made statistical analysis in terms of Gaussian modelling, ANOVA

test and probabilistic model. After applying ANOVA concluded that the recovery rate for all the countries are significantly higher than the mortality rate except for the UK where the mortality rate is significantly higher than the recovery rate. Gaussian modelling applied here was able to predict the original peak values of confirmed cases of countries. Using the probabilistic model, we were able to predict that there is around 5% probability that a person in India will be tested positive for COVID-19 on 100 tests.

Naushad Mamode Khan [12] has proposed in his paper an integer-valued time series model to analyse the series of COVID-19 cases in Mauritius wherein the corresponding innovation term accommodates for covariate specification. In this set-up, sanitary curfew followed by sanitization and sensitization campaigns, time factor and safe shopping guidelines have been tested as the most significant variables, unlike climatic conditions. The over-dispersion estimates and the serial auto-correlation parameter are also statistically significant. This study also confirms the presence of some unobservable effects like the pathological genesis of the novel coronavirus and environmental factors which contribute to rapid propagation of the zoonotic virus in the community. Based on the proposed COM-Poisson mixture models, we could predict the number of COVID-19 cases in Mauritius. The forecasting results provide satisfactory mean squared errors. Such findings will subsequently encourage the policymakers to implement strict precautionary measures in terms of constant upgrading of the current health care and wellness system and re-enforcement of sanitary obligations.

In the additive Holt-Winters' seasonal exponential smoothing model, it is theoretically possible for smoothing parameters in the usual $(0, 1)$ interval to produce 'non-invertible' models. A sample of 406 monthly series reveal that this is a real concern. In the multiplicative model, reasonable estimation procedures produce smoothing constants outside the additive invertible region. When this occurs, the impact on forecasts of values in the distant past is much larger than for recent values and forecasts are poor.

Robert G. Brown [4] has used exponential smoothing as a formalization of the familiar learning process, which is a practical basis for statistical forecasting. Higher orders of smoothing are defined by the operator $St_n(x) = \alpha St_{n-1}(x) + (1 - \alpha) St_{n-1}(x)$,

where $S_t^0(x) = x_t$, $0 < \alpha < 1$. If one assumes that the time series of observations $\{x_t\}$ is of the form $x_t = n_t + \sum_{i=0}^N a_i t_i$, where n_t is a sample from some error population, then least squares estimate of the coefficients a_i can be obtained from linear combinations of the operators $S, S^2, \dots, S^{N+1}$. Explicit forms of the forecasting equations are given for $N = 0, 1$, and 2 . This result makes it practical to use higher order polynomials as forecasting models, since the smoothing computations are very simple, and only a minimum of historical statistics need be retained in the file from one forecast to the next.

The Holt-Winters forecasting procedure [7] is a simple widely used projection method which can cope with trend and seasonal variation. However, empirical studies have tended to show that the method is not as accurate on average as the more complicated Box-Jenkins procedure. This paper points out that these empirical studies have used the automatic version of the method, whereas a non-automatic version is also possible in which subjective judgement is employed, for example, to choose the correct model for seasonality. The paper re-analyses seven series from the Newbold-Granger study for which Box-Jenkins forecasts were reported to be much superior to the (automatic) Holt-Winters forecasts. The series do not appear to have any common properties, but it is shown that the automatic Holt-Winters forecasts can often be improved by subjective modifications. It is argued that a fairer comparison would be that between Box-Jenkins and a non-automatic version of Holt-Winters. Some general recommendations are made concerning the choice of a univariate forecasting procedure. The paper also makes suggestions regarding the implementation of the Holt-Winters procedure, including a choice of starting values

Peter R. Winters [10] had commented in his paper that the growing use of computers for mechanized inventory control and production planning has brought with it the need for explicit forecasts of sales and usage for individual products and materials. These forecasts must be made on a routine basis for thousands of products, so that they must be made quickly, and, both in terms of computing time and information storage, cheaply; they should be responsive to changing conditions. The paper presents a method of forecasting sales which has these desirable characteristics, and which in terms of ability to forecast compares favorably with other, more traditional methods. Several models of the exponential forecasting system are presented, along with several examples of application.

CHAPTER-3

\

AIM AND SCOPE:

AIM OF THE PROJECT:

In this paper the COVID data for India was collected from WHO website from APR 2020 to Jan 2021 and statistical analysis was conducted for the following methods:

1. Run test
2. Mann Kendall tests
3. Moving average
4. Rescale range analysis

SCOPE OF THE PROJECT:

Further more precise forecasting may be provided by continuously updating these data, adding interventions and other real aspects, and applying the model to other countries. These findings of this research works are helpful for future prediction of the confirm cases of the COVID-19 which is the helpful for the government entities and other healthcare organizations to try to mitigate the problem of the outbreak. Moreover, many statistical software's are helpful to understand the forecasting methods and the differences. In this work, worldwide data is used and tested for various models to forecast the future scenario. The dataset may be increasing with time, the methodologies developed in this work are just need a re-run in the future to see the future value of the confirm cases anytime.

CHAPTER-4

Research Methodology

3.1. Runs Test

The first version of this nonparametric test was presented by Mood (1940) and is based on the same run's statistic as the two-sample test by Wald and Wolfowitz (1940), which is why this test is sometimes mistakenly referred to as the Wald and Wolfowitz runs test. However, the article by Mood makes reference to the article by Wald and Wolfowitz and the asymptotic distribution of the statistic uses also the results given by these authors.

A run is a sequence of identical events, preceded and succeeded by different or no events. The runs test used here applies to binomial variables only. For example, in ABBABBB, we have 4 runs (A, BB, A, BBB).

A sample will be considered as randomly distributed if no particular structure can be identified. Extreme cases are repulsion, where you have all observations of one kind on the left, and all the remaining observations on the right, and alternation where the elements of the two kinds are alternating as much as possible.

In the case of the two-tailed (or two-sided) test, the null (H_0) and alternative (H_a) hypotheses are:

- H_0 : data are randomly distributed.
- H_a : data are not randomly distributed.

In the one-tailed case, you need to distinguish the left-tailed (or lower- tailed or lower one-sided) test and the right-tailed (or upper-tailed or upper one-sided) test. In the left-tailed test, the following hypotheses are used:

- H_0 : data are randomly distributed.
- H_a : there is repulsion between the two types of events.

In the right-tailed test, the following hypotheses are used:

- H_0 : data are randomly distributed.

- H_a : The two types of events are alternating.

The expectation of the number of runs R is given by:

$$E(R) = 2mn/N$$

Where m is the number of events of type 1, and n the number of events of type 2, and N is the total sample size.

The variance of the number of runs R is given by:

$$V(R) = 2mn (2mn - N) / [N^2 (N-1)]$$

The minimum value of R is always 2. The maximum value is given by $2\text{Min}(m, n) - t$, where t is 1 if $m=n$, and 0 if not.

If r is the number of runs measured on the sample, it was shown by Wald and Wolfowitz that asymptotically, when m or n tend to infinity,

$$\frac{(r - E(R))}{\sqrt{V(R)}} \rightarrow N(0, 1)$$

Where $N(0, 1)$ is the standard normal distribution.

3.2. Mann- Kendall trend test

A nonparametric trend test has first been proposed by Mann (1945) then further studied by Kendall (1975) and improved by Hirsch *et al* (1982, 1984) who allowed to take into account a seasonality.

The null hypothesis H_0 for these tests is that there is no trend in the series. The three alternative hypotheses that there is a negative, non-null, or positive trend can be chosen.

The Mann-Kendall tests are based on the calculation of Kendall's tau measure of association between two samples, which is itself based on the ranks with the samples.

In the particular case of the trend test, the first series is an increasing time indicator generated automatically for which ranks are obvious, which simplifies the calculations. The S statistic used for the test and its variance are given by:

$$S = \sum_{i=1}^{x-1} \sum_{j=i+1}^x \text{sgn}(x_j - x_i)$$

$$\text{Var}(S) = \frac{n(n-1)(2n+5)}{18}$$

Where n is the number of observations and x_i (1 to n) are the independent observations.

An exact p-value can be calculated if there are no ties in the series and if the sample size is less than 50. If an exact calculation is not possible, a normal approximation is used, for which a correction for continuity is optional but recommended.

The Mann-Kendall trend test requires that the observations are independent (meaning the correlation between the series with itself with a given lag should not be significant). In the case where there is some autocorrelation in the series, the variance of the S statistic has been shown to be underestimated. Therefore, several improvements have been suggested. There are two alternative methods, the first one published by Hamed and Rao (1998) and the second by Yue and Wang (2004). The first method performs well in the case of no trend in the series (it avoids identifying a trend when it is in fact due to the autocorrelation) and the second has the advantage of performing better when there are both a trend and an autocorrelation. Both methods imply the computation of Sen's slope estimator (Sen, P. K. (1968)). Before running a Mann-Kendall trend test, it is of course recommended to first check the autocorrelations of a series.

3.3 Moving Average

This model is a simple way to take into account past and optionally future observations to predict values. It works as a filter that is able to remove noise. While with the smoothing methods defined below, an observation influences all future predictions (even if the decay is exponential), and in the case of the moving average the memory is limited to q . if the constant l is set to zero, the prediction depends on the past q values and on the current value, and if l is set to one, it also depends on the next q values. Moving averages are often used as filters, and not as way to do accurate predictions. Where l is a constant, which, when set to zero, allows the prediction to depend on the q previous values and on the current value.

If l is set to one, the prediction also depends on the q next values. These are the weights. Weights can be either constant, fixed by the user, or based on existing optimal weights for a given application.

3.4. Rescaled Range Analysis

Rescaled Range (R/S) analysis is a technique which is frequently applied to real time problems to detect any biases in behaviour overtime. This method is used to find the memory effect otherwise known as persistence in the trends. Rescaled Range Analysis (R/S Analysis) is a non-parametric methodology developed by H.E. Hurst, a British hydrologist in 1951. He applied this methodology to study the long-term storage capacity of reservoirs is generally used for distinguishing random time series from biased random time series and to study the persistence of trends and also presence of periodic and non-periodic cycles in a time series.

Rescaled The Range $(R/S)_n$ computed by Hurst for a time series $X_1, X_2, \dots, X_n$ of length n is given by

$$(R/S)_n = C * n^H \quad (1)$$

Where C & H are constants and H is called Hurst coefficient

Equation (1) can be written as

$$\log (R/S)_n = \log C + H * \log (n)$$

H is obtained as the slope of $\log (R/S)_n$ versus $\log(n)$ plot

A Hurst exponent ranges between 0 & 1 and measures three types of trends in a time series: Persistence, randomness, and Anti-persistent or mean reversion.

- If $0.5 < H \leq 1.00$, implies a persistent time series. i.e.) a time series characterised by long memory effects. This means if the trend has been positive in the last observed period, the changes are it will continue to be positive in the next period. Conversely if it has been negative in the last period it is more likely to continue to be negative in the next period.
- If $H=0.50$, then the time series is a random time series and implies the absence of long-term statistical dependence.

- If $0 \leq H < 0.5$, implies anti-persistent time series. An anti-persistent system reverses itself more frequently than a random one.

3.5 EMPIRICAL FRAMEWORK OF RESCALED RANGE

The sample of size N is split into A continuous sub periods of length n (where n is an integer which evenly divides into the sample length). Each sub period may be denoted as I_a (where $a = 1 \dots A$) and each element of I_a may be denoted as $N_{k,a}$ (where $k = 1 \dots n$). For each I_a , we must rescale the data by subtracting the sub period mean:

$$Y_{k,a} = (N_{k,a} - e_a) \quad k = 1 \dots n \quad (2)$$

where e_a is the average value of the elements in subperiod I_a and may be estimated as

$$e_a = \frac{\sum_{k=1}^n N_{k,a}}{n} \quad (3)$$

The resulting series ($Y_{k,a}$) has a mean of zero and if we cumulatively sum $Y_{k,a}$ we get the series ($X_{k,a}$) in which the last value will always be zero since the series by construction has mean value zero. The range of the cumulative series $X_{k,a}$ may be estimated as:

$$R_{Ia} = \max(X_{k,a}) - \min(X_{k,a}) \quad \text{where } 1 \leq k \leq n \quad (4)$$

R_{Ia} may be normalized to create the rescaled range by dividing each range by sample standard deviation (S_{Ia}), giving R_{Ia} / S_{Ia} , where S_{Ia} may be estimated as:

$$S_{Ia} = \frac{\sqrt{\sum_{k=1}^n (N_{k,a} - e_a)^2}}{n} \quad (5)$$

This process results in A rescaled range and the average rescaled range across the whole sample for subperiods of length n may be estimated as:

$$(R/S)_n = \frac{\sum_{a=1}^A (R_{Ia} / S_{Ia})}{A} \quad (6)$$

CHAPTER-5

RESULTS AND DISCUSSIONS

The COVID data for India was collected from WHO website from APR 2020 to Jan 2021 and Statistical analysis was conducted using the following methods

1.RUNS TEST

(a)Recovered cases

Table 5.1 :One-sample runs test / Two-tailed test - Recovered	
R	2
r Expected value	35.759
p-value (Two-tailed)	<0.0001
alpha	0.050

Test interpretation:

H₀: Data are randomly distributed

H_a: Data are not randomly distributed

As the computed p-value is lower than the significance level $\alpha=0.05$, one should reject the null hypothesis H₀, and accept the alternative hypothesis H_a.

(b)Confirmed cases

Table5. 2 :One-sample runs test / Two-tailed test - Confirmed	
R	2
r Expected value	35.759
p-value (Two-tailed)	<0.0001
alpha	0.050

Test interpretation:

H₀: Data are randomly distributed

H_a: Data are not randomly distributed

As the computed p-value is lower than the significance level $\alpha=0.05$, one should reject the null hypothesis H₀, and accept the alternative hypothesis H_a

(c)Death Cases

Table 5.3 :One-sample runs test / Two-tailed test - Death	
R	2
r Expected value	37.460
p-value (Two-tailed)	<0.0001
alpha	0.050

Test interpretation:

H₀: Data are randomly distributed

H_a: Data are not randomly distributed

As the computed p-value is lower than the significance level $\alpha=0.05$, one should reject the null hypothesis H₀, and accept the alternative hypothesis H_a.

4.2. Mann Kendall test

(a)Recovered Cases

Table 5.5:Mann-Kendall trend test/two – tailed test - Recovered	
Kendall's tau	1.000
S	3504.000
Var(S)	74523.333
p-value (Two-tailed)	<0.0001
alpha	0.050

Test interpretation:

H₀: There is no trend in the series

H_a: There is a trend in the series

As the computed p-value is lower than the significance level $\alpha=0.05$, one should reject the null hypothesis H₀, and accept the alternative hypothesis H_a.

(b)Confirmed cases

Table 5.4:Mann-Kendall trend test/two – tailed test -Confirmed	
Kendall's tau	1.000
S	3738.000
Var(S)	74401.333

p-value (Two-tailed)	<0.0001
alpha	0.050

Test interpretation:

H0: There is no trend in the series

Ha: There is a trend in the series

As the computed p-value is lower than the significance level $\alpha=0.05$, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.

(c)Death Cases

Table 5. 6:Mann-Kendall trend test/two – tailed test-Death	
Kendall's tau	1.000
S	2546.000
Var(S)	65441.333
p-value (Two-tailed)	<0.0001
alpha	0.050

Test interpretation:

H0: There is no trend in the series

Ha: There is a trend in the series

As the computed p-value is lower than the significance level $\alpha=0.05$, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.

4.3. Moving average

(a)Summary statistics: For Recovered Cases

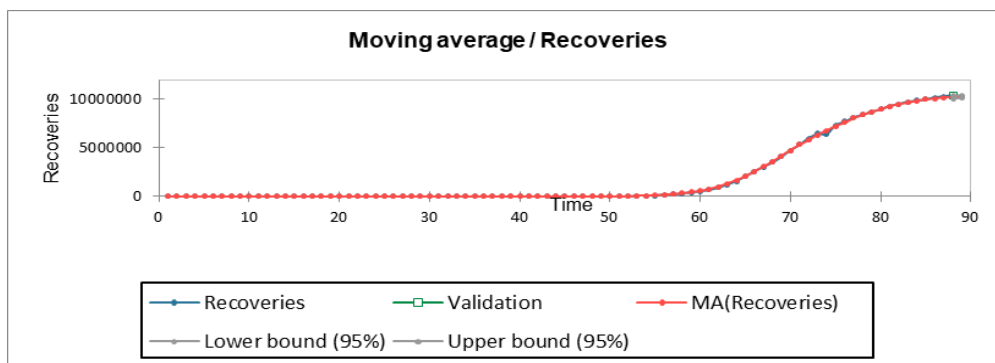


Fig 5.1 Moving Average of Recoveries

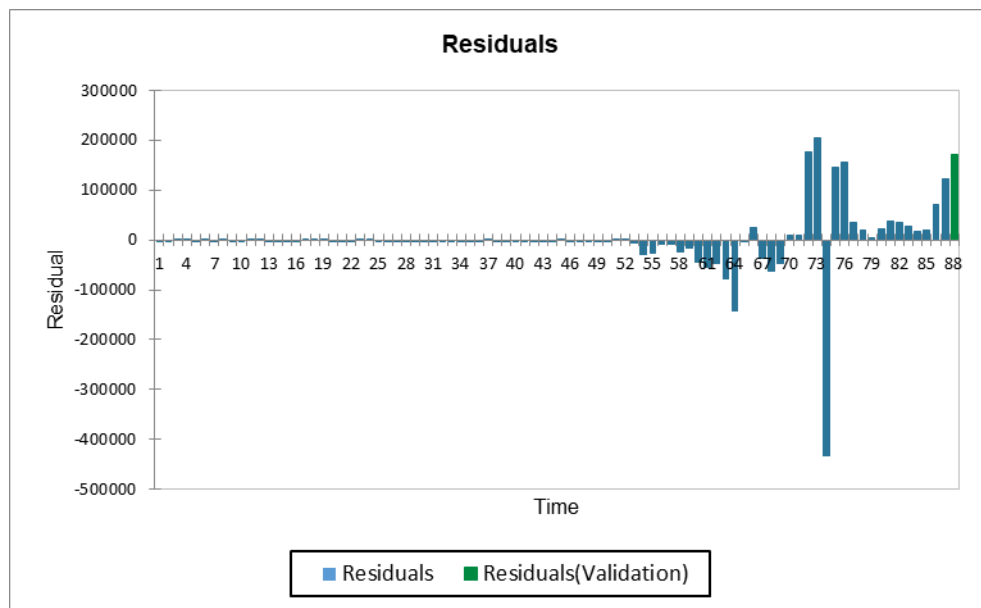


Fig 5.2 Residual plot of Recoveries

(b) Summary statistics: For Confirmed Cases

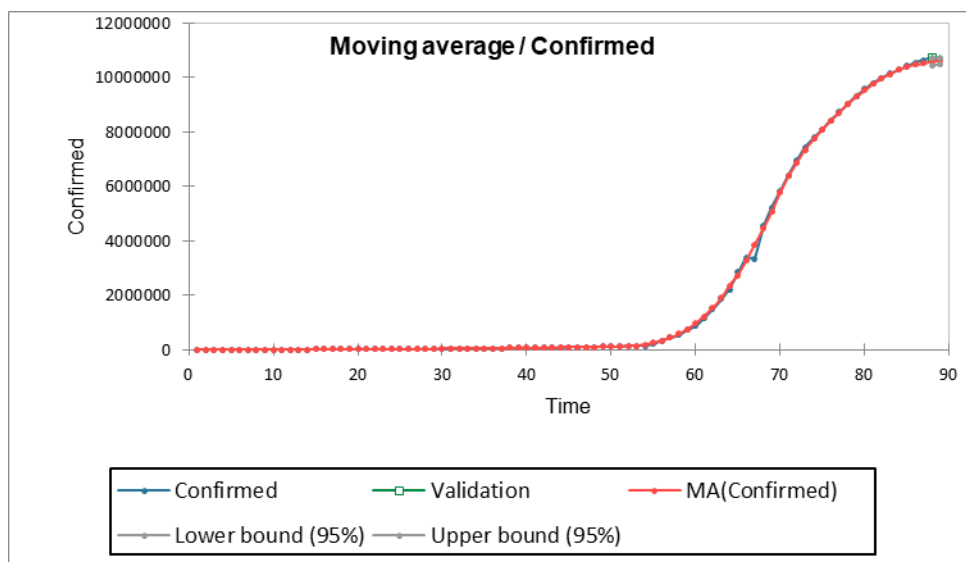


Fig 5.3 Moving Average of Confirmed Cases

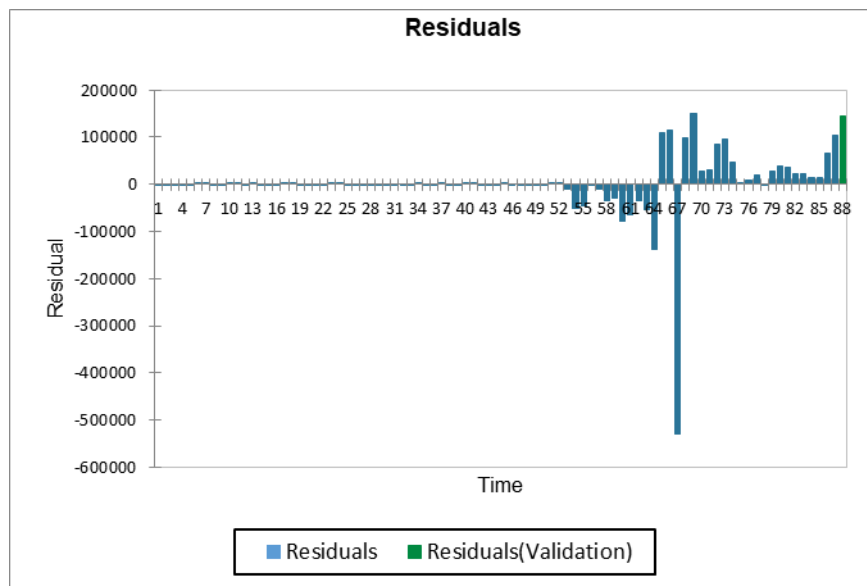


Fig 5.4 Residual plot of Confirmed cases

© Test interpretation:

H₀: There is no trend in the series

H_a: There is a trend in the series

As the computed p-value is lower than the significance level $\alpha=0.05$, one should reject the null hypothesis H₀, and accept the alternative hypothesis H_a.

(c) Summary statistics: For death Cases

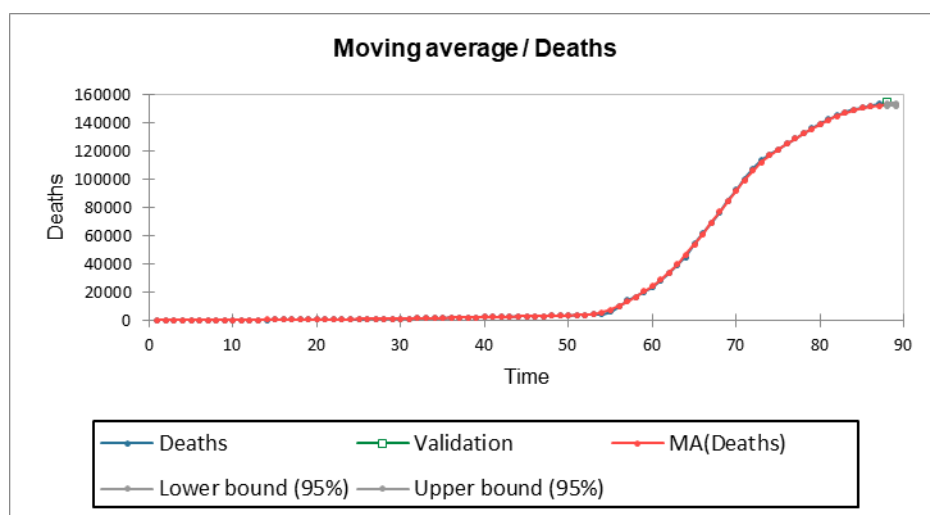


Fig 5.5 Moving Average of Death Cases

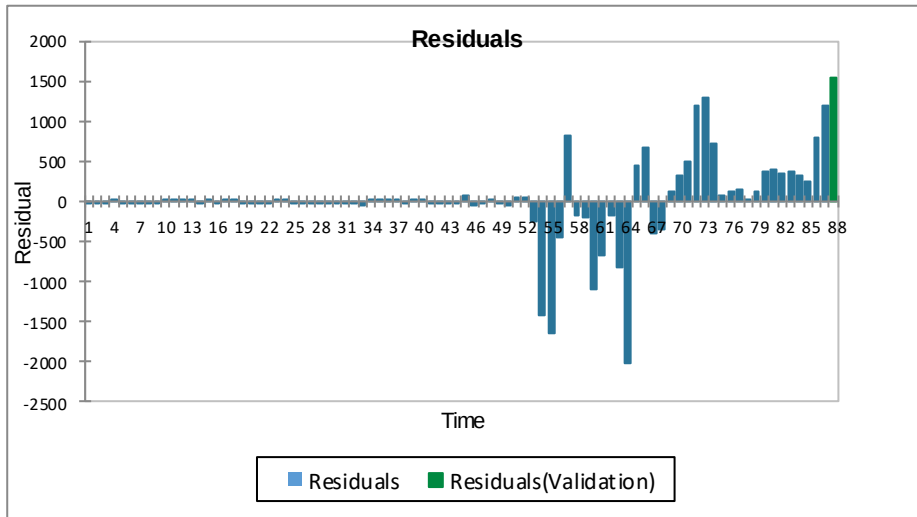


Fig 5.6 Residual plot of Death cases

Table 5.7 Error Measures of Moving Average

	Recovered	Confirmed	Death
Observations	87	87	87
SSE	376045519347.343	418719603587.465	19970020.188
MSE	4322362291.349	4812869006.752	229540.462
RMSE	65744.675	69374.844	479.104
MAPE	4.193	3.132	2.906
MPE	-2.816	-1.927	-1.728
MAE	25575.803	25805.468	243.626
R²	1.000	1.000	1.000

4.4. Rescaled Range Analysis:

Data collection is based on WHO report for the number of confirmed cases, recovered case and death cases in India from March 2020 to January 2021 consisting of 90 data points.

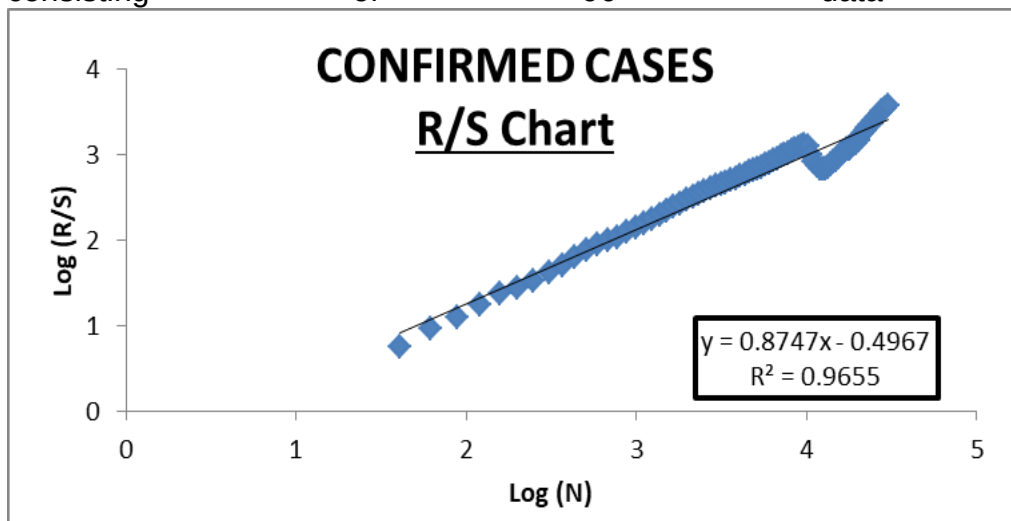


Fig 5.7 : R/S Chart for confirmed cases

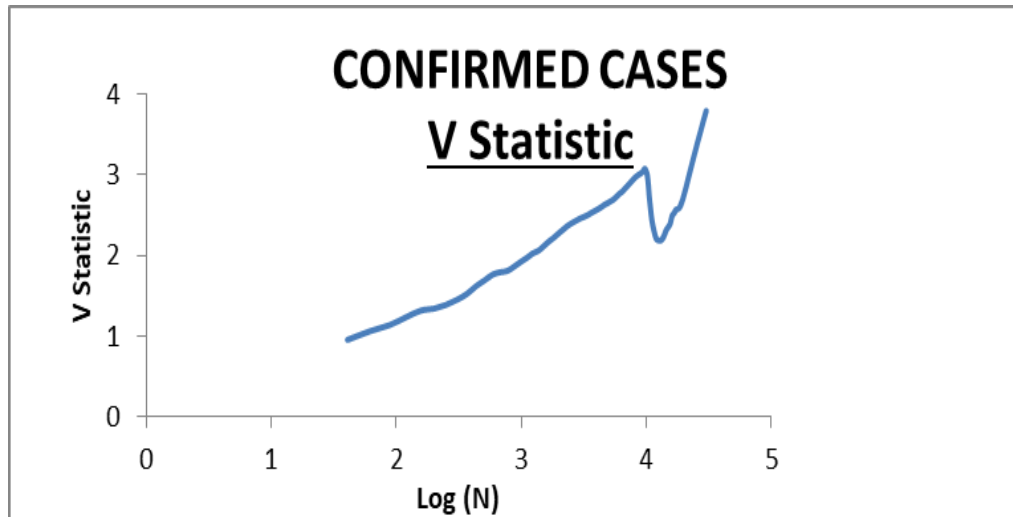


Fig 5.8 : V Chart for confirmed cases

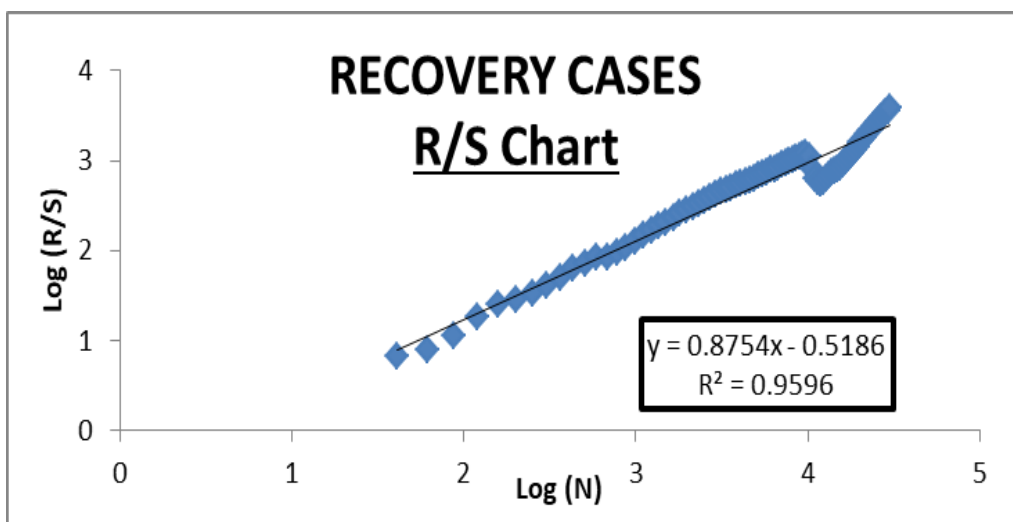


Fig 5.9 :R/S Chart for recovered cases

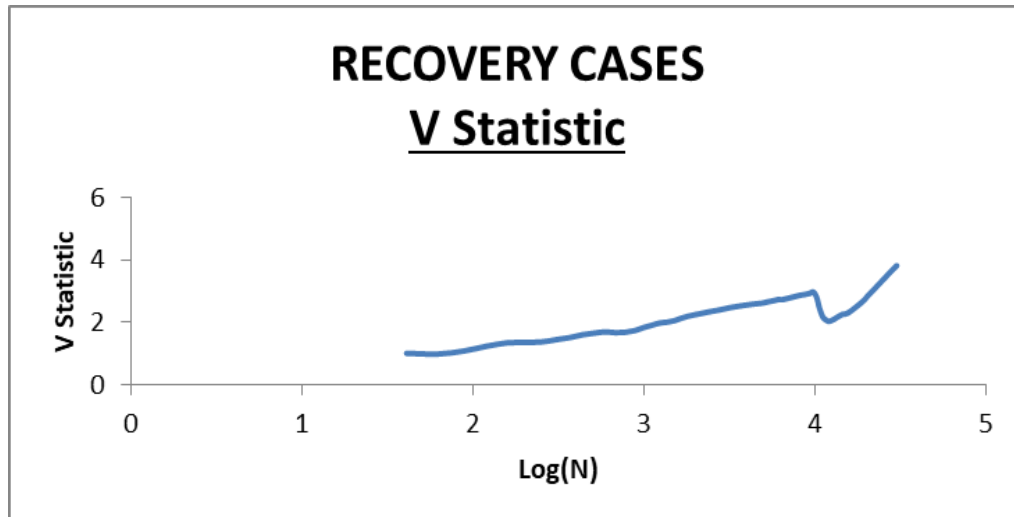


Fig 5.10 : V Chart for recovery cases

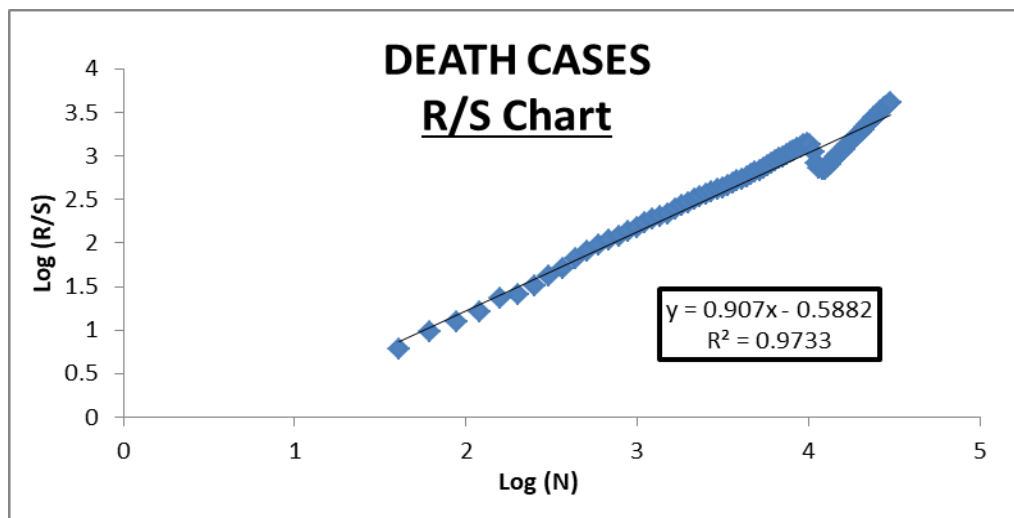


Fig 5.11 : R/S Chart for death cases

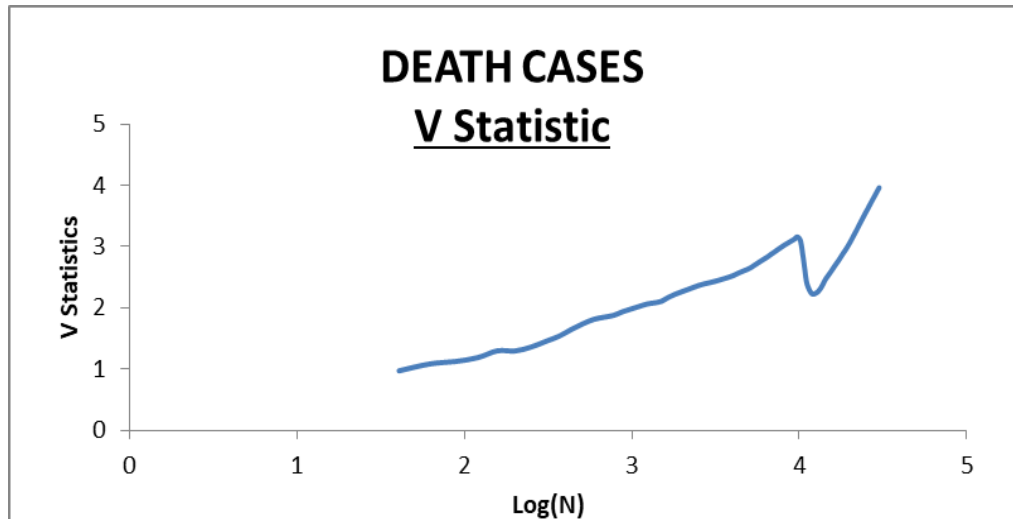


Fig 5.12 : V Chart for death cases

Line of best fit

1. The line of best fit for the confirmed cases was found to be $y = 0.8747x - 0.4967$
2. The line of best fit for the recovered cases was found to be $y = 0.8754x - 0.5186$
3. The line of best fit for the death cases was found to be $y = 0.907x - 0.5882$

Hurst Exponent

1. $H = 0.8747$ for the confirmed cases
2. $H = 0.8754$ for recovered cases
3. $H = 0.907$ for death cases

CHAPTER-6

CONCLUSION

From the SIR models of COVID-19 in India, the following conclusions are made;

1. Using Run test, it was found that the data for analysis for all the three cases were not randomly taken.
2. Using Mann Kendall test, it was concluded there is a trend in the given series for all the three cases.
3. Using Moving average method, it was found that the MAPE of the predicted value was

Recovered	Confirmed	Deaths
4.193	3.132	2.906

for April 2021 in which death gave the lowest error measure.

4. Using Range Analysis, $H = 0.8747$ which shows that the confirmed cases time series is a persistent series (if it increasing in one time period it will continue to increase in the next time period)
5. Using Range Analysis, $H = 0.8754$ which shows that the confirmed cases time series is a persistent series (if it increasing in one time period it will continue to increase in the next time period)
6. Using Range Analysis, $H = 0.907$ which shows that the confirmed cases time series is a persistent series (if it increasing in one time period it will continue to increase in the next time period)
7. Of all the three ,it was found that death cases was more persistent than the other cases.

REFERENCES

1. **Archibald B.C. (1990).** Parameter space of the Holt-Winters' model. *International Journal of Forecasting*, **6**, 199-209.
2. **Box G. E. P. and Jenkins G. M. (1976).** Time Series Analysis: Forecasting and control. Holden-Day, San Francisco.
3. **Brockwell P.J. and Davis R.A. (1996).** Introduction to Time Series and Forecasting. Springer Verlag, New York.
4. **Brown R.G. (1962).** Smoothing, Forecasting and Prediction of Discrete Time Series. Prentice-Hall, New York.
5. **Brown R.G. and Meyer R.F. (1961).** The fundamental theorem of exponential smoothing. *Operations Research*, **9**, 673-685.
6. **Chatfield, C. (1978).** The Holt-Winters forecasting procedure. *Applied Statistics*, **27**, 264-279.
7. **Holt C.C. (1957).** Forecasting seasonals and trends by exponentially weighted moving averages. ONR Research Memorandum 52, Carnegie Institute of Technology, Pittsburgh.
8. **Makridakis S.G., Wheelwright S.C. and Hyndman R.J. (1997).** Forecasting: Methods and Applications. John Wiley & Sons, New York.
9. **Shumway R.H. and Stoffer D.S. (2000).** Time Series Analysis and Its Applications. Springer Verlag, New York.
10. **Winters P.R. (1960).** Forecasting sales by exponentially weighted moving averages. *Management Science*, **6**, 324-342
11. **Soumya rajan nayak.** A statistical analysis of COVID-19 using gaussian and probabilistic model
12. **Naushad Mamode khan.** Studying the trend of the novel coronavirus series in Mauritius and its implications.

