Mathematical and Statistical Analysis on the effect of COVID-19 on the Internet Infrastructure and a case study on its effect on people

A project work submitted to the Pondicherry University

in partial fulfillment of the requirements for the

award of the degree of M. Sc. in Physics

by

RINDHUJA TREESA JOHNSON

Regd. No: 17380319

Department of Physics

Pondicherry University

R.V.Nagar, Kalapet,

Puducherry – 605 014

INDIA

June 2022

CERTIFICATE

This is to certify that the work embodied in this project entitled "**Mathematical and Statistical Analysis on the effect of COVID-19 on the Internet Infrastructure and a case study on its effect on people**" has been carried out by **Ms.RINDHUJA TREESA JOHNSON** during the academic year 2021-2022 under my supervision in partial fulfillment of the requirements for the award of the degree of M.Sc.,in Physics and the same has not been submitted for the award of the degree or diploma of any university or institution.

PLACE: PUDUCHERRY

DATE: 23 JUNE 2022

(Supervisor)

(Dr. A. RAMESH NAIDU)

DECLARATION

I hereby declare that the work embodied in this report entitled, "**Mathematical and Statistical Analysis on the effect of COVID-19 on the Internet Infrastructure and a case study on its effect on people**" submitted to the Department of Physics, Pondicherry University, Puducherry in partial fulfillment of the requirements for the award of the degree of M.Sc., in Physics has been carried out under the supervision of **Dr.A.RAMESH NAIDU** Associate professor, Department of Physics, Pondicherry University, Puducherry. This work has not been submitted for the award of the degree or diploma of any university or institution to the best of my knowledge.

Place: Puducherry

Date : 23 JUNE 2022

(RINDHUJA TREESA JOHNSON)

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. A.RAMESH NAIDU, Associate Professor, Department of Physics, Pondicherry University for his constant support, valuable suggestions and encouragement to explore for better understanding of concepts during the project period.

I am very much thankful to the Head of the Department of Physics and the faculty members of the Department of Physics, Pondicherry University for their valuable suggestions and ideas.

I extend my gratitude to the research scholars in the Department of Physics who have guided me throughout the work and helped in finding better insights with much enthusiasm.

I am grateful to the non-teaching staff of the Department of Physics, Pondicherry university for their constant help during the course of my studies.

I also thank all my friends and classmates who have suggested me the sources for readings and for helping me with the necessary software supports. I am also indebted to my parents and my siblings who have supported me throughout the work and also others who helped indirectly to carry out this project work.

– RINDHUJA TREESA JOHNSON

ABSTRACT

COVID – 19 has created a new era in the world. When most of the sectors have witnessed a nosedive, the Internet served as a medium which could save them all from oblivion. This project emphasis on how COVID – 19 has changed the Internet usage trends and technologies in India. The study gives a detailed statistical and mathematical study on the different parameters related to Internet technologies such as the OFCs, 4G networks, usage durations. Further, in order to get a picture on how the sudden demand and increased exposure to internet has affected the people, a survey was conducted which gathered the personal views of people from different fields. The data retrieved also gave insightful information on the internet involvement in their lives. And adding to it, a quick study on how all these demands could have affected the 5G era in India has also been discussed along with the different other schemes that were implemented by the government to enhance the technology as an emergency of the time.

Contents

Chapter 1

INTRODUCTION

The COVID – 19 pandemic out-break has brought in a war-like scenario to the world. The economic, social and individual damages that came along interrupted the ever-progressing world. When a war or a natural disaster is far more predictable, the communicable diseases come as a surprise. The unexpected arrival of COVID – 19 has traumatized people of all nations without any exception. However, the revival from these hard times were highly correlated to the financial, medical and infrastructural growths of each country.

In a country like India, where majority of the population just have enough resources and money to run their daily errands the conditions could have been worse. The economy was severely hit, which was evident from the Gross Domestic Product (GDP) witnessing a contraction of 7.3% in the fiscal year 2020 – 2021.

There is no doubt that the COVID – 19 has impacted on us in such a way that it's after-effects could last for a life time or even more. However, in this project, I would like to shine light on the thoughts of how this mishap would have brought

unexpected changes to our lifestyle and to our standard of living.

We live in a world where we speak using fingertips, where our ideas could reach the opposite side of the planet within milliseconds, where we could travel virtually or physically, from place to place in matter of seconds or hours, as if distance has either become imaginary or diminished! The Science and Technology has always been on an exponential growth from the discovery of wheels to the present Artificial Intelligence. When we give a closer look at the innovative technologies that bloomed throughout the centuries and eras, we could understand that it was always the necessity that led them on.

Internet was popular from its very first day. The ARPANET (Advanced Research Projects Agency Network) developed by the US Defense Projects Agency in 1969 serves as predecessor to the present-day Internet technology. The ARPANET was developed for the communication between four computers during the cold war in US. The idea of Transmission Control Protocol/ Internet Protocol (TCP/IP) laid the foundation for it, which still is one of the fundamental network interconnection rules-set. The adaptation of this technology was much quicker than anyone would have even imagined of. With this basic idea, the further developments took place eventually. The invention of World Wide Web in 1989 by Berners-Lee at CERN unwrapped the endless opportunities in the world of internet. His idea was to connect the academic institutions around the world for better and faster exchange of information, however, in reality he connected the whole world and thus forming a new era of life. [1]

This project aims to put up an analogous situation in this respect. With the

outbreak of the pandemic, we were forced to live in isolation and were restricted from going to offices or schools. This resulted in an emergency where we had to come up with alternatives to mend the broken chain of education and work-life. We have developed so much that we could never imagine of leaving behind two years so that we could start fresh. The idleness and loneliness could not be tolerated; we had to come up with alternatives.

It was not a new idea that came handy to tackle the problem at hand. We had the idea well before all these, but we were not prepared. The idea of online education and work-from-home were exercised by many educational institutions and companies long before. There were plenty of online platforms that could facilitate the distance education via pre-captured video lectures and standard notes. The case with distance working was also not new. Multi-National Companies and IT companies had numerous ways of connecting with their employees across countries.

This tells us that the real task at hand was not to come up with an idea, but was to implement it on such a large scale. In India, even though the number of internet users haven't been increased due to the COVID – 19, there is a significant change in the different technologies adapted for the same. The Optical Fiber Cable technology, for example, was on a constant increase till the fourth quarter of the 2019, however, we could gather a significant increase in the slope once it reached 2020. There are such major and minor changes that accompanied the COVID -19 in India which would have aided in a surge of the internet technology.

This project focuses on the different Internet infrastructural developments or

improvements or any significant changes that could have been aided by to the emergency situation that arose due to the pandemic. It mainly focuses on statistical and mathematical inferences that could be obtained from the data available from the open resources. The whole of this work has utilized the Python libraries accessed in a Jupyter Notebook. The data is collected from Telecom Regulatory Authority of India (TRAI) website and is dating from 2016 January to 2021 December. The monthly data is grouped into Quarterly data (Each quarter consists of the cumulative or average of the three months' data) and used for different analysis. The major analysis includes Extrapolation of graphs, Principal Component Analysis (PCA), Singular Value Decomposition (SVD), L*ö*wdin Orthogonalizations (Symmetric and Canonical matrices) and Correlations. [2]

Furthermore, the project includes a survey conducted among 354 random people who are native to Kerala state on their personal experiences with the internet usage during the lock downs and social distancing phases. It mainly focuses on the role of internet in the lives of people during these hard times. Different statistical analysis has been done on this primary data to give insights on how the Internet served during the COVID – 19.

Chapter 2

COVID – 19 AND INTERNET IN INDIA

2.1 COVID – 19 IN INDIA

With a population of over 1.4 billion spread across an area of 3.287 million km^2, India is a thickly populated country which is vulnerable to massive outbreaks of communicable diseases. This was quite evident when the COVID – 19 had its various waves hit the country. The first remote case was reported on 30th January 2020 in the state of Kerala. The cases were on a steady increase from the date which resulted in the first nation-wide lockdown on 25^{th} March 2020. The first phase of this lockdown lasted for 33 days. This first wave lockdown was unexpected and resulted in economic crashes throughout the country. There were numerous issues that sprang up around this sudden closure of all services including the termination of all modes of transportation at all levels. Students

were stuck in different hostels devoid of means of transportation, so were the people in different countries. The Vande-Bharat mission was implemented by the Central government to bring those stranded in the gulf countries back to India. This has been recognised as one of the greatest civilian repatriation programs in the history of the world and it is ongoing to date rescuing Indian citizens stuck in Ukraine and Russia.

The first lockdown, or more widely known as the First wave, unlocks started on 16^{th} April 2020 when a very few restrictions were lifted. The second and third unlock phases were on 20^{th} April 2020 and 31^{st} May 2020. In public's point of view, these unlocks did not affect them much as they were still not allowed to go for work or school or any other social activities. At this point, when people were forced to stay within their homes arose uneasiness and need for external aid for supporting their mental and physical well-being. Different governmental schemes to provide necessary aid for people in despair including the tele counselling options and advisory notices on how to handle the lockdown time were established during this time. However, what people sought mostly to was the availability of internet and smartphones during these times. They engaged themselves being online entertaining themselves, posting and viewing recent updates on social media, video and voice calling over internet and social media and so much more. This takes us to the importance of the internet during COVID – 19 and thereafter.

The public restrictions from the first wave started unlocking from 8^{th} June 2020 and extended till November 2020 when people could gather for funerals,

weddings or other social occasions, although with restrictions on the head count.

With the relaxation in the restrictions the second wave came in, which was not unexpected this time. The COVID – 19 cases went on a surge from the mid of February 2021 in various parts of the country starting from the dense metropolitans, expanding to the suburbs and across the country. During this second wave, a nation-wide lockdown was not imposed, however, the states and union territories were given the right to monitor and handle their situations region-wise so as not to disturb the country's economy as a whole. By the mid May 2021, 35 states and union territories went in for their second wave lock downs which started unlocking relatively sooner, by June 2021.

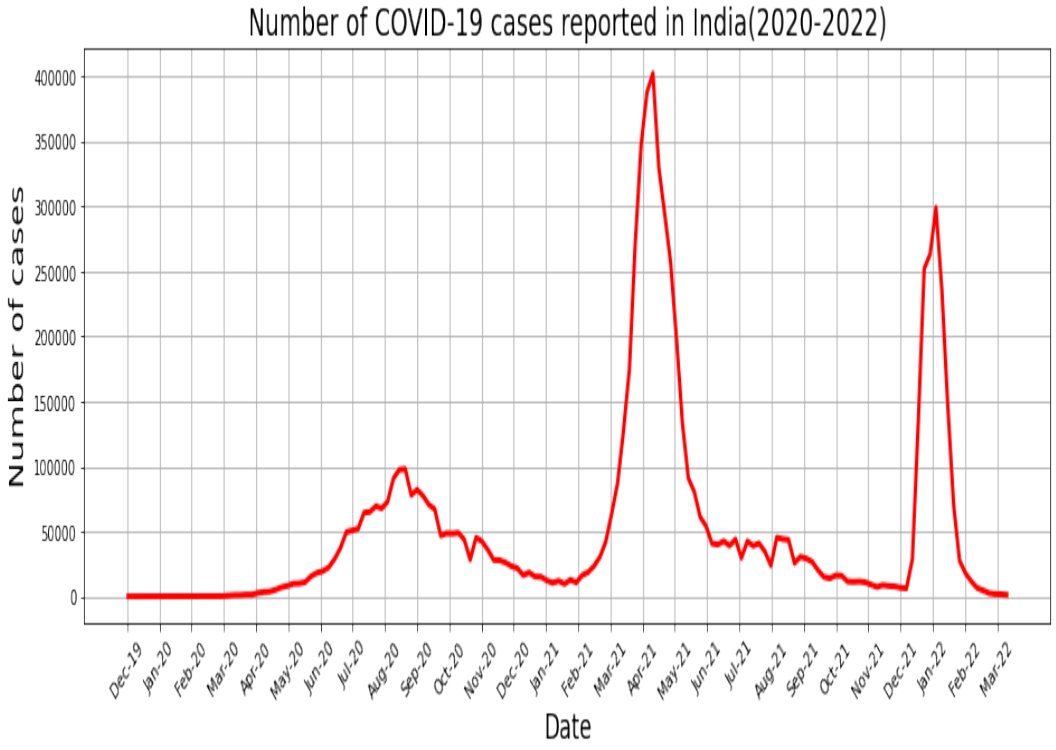

Figure 2.1: COVID - 19 cases reported in India [3]

We can observe that the first wave actually took place after the uplifting of

the lock downs which can be easily understood as the people retreated to their normal lives. However, the second lock downs were a result of the second wave hitting the country. It should be noted that the lockdown was successful declining the cases sharply. The first wave witnessed a broader curve which denotes that the cases could only be gradually decreased because the situation was out of hands as the public restrictions were lifted. It took almost a year (Jan 2020 to Jan 2021) to come back to a minimum number of cases. However, for the second wave, the restrictions were put in place to keep the incline under control and it was successful to a good measure. The second wave was subsided within five months of its appearance (February 2021 to July 2021).

The third wave was observed in December 2021 and lasted for even lesser time (February 2022). Even though the number of cases were almost as high as the second wave, there were no extra precautions or lockdown actions that were implemented during this third wave. We have to realize that people have become more aware and were ready to face this wave because it did not affect the life of the public at all.

The role of infrastructural developments in every field has to be recognized to appreciate this effect completely. In this thesis work, we will be concentrating on the internet infrastructure that would have contributed positively.

2.2 INTERNET IN INDIA

2.2.1 A Brief History

With over 658 million internet users, as of January 2022, with a 47% internet penetration rate, India has a rich history in internet. It all began in 1986 when the Department of Electronics (DoE) and United Nations Development Program (UNDP) combinedly formed the Educational Research Network (ERNET) which provided access to internet among the prestigious educational institutions and research centers in India. When ERNET was dedicated entirely for the educational communities, the public had to wait for another decade to get this access. [4]

On 15^{th} August 1995, the Videsh Sanchar Nigam Limited (VSNL) was officially launched, presently known as the Tata communications Ltd. It provided a dial-up connection worth Rs.100 per hour with a download speed of 9.6 kbps. VSNL used the MODEM technology to convert the analog signals from a cable into digital signals to be processable by the computers and vice versa. Also, it was during this time when the World Wide Web also was getting stronger and widely used which promoted the internet access in India. The web browsers, namely – Mosaic and Netscape Navigator - were the popular web browsers of the time. Later in the same year Gateway Internet Access Service (GIAS) was established in the four major metropolitans – Mumbai, Delhi, Kolkata and Chennai - followed by Bangalore and Pune by the end of 1995. However, VSNL had to experience a massive failure in its first attempt due to unexpected demand that

awaited this launch along with inadequacies in the hardware requirements and network issues. But in six months' time VSNL managed to get more than 10,000 users, which is a success considering the novice nature and the cost of internet in the country in 1995. [5]

In 1997, Integrated Services Digital Network (ISDN) access was provided for public. This allowed exchange of voice, video and other network services along with the texts which was the only possibility earlier. In the same year, ICICI bank started the internet banking for the first time in India. By 1998, the number of internet subscribers reached above 90,000. In 1998, NASSCOM was set up and right to startups on private Internet Service Providers (ISPs) were given by the government. Incidentally, it was the same year the first ever cyber hacking was also reported in India! In 1999, the Railway centralized their reservation system throughout the country.

The new century was marked by legendary events in Internet. The cable internet was made possible for the first time in India. Also, the much need IT Act 2000 came into existence constitutionally. The remarkable social media predecessors Yahoo, MSN and eBay found their base in India also in 2000. The official Indian Railway ticket booking system IRCTC was launched in the same year, followed by airline ticket booking service in 2002. The BSNL began providing Broadband services from 2004. According to the Telecom Regulatory of India (TRAI), broadband connection is defined as "an always-on internet connection with download speed of 256 kbps or above". The speed limit was updated to 256 kbits in 2010. Yet another remarkable event was when Google opened their office in India in

2004. It was followed by Orkut in 2005 and Facebook in 2006.

The cellular internet took its form when the 2G spectrum was first allocated in 2008, followed by 3G in 2009 and Wi-Max in 2010. In 2012, Bharti Airtel initialized providing dongle based 4G services and two years later mobile 4G services. The launching of Reliance Jio in 2016 revolutionized the internet in India. The costs for data were cut-off by the other Internet Service Providers to compete with them and stay in market. Also, railway started to provide free Wi-Fi services known as Rail-Wire in railway stations. Eventually, in 2019 the internet users touched the 500 million mark. Lately, the internet usage trend and the number of users is affected by the COVID – 19 pandemic. The sudden demand of internet for the online schooling and work-from-home gave pressure on the infrastructure and services that were available.

2.2.2 Programs Initiated by Govt. Of India to Improve Internet Service

The government has taken a variety of measures to improve the internet experience available for the citizens. The Digital India program, Bharat Broadband Network Limited (BBNL, or known as Bharat-Net) and National Infrastructure Pipeline (NIP) are some of the projects that are of current importance.

DIGITAL INDIA

Digital India aims in realizing the dream of making India completely a digital community. According to the Digital India website “Digital India is a flagship

program of the Government of India with a vision to transform India into a digitally empowered society and knowledge economy." It aims in bringing in the prosperity of e-governance to the citizens. The Digital India platform provides almost all of the possible government services to be done electronically without visiting a government office or institution. Digital India program brings together a variety of electronic related services to be made available for the citizens. There are mainly nine pillar programs that are complex in itself and has broad implementation strategies, namely - Broadband Highways, Universal Access to Mobile Connectivity, Public Internet Access, Program, e – Governance, Reforming Government through Technology, e – Kranti; electronic delivery of services, Information for All, Electronics Manufacturing, IT for Jobs and Early Harvest Programs. A various applications and websites – Aarogya Setu, UMANG, Aatmanirbhar Bharat, COWIN - were designed under this Digital India program. Responsible AI for Youth is a part of the Digital India program which aims igniting the young minds from school level to understand and pursue the innovative world of Artificial Intelligence. [6]

BHARATNET (BBNL)

The BBNL or BharatNet program focuses on availing high speed broadband internet to all the citizens via the Optical Fiber Cable (OFC) technology. A total of 5,74,332 km of OFC has been laid across 1,84,646 Gram Panchayats as of May 2022. Out of this 53,599 Grama Panchayats have an active Wi-Fi service which sum to a total of 2619.10 TB per month usage of data. National Optical Fiber

Network (NOFN) is the outcome of the BharatNet project. It could connect the states, districts and the grama panchayats in the country. It utilizes the Gigabit Passive Optical Network Technology (GPON). [7]

GIGABIT PASSIVE OPTICAL NETWORK (GPON) TECHNOLOGY

PON is a fiber technology which uses the optical fiber cables to distribute access to internet to its subscribers. The main idea behind this is to use a single transmission line which can be further split into numerous lines on reaching destinations, i.e., a single OFC can be drawn from the internet service provider's hub and it is divided until it reaches the termination or endpoint. Basically, it can be described as point-to-multipoint system. The endpoint is defined depending on the endpoint as – Fiber-to-the-curb (FTTC), Fiber-to-the premise (FTTP) or Fiber-to-the home (FTTH) and Fiber-to-the building (FTTB). At the respective endpoint, the single cable consisting of a wide band of wavelength is split into required sizes using a splitter. This splitting is the characteristic of the Passive Optical Network.

GPON is an upgraded version of the PON technology. A GPON technology includes a shared network equipment in the central office (Optical Line Terminal) and a dedicated optical unit at each subscriber location (Optical Network Terminals) which is connected by fiber using Passive Optical Distribution Network (ODN). This technology allows the transmission and usage of voice, data and IPTV simultaneously. Besides, it supports high bandwidth transmission so as to avoid the bandwidth congestion, long distance service coverage, integrated services at highest standards and a fully optical architecture. GPON allows a fast,

flexible and a multi-terminal fiber deployment at the lowest possible cost. [8]

NATIONAL INFRASTRUCTURE PIPELINE (NIP)

National Infrastructure Pipeline is a five-year long project from the fiscal year 2019 to 2025 which aims in exercising all the necessary steps to provide the best infrastructure in all areas of life thus elevating the quality of life. The finalized report on this project was released on 29^{th} April 2020 keeping in light of the emergency that arose in the form of COVID-19 pandemic. The various sectors in which NIP concentrates include Transport, Logistics, Energy, Water Sanitation, Communication, Social Infrastructure and Commercial Infrastructure. The internet and other telecommunication infrastructure is included in the Communication sector. The government has kept aside $14.81 billion for the telecommunication sector alone. In this sector, various independent projects have been planned and are at different stages. A few among them which were enforced with an immediate effect or accelerated to avail internet to remote areas due to the COVID – 19 lock downs are given below. [9]

- Telecommunication (Fixed Network): There are 25 projects worth $ 5.58 billion that comes under this category
  - Bharat-Net Project: This is the most significant project which we have seen in detail in section 2.2.2.2. The project is worth $ 4.76 billion and began on 25^{th} October 2011 (Planning) and is supposed to be completed by 31^{st} December 2022.

  - Provision of submarine OFC connectivity between Andaman Nicobar island and mainland Chennai: The project is worth $ 125.09 million. The project commenced on 01^{st} July 2018 and is intended to be complete by 20^{th} June 2022.
  - Provision of submarine OFC connectivity between mainland and Lakshadweep islands: The project is worth $ 110.53 million. It was commenced on 09^{th} December 2020 and is planned to be operational by 31^{st} March 2024.
  - Free space Optical communication Network project:

    Project commencement – 01^{st}April 2020;

    Project Completion – 2025 (tentative).

    Project worth - $ 70.57 million. Area of coverage – Andhra Pradesh

- Telecommunications (Telecom Services): It consists of 53 opportunities and is worth $ 8.3 billion.
  - 4G services by MTNL: This project involves establishing 4G service in the states of Delhi and Maharashtra. The project commenced on 01^{st} November 2020 and is supposed to be completed by 31^{st} March 2024. The project is worth $ 1.16 billion.
  - 4G services by BSNL: It is a nationwide project to avail 4G BSNL services. The project was commenced on 01^{st} November 2020 and is estimated to be operational by 31^{st} March 2024. The project is worth $ 3.76 billion.

- Provisions for mobile services in uncovered remote villages:
  - Arunachal Pradesh and 2 districts of Assam:

    Project Commencement - 09^{th} December 2020;

    Project Completion - 31^{st} March 2024.

    Project worth - $ 133.91 million.
  - Meghalaya: Project commencement – 04^{th} September 2020;

    Project completion – 31^{st} March 2024;

    Project worth – $ 48 million.
  - 502 villages of 4 districts of Bihar, Madhya Pradesh, Uttar Pradesh and Rajasthan:

    Project Commencement – 01^{st} October 2020;

    Project Completion – 31^{st} March 2024;

    Project Worth - $ 28.52 million.
  - 354 villages in Jammu & Kashmir and Ladakh along the borders:

    Project commencement – 28^{th} April 2020;

    Project Completion – 31st March 2023;

    Project worth -$ 22.25 million.

The National Infrastructure Pipeline is a promising project for the remote regions of the country which are not accessible easily. Along with the communication sector, the project has an Information Technology sector which concentrates on the software requirements and data center establishments. It is worth $ 489.69 million and is promising for the advancement of the IT sector which is equally

essential to improve the overall quality of life of the citizens.

The different projects planned and executed by the government supported by different Internet Service Providers provide improved conditions for the citizens at a lower cost and most importantly at a higher standard.

Chapter 3

AN ENVISION ON 5G TECHNOLOGY

3.1 CELLULAR SPECTRUM

Electromagnetic spectrum consists of radiation ranging from the extremely low frequency radio waves (3kHz to 300GHz) to the extremely high frequency gamma waves (greater than 1019 Hz). The cellular communication utilizes the radio frequency range for transmission of signals. The cellular spectrum denotes the frequency range or the bandwidth allocation provided for different purposes and to different operators. An efficient allocation of this vast band is required to make sure that the spectrum is utilized to its full potential and no wastage of frequency happens. The cellular spectrum allocation is done on different levels starting from the international level, which is further reallocated nationally to be allotted for different operators and different type of connections. The radio-

frequency ranging from 3kHz to 300GHz has to be allocated and monitored. This is taken care by the International Telecommunication Union (ITU). In this frequency range, the cellular spectrum is in the range between 600MHz to 39GHz. The ITU-Radio-communication Sector along with the Radio-communication Bureau is responsible for the management of the frequency spectrum and satellite orbit resources. Apart from the responsibility to allocate the spectrum, the ITU-R is also in-charge of monitoring the regulations and make sure that all the users abide by the Radio Regulations (RR). They are an international treaty which determines how the spectrum allocation is done, how the radio services, satellite services, space research and Earth explorations can be coordinated, how the equipment specifications should match globally, for example, the operating frequency of the ac appliances is set to be from 50Hz to 60Hz globally so that the appliances made will not be region specific, but can be used anywhere. [10]

The cellular spectrum allocation is based on the technology that is being employed. A 2G technology will not require a broad spectrum as required by 3G and so on. The better the technology the bandwidth of the frequency required will be higher as more data can be sent and received at once. Besides, the spectrum is further allotted differently for the up-links (from cell phones to towers) and down links (from towers to cell phones). The cellular carrier should possess license to both the up-link frequency bands and down-link bands to make possible a faster two-way communication, where at one end the user sends the message (up-link) and at the other end user receives the message (down-link).

Now, we are in an era where the internet does not just contain the mobile

phone users. We have a broad range of objects which come under the Internet of Things (IoT) category that requires high speed internet without any delay and instantaneous connectivity to the server for efficient functioning. The 5G technology comes with this ability to provide maximum efficiency in terms of speed and time unbound connection. Interestingly, unlike the other technologies, 5G bands are scattered across the radio wave spectrum to make it available for different purposes. It is categorized into low band (600MHz to 1GHz), mid-band or C-band (1GHz to 6GHz) and high-band or mm Wave (29 GHz to 39GHz).

Apart from dividing the spectrum into different bands, a guard band has to be allotted in between different frequency bands to make sure that the close frequencies do not interfere resulting in a chaotic transmission. Other than this, different time slots can be provided for different signals.

The Frequency Division Duplex (FDD) is the technique in which several message signals are sent via a single band by allotting a frequency band for each signal which is separated from the adjacent signals using a guard frequency. This makes sure that the two signals do not interfere. Along with its clarity in the data transmitted and received, the FDD makes sure that the communication is much faster as different bands are used for the up-link and down-link.

The Time Division Duplex (TDD) is a converse technique in which the signals are sent at a same frequency, however, the rate of transmission is maintained by allotting different time slots for the signals. The time delay between the transmission and the receiving of the signal, known as latency, is relatively lesser for the TDD technique. The consecutive signal transfer and any inefficiency in syn-

chronization of the up-links and down-links can result in interference or even data loss. However, due to its higher speed (lower latency) in data transmission, the 5G technology is based on the time duplexing. With proper synchronization systems and precise timing it is possible to achieve efficient outcome.

3.2 SPECTRUM ALLOCATION IN INDIA

Initially, when the voice call technology came into use, the cellular spectrum devoted was just of bandwidth 800MHz. It was mainly used for voice communications. The Code Division Multiple Access (CDMA) in India utilized 20MHz of this 800MHz available 2G across the world. The Global System for Mobile Communication (GSM) was allotted 25MHz of the 900MHz band and 75MHz of 1800MHz. These two technologies, CDMA and GSM, were the 2G technology implementation techniques. Further, when the number of cell phones users increased, this was no longer sufficient. Also, 3G came into existence which utilized the 2.1 GHz band. However, with the 3G, revolution took place in the internet with the band requirements for video streaming came into picture. These bands could no longer handle the ever-increasing demand for frequencies. It was during this time of crisis, the microwaves in the range of 2GHz to 4GHz known as the S-band (Satellite - band) was of particular interest due to their ability for satellite communication and radars. It was allotted by the ITU – R for terrestrial mobile communication services in 2000. This has been used by the weather forecasters and communication satellites ever since. The 2.5MHz band in this

S-band was given to India Space Research Organization (ISRO) by the Wireless Planning Commission (WPC) under the Department of Telecommunications for mobile satellite services (MSS), however, was left unused for years. The demand for this valuable band increased and the mobile operators constantly reminded of how the 3G implementations could be bettered utilizing the 2.5GHz to 2.69 GHz band. Later on, the scope for satellite phones were outrun by the 4G technology. So instead, 2.5GHz band of the S – band was allotted to 4G. Further, 150MHz of this band belongs to ISRO and in 2009, BSNL and MTNL were given 20MHz each of S-band to enable wireless government services. This, however, was a tragedy as the 2.3GHz band provided to the network providers were not authorized by the ITU and was not compatible for the available devices. [11]

With the emergence of the smartphones, more frequency bands for 4G were allotted. These bands include 800MHz, 900 MHz, 1.8GHz, 2.1GHz and 2.3GHz. The lower frequencies were used for transmission of 4G/LTE signals across rural and rural areas due to their longer wavelengths.

Currently, the auctions for the 5G spectrum allocations are under-progress. According to TRAI resources, the whole mobile communications spectrum including some new bands are to be put up for auction. These bands are 700MHz, 800MHz, 900MHz, 1.8GHz, 2.1GHz and 2.3GHz among the existing. And 600 MHz, 3.3GHz – 3.67GHz and 24.25GHz to 28.5GHz which are to be allotted for the first time.

3.3 5G AND ITS RELEVANCE

5G or the Fifth-generation is the latest cellular connection technology which promises fastest speed signal transmission and reception with ultra-low latency, greater capacity or higher bandwidth and most of all, better reliability. The theoretical assumptions show that 5G could provide a minimum download speed of 1GHz and a latency lesser than 1 millisecond. The highlight of the 5G technology is that it could make the IoT a global success. Since 5G technology could connect more devices with greater bandwidths the IoTs could be operated more efficiently without any noticeable latency.

The previous generations were mostly used just for uploading and downloading data in the form of text, voice or video, however, 5G technology is capable uniting devices that consists of embedded sensors (IoT) to other devices which could utilize the data retrieved using the sensors. This requires high accuracy of synchronization as the functioning of other devices could be dependent on the latency of the data reception from the sensors. For example, if we consider a self-driving car, we have to make sure that the car sensor senses a barrier, sends the message to the system which recognizes the barrier and returns a message to do an action like braking or turning. If the latency between the transmission of sensor message and reception of response message is longer than the time taken for the car to run into the barrier, then the technology will be a failure. This is where the high efficiency provided by the 5G is highly relevant.

The 5G is categorized into three depending on the frequency bands utilized

by them –

- Low-band 5G: It is the range of frequencies below 1GHz spreading to low frequency waves of 600MHz. This provides speeds almost in the range of the 4G. This is used for enabling 5G connection across a country which can make sure that the connectivity is available even in the remote and rural regions.

- Mid-band 5G and C-band: The frequency band ranging from 1GHz to 6GHz is referred toas the mid–band 5G. This range of frequencies offer faster network speeds than the low-band 5G and can cover greater area than the high-band. The mid-band satisfies the optimal conditions that could provide the best service in terms of speed, range, coverage, penetration and capacity. This band can be utilized in regions where the demand is too high. The band of frequencies from 2.4GHz to 5GHz, known as the C – band is the frequency band that could provide the much greater speed and coverage, and is therefore, the most favorable band for 5G technology.

- High-band 5G or mm Wave: The high-band 5G consists of high frequency radio waves ranging from 24GHz to 39GHz. With very high frequencies, these frequency bands can transmit data much faster and with negligible latency. However, with higher frequency, the ability to penetrate the usual obstacles like trees and buildings will be diminished. As a result, high speed transmission could be made possible only for a short distance.

The 5G technology could revolutionize the already internet-dependent world.

The ability to connect with objects that could be of daily relevance and the ever-increasing demand for better speed and reliability on the internet services make this technology something to be looking forward to.

3.4 IMPLEMENTATION OF 5G IN INDIA

According to the Telecom Regulatory Authority of India (TRAI), the spectrum for 5G is all ready for distribution among the network providers. All the frequency bands that are available for India as per ITU – R allocation would be open for auction. This includes the existing bands that are already in use (700MHz, 800MHz, 900MHz, 1.8GHz, 2.1GHz and 2.3GHz) along with a few bands (600MHz, 3.3GHz – 3.67GHz and 24.25GHz to 28.5GHz) made particularly available for 5G technology.

The preparations to roll out the 5G services begin with the auction which allots different frequency bands for the network providers. It is forecast that the services will be made available from later this year 2022 or latest by early 2023. There are several collaborations expected from foreign nations like Australia and Japan who have already implemented the technology. Within the country, the auction will be of great importance because a better band will be required for better quality of service and 5G is all about making the online experience better. Once the auctions are done, it is expected that the services will be rolled out in a time of six months or even quicker. Many service providers have already conducted their trial and tests on the 5G technology to be implemented. This test

was conducted on the specially allocated bands exclusively for the research purpose. This would accelerate the implementation speed of the actual 5G services once the spectrum allocation is documented.

Many public trials were conducted by the leading network providers in the country. Airtel has claimed to have completed the cloud gaming trial and rural 5G test already. Reliance Jio has come up with connected drones that used 5G technology. Further, the company was successful in integrating energy utilities through sensors. Vodafone-Idea has also announced their 5G trials in different areas which includes smart cities, healthcare, education, agriculture and cloud gaming. These network providers use external collaborations from different firms to conduct these trials which involve knowledge in various domains.

It is evident from the above discussions that 5G technology is at our doorstep and it is just the matter of a few months until when we will be connected to our cars, homes and offices in a way that our presence is not relevant anymore. The current delay for the auction is due to the extension in trial period offered by the government on the service providers request which would be end on 26th November 2022. Moreover, there are concerns raised by the service providers on the high base amount set for the 5G spectrum. In addition to that, there are concerns related to how effective will be 5G utilities in the Indian market. The network providers are not confident that the users are capable of utilizing the enormous opportunities that come up with this technology and worry about their losses if it does not kick in as per expectation. Combining all these concerns, there is a possibility that the bid prices may not go up as per the government's

expectations, which also could be a factor that delays the auction. Despite all these uncertainties, it is quite sure that 5G technology could not be delayed any further and is ready for disposal.

Chapter 4

STATISTICAL AND MATHEMATICAL METHODS

4.1 EXTRAPOLATION

Extrapolation is the method in which an existing curve is extended by predicting the unknown points. For a purely linear function, we can assume to a high degree that the extrapolated point will be the original point. However, with the practical data, which is always chaotic, the accuracy will decline resulting in completely different trends. Basically, any linear graph can be extended using the function that defines it.

A linear function can be given by –

$$y = mx + c$$

Assume that

$$(x_1, y_1) \tag{4.1}$$

and

$$(x_2, y_2) \tag{4.2}$$

are two known points on the linear graph as shown in Fig.4.1.

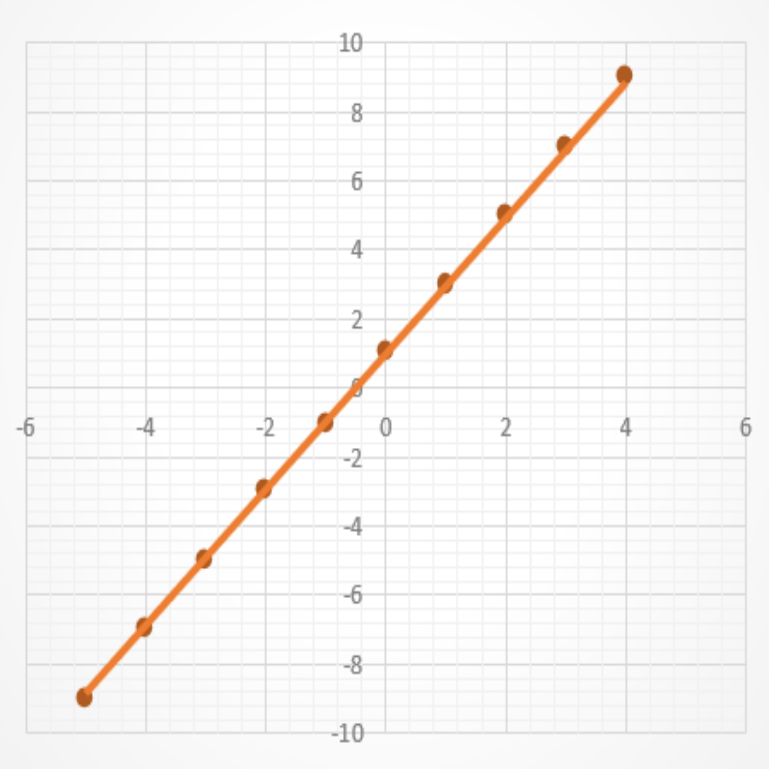

Figure 4.1: A simple linear graph

The slope (m) of this graph and the y-intercept (c) can be calculated as –

$$Slope, m = \frac{y_2 - y_1}{x_2 - x_1} \tag{4.3}$$

$$y = mx + c \tag{4.4}$$

$$y - intercept, c = y, atx = 0 \tag{4.5}$$

From Fig.4.1, Let $(x_1, y_1) = (3, 7)$ and $(x_2, y_2) = (4, 9)$

Using (4.4) and (4.5), m = 2 and c = 1

Now, we get the function for the graph as –

$$y = 2x + 1 \tag{4.6}$$

With this function we can find any point for any given x and extend the graph along both the directions (positive and negative x-values). It is illustrated in Fig.4.2.

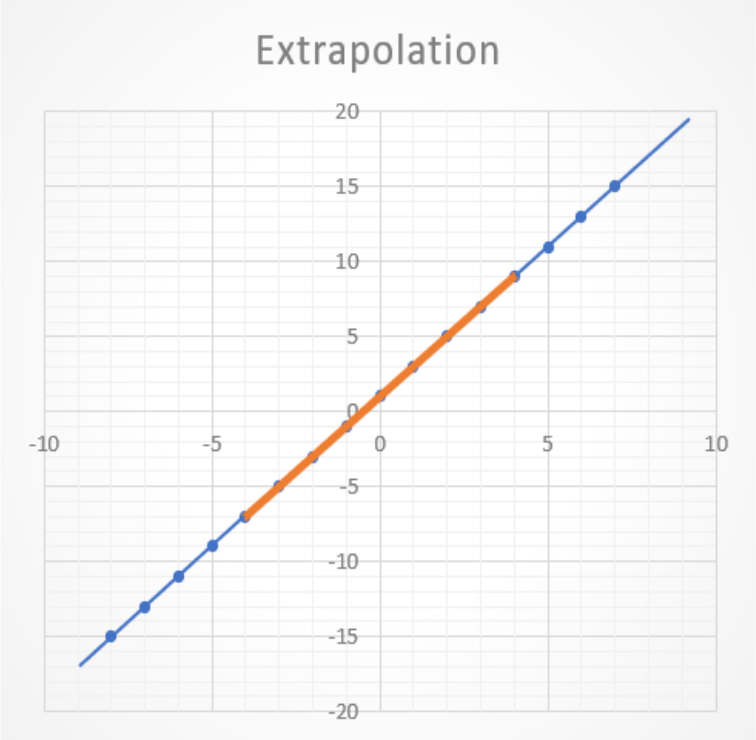

Figure 4.2: Extrapolation of Linear graph

In this project, the extrapolation is done on different parameters with respect to the date axis. The extrapolation could give a insight on the difference between the predicted trend and actual trend. Moreover, most of the real-life data sets consists of complex functions with higher degrees of polynomials. It is impossible to predict the actual value unless all the factors that influences the trend is properly studied and included in the function polynomial, which is also not possible.

Consider a polynomial trend observed as shown in Fig.4.3.

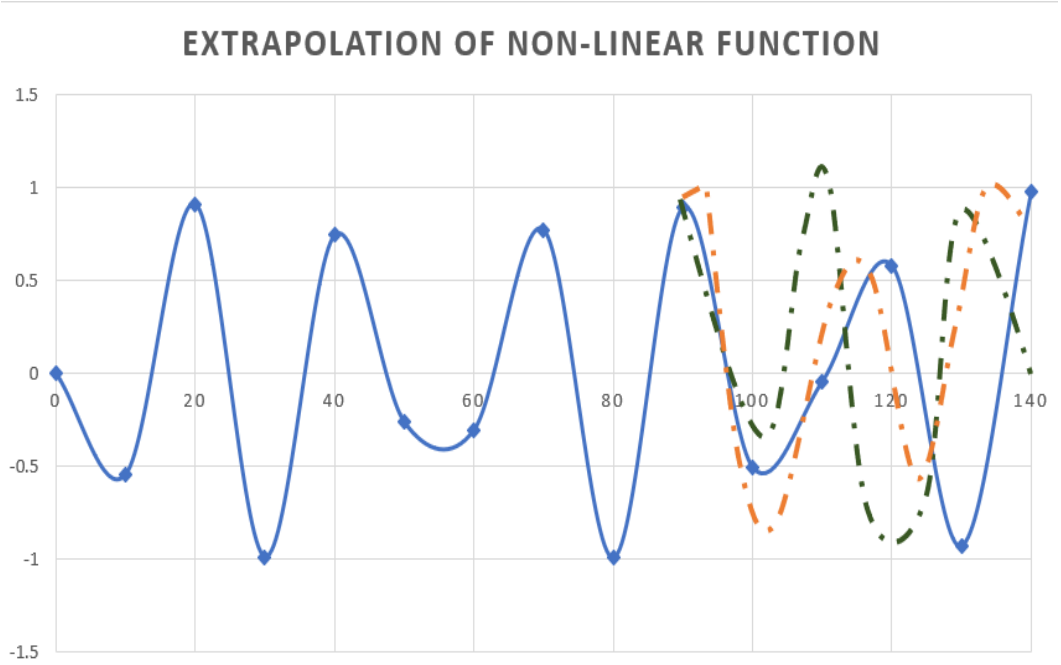

Figure 4.3: Extrapolation of a non-linear graph

The green and orange dashed lines show two possibilities how this sinusoidal type non-linear function could be extrapolated. There are such infinite possibilities to extend the graph from a given point and therefore, it is not an apt technique for analysis when it comes to non-linear functions. In this work, only

nearly linear functions are extrapolated to observe the change in trends. Also, the graphs are extrapolated linearly using InterpolatedUnivariateSpline () function from the Scipy.interpolate library which could give a smoothened spline curve.

4.2 SPLINE CURVES

All the real-life data when plotted and joined to get a graph will end up with corners and sharp edges. Such graphs are difficult to interpret and no curve fitting techniques could be implemented on the data set. In order to avoid this problem, the graphs are 'smoothened' using mathematical tools. The plotting of spline curves is such a technique adopted for most graphic purposes. Usually, smoothening of the curves is done by approximating infinite number of infinitesimal lines that are connected so as to smoothly join two points. However, using the spline curves, we assign a few knots or control points which are connected by a function.

Spline curves are basically nothing but polynomial functions that can give smooth curves. The most widely used polynomial is of the third degree or the cubic polynomials for graphical spline curves. There are two reasons behind this choice of cubic polynomials – it is the lowest degree polynomial which can produce inflection in graphs, i.e., change in the concavity of the graph or the trend of the graph reverses. Mathematically, it is the point where the second derivative of a function is zero such that two points taken on either sides of the graph change their sign from positive to negative or vice versa. Also, the cubic

polynomials are smooth naturally making them the best option for smoothening (Fig.4.4). [12]

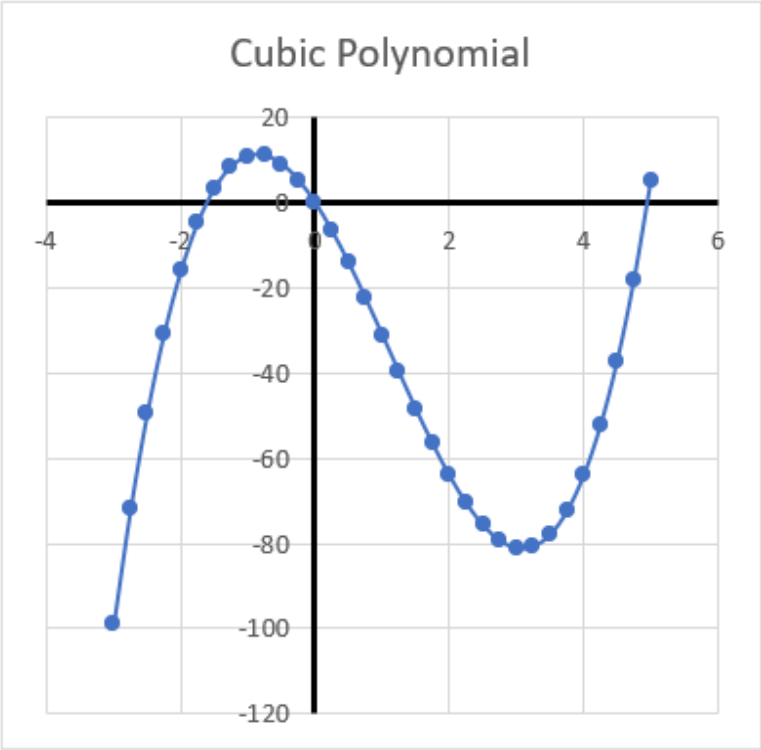

Figure 4.4: Cubic Polynomial

The cubic polynomial constitutes of four coefficients a, b, c and d in the polynomial as –

General Form:

$$y = ax^3 + bx^2 + cx + d \tag{4.7}$$

The four coefficients can be solved from four linear equations which are obtained by applying constraints depending on the function. For example, in the above graph when

$$Point(x, y) = (0, 0), d = 0 \tag{4.8}$$

$$Point(x,y) = (1,-31), a+b+c+d = -31 \tag{4.9}$$

$$Point(x,y) = (-1,11), -a+b-c+d = 11 \tag{4.10}$$

$$Point(x,y) = (5,5), 125a+25b+5c+d = 5 \tag{4.11}$$

This can be expressed in matrix form and the equation for the solution can be obtained.

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{bmatrix} Y = \begin{bmatrix} 0 \\ -31 \\ 11 \\ 5 \end{bmatrix} X = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \tag{4.12}$$

The solution for the coefficient matrix is given by –

$$X = A^{-1}Y \tag{4.13}$$

These four points of matrix X act as the knots or the control points that connect two data points using a cubic polynomial function.

As the complexity of the data points increases, a simple cubic polynomial would not suffice the actual smoothening by taking into consideration of the all the data point. In such cases, piece-wise polynomial curves technique is utilized. In this method, we mainly account to the fact that most data points will be connected by more than one inflection points. Using higher degree polynomials might be a convenient method, however, most of the times this would result in

sharp graphs rather than the required smooth ones. Therefore, we use different cubic polynomial curves to join the knots rather than using a single function. Each control point will act as a boundary for two different cubic polynomials and thus create a much accurate and smoother curve.

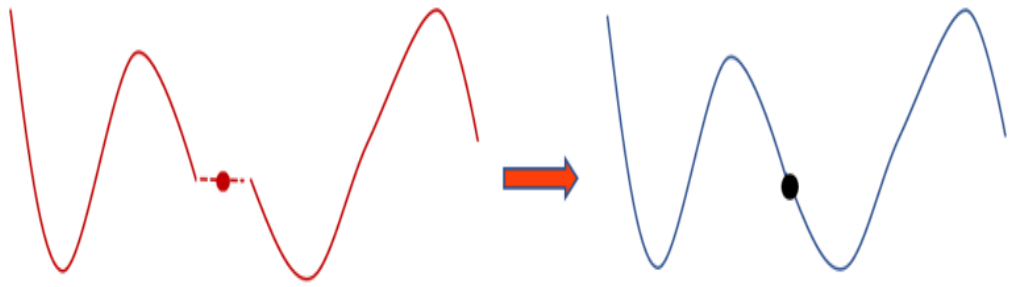

Figure 4.5: Connecting control points using cubic polynomials

As clear from the Fig.4.5, each point is connected by different cubic polynomial curves. However, the solution will require too many coefficients. For one cubic polynomial we have seen that we require four coefficients that could solve it. A real data will have a large number of data points which will require even more control points and therefore four times as many coefficients as the control points. We will have to do the previous exercise of finding the X matrix using $X = A^{-1}Y$ formula for each of the polynomial curve. This is done easily and effectively by the computers. For example, 10 control point graph will have 9 cubic polynomials that join them. In such a case, we will obtain a 36×36 matrix for A, a 36×1 matrix each for X and Y.

Along with this, the piece-wise curves that are joined at the control points should satisfy the conditions for continuity. Each point should follow three con-

tinuity conditions, namely –

- Continuity of the values, i.e., the control pint values at the boundary of the two cubic polynomials should be the same. This is the zeroth order differential continuity.
- Continuity of the slopes, i.e., The slopes of the two cubic polynomial curves should be same at the control point. This is the first order differential continuity.
- Continuity of the curvatures, i.e., the concavity of the control point should be matching as obtained from both the cubic polynomial curves. This is the second order differential continuity.

In order to satisfy these continuities, apart from solving the linear equations we need to parameterize each of the piece-wise curves. Most of the times, the parameterization is such that the variation of the function with respect to the variable lies within a limit. This is the basic spline curve computation technique. However, using the above approach is dependent on one set of linear equations and is controlled by all of the control points. This reduces our freedom to change individual or local points and observe the change in the curves. The interpolation of the control points can give us the access to this local control on the shape of each curve. There are many methods used for this purpose which includes Hermite cubic splines, Catmul-Rom splines and Cardinal splines.

In this work, the interpolation in done to obtain more accurate spline curves. The SciPy InterpolatedUnivariateSpline function is utilized. This function has

parameter 'k' (order of the polynomial) which can be manually set to any order from 1 to 5.

4.3 CORRELATION

In any data analysis, the relation between two variables plays an important role in determining the dependent factor and its relation between the two variables. Correlation is a statistical method to quantify the relation in the trends of two variables. The correlation indicates how two variables vary with the independent unidirectional variable (e.g.: time, date).

4.3.1 Correlation using covariance

Correlation coefficients give us the quantitative measure of the linear relationship between the variables. Usually, the coefficients lie within –1 to +1.

- A zero for correlation coefficient value implies that the two variables do not have a linear relationship.
- If the coefficient is +1, it implies that the two variables are perfectly positively correlated, i.e., with increase in the first variable, the second variable also has a similar increase. (Fig.4.6)
- If the coefficient is –1, it implies that the two variables are perfectly negatively related, i.e., when there is an increase in the first variable, the second variable experiences a decrease in the similar rate. (Fig.4.7)

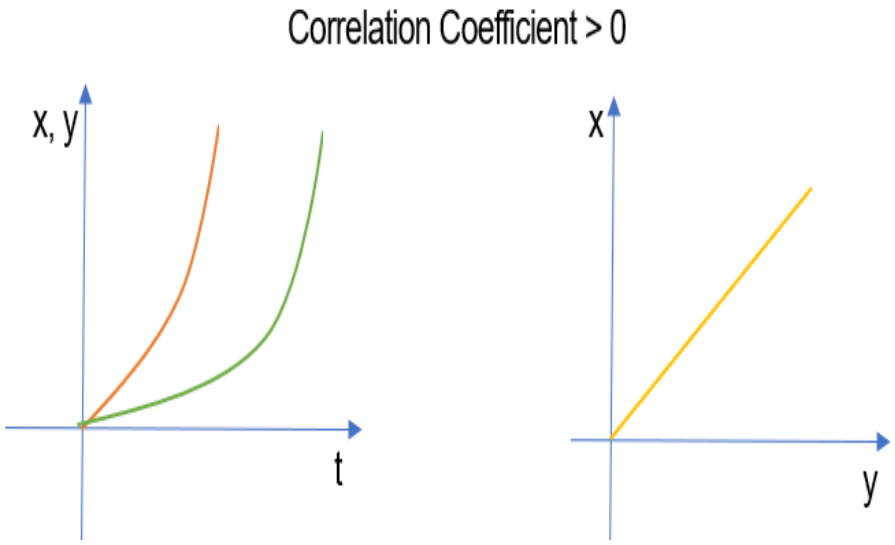

Figure 4.6: Positive correlation

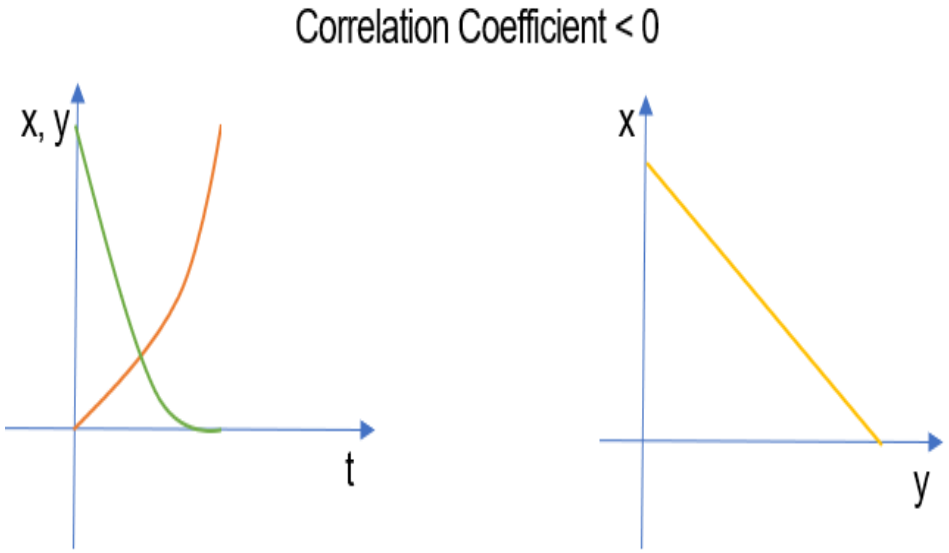

Figure 4.7: Negative Correlation

Positive values for coefficients denote that the two variables have a similar trend and negative values for coefficients denote that the two variables have opposite trends.

However, it should be noted that the correlation only gives an idea on how the

trends of two variables are related and they do not give any physical or mathematical relation between the two variables. More precisely, the correlation does not infer to causation. Even when two variables might seem to have the same trend, we cannot prove in anyway that the one of them causes the other or vice versa. Both the variables will be considered independent to each other. Statistically, the correlation is defined as the ratio of covariance of the two variables to the product of their standard deviations. Covariance of two variables is defined as the direction of relationship between the variables, i.e., how the two variables deviate from each other. Covariance shows the deviation whereas the correlation shows the relation between the variables. Covariance is different from variance. The latter determines the deviation of the values of a single variable from its mean value whereas, the former determines how two random variables deviate from each other. We know that when we consider two random variables, their values can be in different ranges, but this is nullified as the covariance is calculated by subtracting each of the value from its own respective mean value. Covariance of two variables x, y is given by –

$$Cov_{x,y} = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{N-1}$$

where,

$cov_{x,y}$ is the covariance of x and y

$\overline{x}$ and $\overline{y}$ are the mean of the variables x and y

x_i and y_i are the values of the variables x and y

N is the number of samples or the population

The standard deviations of the two variables are obtained as –

$$\sigma_x = \sqrt{\frac{\sum_{i}^{N}(x_i - \overline{x})^2}{N-1}} \tag{4.14}$$

$$\sigma_y = \sqrt{\frac{\sum_{i}^{N}(y_i - \overline{y})^2}{N-1}} \tag{4.15}$$

σ_x and σ_y are the standard deviations of x and y respectively

The correlation between the variables is given by –

$$Correlation = \rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} \tag{4.16}$$

There are several data – specific methods to compute the coefficients of correlation. In this work, the Pearson Correlation coefficients and Spearman's coefficients are used.

4.3.2 Pearson Product – Moment Correlation

The Pearson product – moment correlation coefficient, or simply, Pearson correlation coefficient is the quantitative measure of the linear association between two variables. It is the technique adopted in fitting a linear graph for the data points of the two variables. The correlation coefficient indicates how well does the individual data points fits to this best fitting line. The Pearson's correlation

coefficient ranges from –1 to +1 (Fig. 4.8). The interpretation for the coefficient is same as that of the normal correlation as explained in the section 4.3.1. [13]

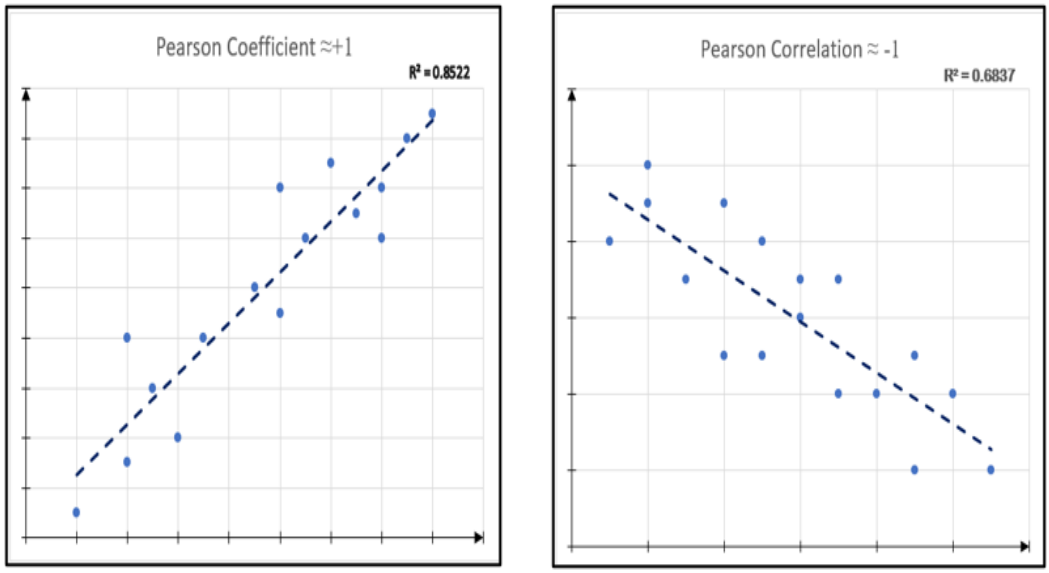

Figure 4.8: Pearson correlations

Even though this sounds as an easy approach in associating the linear relation between two variables, the Pearson method cab be applied only to variables which obey a certain conditions or assumptions. They are –

- The variable data type: Both of the variables under consideration must be of interval type or ratio type data. Interval data is defined as a continuous variable which has values at equal intervals. In this work, all the data are equally spaced along the time axis (Quarterly or monthly data). Ratio data is also a continuous variable with values at equal intervals, moreover, it has a true zero, i.e., the particular variable is absent at that point.

- Paired variables: The two variables should have corresponding values for each other. Data points of the variables should be an ordered pair of data

from each variable.

- Independence of cases: The rows of the variables should be independent. There should not be any relation within the rows or the data of the variables. If such a relation exists within the variable, then Pearson coefficient is not useful.

- Linear relationship: Pearson coefficient gives the linear association between two variables. Therefore, it should be noted that the variables have data points that show a linear trend and not any non-linear trend.

- Homoscedasticity: This is the property of a scatter plot in which the data points have a constant variance from the best fit. Pearson correlation is applicable only for such data points. The opposite of this case is the heteroscedasticity.

- Outliers: The presence of uni-variate or multivariate outliers could affect the reliability on the Pearson's correlation. Uni-variate outliers are those values which lie outside a variables' usual range. The multivariate outliers are those data points which shows deviation from the usual trend shown by two variables. It can be such that one data point of the two variables deviates from the linear fit too much than the others in the same data set. It is advisable to remove the outliers so as to get a better insight using the Pearson.

The Pearson coefficient is calculated using the formula –

$$r = \frac{N(\sum xy) - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \tag{4.17}$$

Where, r is the Pearson Correlation Coefficient,

x and y are the variables,

N is the number of samples in the population.

The coefficient of determination is the square of the Pearson correlation coefficient r. This quantity gives the measure of variation present in the model. This can be either represented as a proportion as in Fig. 9 or as percentage.

4.3.3 Spearman's Rank – Order Correlation

Spearman's correlation is the quantitative measure of the strength and direction of the monotonic relationship between two variables. Monotonic variables are those which have a uniform trend throughout, i.e., the data points are either increasing with the interval or decreasing but not both. However, unlike the Pearson's correlation, the Spearman's does not require the data points to be linear. The Spearman's correlation determines how much similar are the trends of the two variables given that the data points form a monotonic graph. Therefore, Spearman's correlation is more flexible with the real-life data sets as we can include non-linearly associated variables also. Moreover, Spearman's can be calculated for the ordinal data also along with interval and ratio data types. [14]

The Spearman's rank – order coefficient can be calculated using the formula –

$$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{4.18}$$

Where x and y are the variables;

$\overline{x}$ and $\overline{y}$ are the mean of the variables.

Spearman's Correlation computes the association between the two variables by ranking them separately and find the correlation between them. If we consider an example of a class of students whose scores in the language papers are to be correlated. Firstly, the students are ranked from 1 to N for each language. Then, the correlation between these rankings is calculated. Consider an individual student, if he/she has a rank 2 in paper A and rank 2 in paper B, we can say that this student has a good aptitude in languages. This is the general idea and it is employed over the whole data set to get the Spearman's Correlation. The Spearman's rank – order coefficient can take values from –1 to +1 and is defined the same way as the other correlation coefficients.

4.4 SINGULAR VALUE DECOMPOSITION (SVD)

Singular Value Decomposition or SVD is a type of decomposition method applied on matrices to study their properties. The basic decomposition method is the eigen decomposition which is applicable only to square matrices. SVD, on the other hand, is applicable to any rectangular matrix $(m \times n)$, given that $m \gg n$. This is the reason of the why SVD is more widely used as the data we come across is mostly in rectangular shape. [15] [16]

The SVD is defined mathematically as –

$$A = U \sum V^T \tag{4.19}$$

Where U ($m \times m$) is the left orthogonal unitary matrix of A;

$\sum(m \times n)$ is the diagonal singular matrix of A;

V ($n \times n$) is the right orthogonal unitary matrix of A;

A is the data set matrix of dimension $m \times n$

The SVD decomposes the matrix A into three matrices by definition. In matrix notation, it is expressed as –

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}_{(m\times n)} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{m1} & \cdots & u_{mm} \end{pmatrix}_{(m\times m)} \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \end{pmatrix}_{(m\times n)} \begin{pmatrix} v_{11} & \dots & v_{n1} \\ \vdots & \ddots & \vdots \\ v_{1n} & \cdots & v_{nn} \end{pmatrix}_{(n\times n)}$$

- The matrix $\sum$ consists of the singular values or the square root of the eigen values of AAT arranged in descending order along the diagonals.

- The unitary matrices have the property of obtaining an identity matrix when multiplied by its own transpose. i.e.,

$$UU^T = U^T U = I \tag{4.20}$$

$$VV^T = V^T V = I \tag{4.21}$$

SVD is a technique in which a given matrix A is represented by three other matrices. These matrices are actually the transformations that are done on a group of vectors which are the columns of the data set. In SVD, a group of vectors

are transformed to another group of vectors which are orthogonal to the initial vectors and are normalized. Assume that V is the matrix that represents the initial set of orthogonal vectors and U is the matrix that represents the transformed orthogonal vectors. Let A be the transformation acting on V that converts it into U.

$$U = AV \tag{4.22}$$

Now, in order to orthonormalize U, multiply it with the singular matrix .

$$U \sum = AV \tag{4.23}$$

$$A = U \sum V^{-1} \ \because VV^T = I \tag{4.24}$$

$$A = U \sum V^T \tag{4.25}$$

From the above steps it is clear that A is the dataset that is obtained from transformation between the mutually orthogonal vectors. So, using SVD we obtain the initial condition of the data (V) and the transformation that is responsible for getting the matrix A, i.e., U. The singular value decomposition is done in python using the NumPy library. However, in this work, SVD is calculated for the covariance matrix of the given data so that we can include the correlation within this decomposition.

4.5 PRINCIPAL VALUE DECOMPOSITION (PCA)

Principal Component Analysis or PCA is a dimensionality reduction method which simplifies the understanding of the data set. A large data set consists of

various number of variables which could be related to each other or not. PCA helps us reduce this n-dimensional data into lesser dimension of our convenience. In doing this, we are reducing the probability of over-fitting the data as too many variables could end up in perfect fittings which may not be applicable for a future data because those data might depend on some new features. Basically, in PCA we obtain the vector directions which gives the maximum variance.

There are two different methods by which we can reduce the dimension of a data set. First is the feature elimination method – it involves the manual removal of the variable from basic analysis or correlations. This method is simple and could maintain the interpretability of the remaining variables. However, the contributions done by the eliminated variables will be discarded. Second is the feature extraction method – it involves defining new variables which has inherited properties of the previous variables. The new variables are defined such that the important characteristics from all the old variables are extracted. As a result, we can remove the new variables which are of least relevance without affecting the data fitting. PCA uses this feature extraction method. [17] [18]

PCA is used in cases where the number of variables is to be reduced without losing the significance of the values, i.e., feature elimination is not favored. PCA also makes sure that the variables are independent of each other. The first step in Principal Component Analysis is the standardization of the data set. This makes sure that the contributions from all the variables are treated equally. The standardization of the data is done using a custom defined function which did

the following transformation to the variables –

$$x_{new} = \frac{x - \overline{x}}{\sigma}$$

Where $x_{new}$is the standardized value of variable X;

x is the value of the variable X;

$\overline{x}$ is the mean of the variable X;

σ is the standard deviation of the variable X.

Once the standardization is done, the covariance matrix for the data set is computed. It gives the idea on how the variables of the data set vary about their mean and the relation between the variation in different variables, i.e., the correlation is obtained. The diagonal elements of the covariance matrix give the variance of each variable, whereas, the non-diagonals are symmetric about the diagonal and each element gives the covariance between two variables.

With the covariance matrix, the eigenvectors and eigenvalues of this matrix are calculated. Each eigen vector of the covariance vector gives the direction of maximum variance. These axes are the Principal Components of the data set. The eigen values, on the other hand, are the amount of variance, i.e., the magnitude of the Principal Components. Therefore, the eigenvectors and eigenvalues of the covariance matrix together give the magnitude and direction of the Principal Components.

All the principal components are independent of each other and therefore, they will be perpendicular to each other.

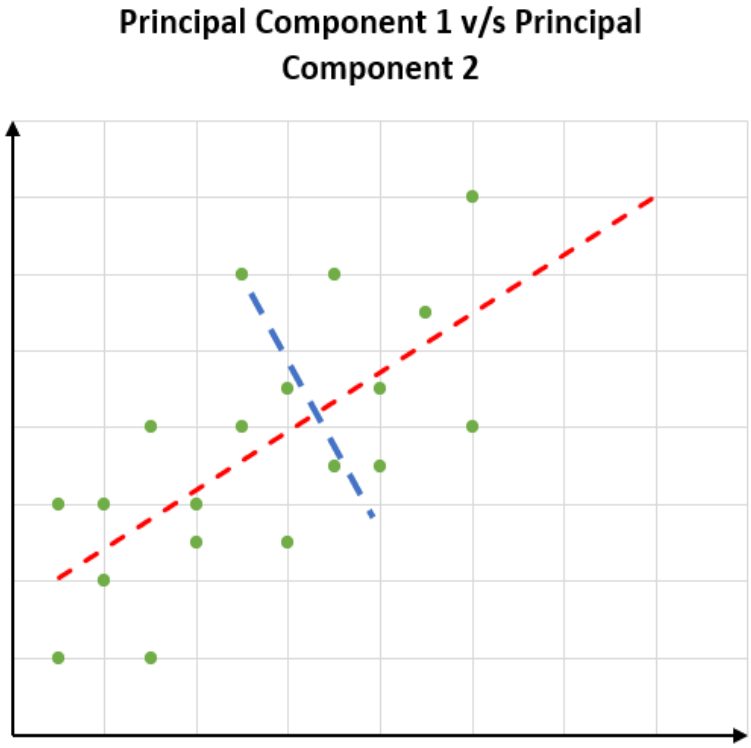

Figure 4.9: Principal Components graph

In this work, the PCA is done using the eigenvalues and eigenvectors obtained from the Singular Value Decomposition as explained in section 4.4. Further, the graph is plotted between the significant principal components.

4.6 LöWDIN ORTHOGONALIZATION

Orthogonalization is the process in which a non-orthogonal set of linearly independent functions is transformed into an orthogonal matrix with orthonormal bases as rows. An orthogonal matrix is defined as matrix with orthonormal bases as its rows, i.e., each row represents one orthonormal base of the matrix representation, such that $AA^T = A^TA = I$. This also applies to the columns of the matrix.

There are different methods of orthogonalization employed depending on the requirement. Gram – Schmidt process is the widely used orthogonalization process. In the process, a matrix of any basis is converted to an orthogonal matrix using the general formula –

Let **V** be a vector space with an inner product with x_a the basis of **V**

$$\{v_n\} = \left\{ x_n - \frac{\langle x_n . v_1 \rangle}{\langle v_1, v_1 \rangle} v_1 - ..... - \frac{\langle x_n . v_{n-1} \rangle}{\langle v_{n-1}, v_{n-1} \rangle} v_{n-1} \right\} \tag{4.26}$$

v_n is the orthogonal basis for vector space **V**.

Löwdin Orthogonalization utilizes the orthogonal property of the eigenvectors and eigenvalues of a matrix for orthogonalization. Any general real or complex vector space can be transformed to an orthogonal vector space using a non – singular linear transformation. [19]

4.6.1 Symmetric Orthogonalization

Consider vector space **A** which is a set of linearly independent vectors of n–dimension which is represented as –

$$A = \{\overrightarrow{a_1}, \overrightarrow{a_2}, \overrightarrow{a_3}, ..., \overrightarrow{a_{n-1}}, \overrightarrow{a_n}\} \tag{4.27}$$

Where each vector represents a column matrix of length n such that **A** is a square matrix $n \times n$. The eigenvalues of the matrix **A** are given by matrix $\sum$ and a new matrix **B** is defined as –

$$B := \sum I_{n \times n} \tag{4.28}$$

$$\begin{pmatrix} b_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_n \end{pmatrix}_{n\times n} = \begin{pmatrix} \sigma_1 & \cdots & \sigma_n \end{pmatrix} \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{n\times n} \tag{4.29}$$

Where **B** is the square diagonal matrix with the eigenvalues of **B** along the diagonals. Further, we define the inverse of the square root of this matrix **B** as –

$$d := B^{-\frac{1}{2}} \tag{4.30}$$

$$d = \begin{pmatrix} \sqrt{\frac{1}{b_1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\frac{1}{b_n}} \end{pmatrix}_{n\times n} \tag{4.31}$$

Also, let the eigenvectors of the matrix **B** be given by the matrix **V**, which is also a $n \times n$ matrix Now, **T** can be defined as the transformation on the matrix **A**, such that it acts on the eigenvalues and eigenvectors of **A** as –

$$T := V d V^T \tag{4.32}$$

This **T** is the symmetric orthogonalization matrix of the vector space **A**.

4.6.2 Canonical Orthogonalization

For the same vector space, **A** as in section 4.6.1, with the eigenvectors given by matrix **V**, eigenvalues given by the square matrix **B** and a unitary matrix **U**, the canonical orthogonalization is defined as –

$$\Delta := V U d \tag{4.33}$$

Where, d is the inverse of the square root of the matrix **B** as defined before. This$\Delta$is the canonical orthogonalization matrix of the vector space **A**. In this work, the orthogonalization is attained using the eigen decomposition of the data set. The usage of singular value decomposition is also used to find the eigenvectors and eigen values. This is suitable for rectangular matrices. Here, however, the covariance matrix is considered for evaluating the symmetric as well as the canonical orthogonalization matrices. This avoids the necessity for SVD as the covariance matrix is in square form and eigen decomposition works well. SVD can be used on the covariance matrix also.

Chapter 5

PYTHON CODES

All the computations and the graphical representations are done using Python 3 kernel in Jupyter Notebook. The libraries used are imported as –

```
import pandas as pd

import numpy as np

import scipy

import matplotlib as plt

import seaborn as sns

import sklearn
```

5.1 CSV FILE TO A PANDAS DATA FRAME

```
data = pd.read_csv('Data.csv')

# This converts the `Data.csv' file to a data frame.

# Data frame is a two { dimensional labeled data
```

```
# structure with columns of different data types.
```

5.2 STANDARDIZATION FUNCTION

```
def normalize_to_csv(x,x_norm):
    var = data[x]
    N_data[x_norm] = (var - np.mean(var))/np.std(var)
    N_data.to_csv('file location')
# This function is called for each variable in the `data'
# dataframe to standardize each variable and is store
# in the new csv file `N_data.csv'
```

5.3 EXTRAPOLATION USING SPLINE CURVES

```
# This is a user { defined function which gives an extrapolation
# for a trend  as well as generate a smooth curve of the concerned
# variable across the quarters.
from scipy.imterpolate import InterpolatedUNivariateSpline
from scipy.interpolate import make_interp_spline
def extrapolation(y,yi,Title,order):
    Quarter_list = list(["Q1 - 2016","Q2 - 2016","Q3 - 2016",
    "Q4 - 2016","Q1 - 2017","Q2 - 2017","Q3- 2017","Q4 - 2017",
    "Q1 - 2018","Q2 - 2018","Q3 - 2018","Q4 - 2018","Q1 - 2019",
    "Q2 - 2019","Q3 - 2019","Q4 - 2019","Q1 - 2020","Q2 - 2020",
```

```
    "Q3 - 2020","Q4 - 2020","Q1 - 2021","Q2 - 2021","Q3 - 2021",
    "Q4 - 2021"])
    rnge = np.array(range(0,24,1))
    #plt.style.use("seaborn-whitegrid")
    plt.figure(figsize=(30,10))
    X_Y_Spline = make_interp_spline(rnge,y)
# make_interp_spline() forms a list of values that are the knots
# for the spline curve.
    X_ = np.linspace(rnge.min(), rnge.max(),150)
# linspace(start_value, end_value, n = number_of_points) { this
# is the function's parameter list as used in the above line.
# The function creates an array of numbers starting from the
# start_value and ending at the end_value such that there
# are `n' number of numbers between these two values.
    Y_ = X_Y_Spline(X_)
    fig = plt.figure()
    ax = fig.add_axes([2,2,2,1])
    plt.plot(X_,Y_,'g',label = "Actual Graph")
    xticks = np.array(range(0,24,1))
    ax.set_xticks(xticks)
    ax.set_xticklabels(Quarter_list)
    plt.tick_params(axis = 'x', rotation = 45)
    plt.xlabel("Quarters")
```

```
    plt.title(Title)

    xi = np.array(range(0,16,1))

    x = np.array(range(15,24,1))

    plt.plot(xi, yi,'b',label = "Pre COVID -19 pattern")

    s = InterpolatedUnivariateSpline(xi, yi, k=order)

# This function gives extrapolation of the data points to a

# certain defined range, here, from 15th to 24th x-values,

# the y-values are estimated.

    y2 = s(x)

    plt.plot(x, y2,'--r',label = "Predicted Graph")

    plt.legend()

    plt.grid()

    plt.show()

# The function extrapolation() is called with the parameters

# corresponding to each variable where `y' is the whole set of

# values of the variable and `yi' is the set of initial variables

# which are further extrapolated to visualize the difference

# between the expected trend and the actual trend.

# Also, this function has a title and the order of spline that

# is to be passed as a parameter.
```

5.4 CORRELATION FUNCTIONS

```
# There are two types of correlations that are calculated and
# plotted { Pearson's correlation and Spearman's correlation
# (section 4.3).
```

5.4.1 Pearson's Correlation Coefficient

```
def correlation(x1,x2):
correlation = np.corrcoef(x1,x2)
# correlation() function returns the covariance matrix of the
# two variables # x1 and x2 in the variable correlation.
# The normalized variables Y1 and Y2 can be plotted on the
# same graph with the same timeline given by the variable X
# using the below function
def plotfunc(X,Y1,Y2,LABEL1,LABEL2):
    plt.figure(figsize = (9,7.5))
    plt.plot(X,Y1,label = LABEL1)
    plt.plot(X,Y2,label= LABEL2)
    plt.legend()
# The correlation is plotted using a linear regression
# function from the seaborn library
plt.figure(figsize=(9,7.5))
sns.regplot(x1,x2)
```

```
plt.ylim(0,)
```

5.4.2 Spearman's Correlation Coefficient

```
# Spearman's correlation can be obtained using the scipy
# library
import scipy.stats
S_coeff,S_pvalue = scipy.stats.spearmanr(Y1,Y2)
# spearmanr() returns the coefficient as well as the
# significance value p
```

5.4.3 Heatmap using correlations

```
# We use the heatmap function from the seaborn library
# to plot a color map # of the correlations between
# different variables.
# To do this, first of all we use the dataframe.corr()
# function obtain acorrelations of the dataframe.
# This is a parameter that is passed to the
# heatmap() along with the color and other format styles.
plt.figure(figsize = (10,10))
sns.heatmap(N_data.corr(), cmap = 'Spectral', square = True,
annot = True)
```

5.5 SINGULAR VALUE DECOMPOSITION (SVD)

```
# Singular value decomposition (SVD) can be done very easily
# for a rectangular matrix using the svd() function in the
# linalg module of numpy library.
# Here, however, we use the covariance matrix of the
# given dataset to calculate the SVD as this can be later
# used for Principal Component Analysis (PCA) and L$\ddot{o}$wdin
# orthogonalizations
cov = np.cov(N_data)
svd = np.linalg.svd(cov)
# The function svd() returns an array of three arrays which are
# the three decomposed matrices { the two orthogonal matrices
# and the eigenvalue matrix
eigen_values = svd[1]
eigen_vectors = svd[0]
r_eigen_vectors = svd[2]
# Now, these eigenvalues and eigenvectors are sorted in the
# decreasing order and the two most relevant eigenvalues
# are extracted
Sort = eigen_values.argsort()[::-1]
eigen_vals = eigen_values[Sort][:2]
eigen_vec = eigen_vectors[:,Sort][:,:2]
```

```
r_eigen_vec = r_eigen_vectors[:,Sort][:2]

transform = np.dot(cov.T,eigen_vec)

variance_ratio = eigen_vals/np.sum(eigen_vals)

# The variance_ratio gives the magnitude of the major

# variable's influence on the data set.
```

5.6 PRINCIPAL VALUE DECOMPOSITION

```
# The Principal Component Analysis is done using an

# advanced python library { scikitlearn. We extract

# the two major variables that are created using the PCA.

from sklearn.decomposition import PCA

pca = PCA(n_components = 2).fit(N_data.T)

assert np.allclose(np.abs(pca.components_),

np.abs(eigen_vec.T))

print(pca.explained_variance_ratio)

# The above attribute gives how significant the two

# variables chosen from the PCA are.

# Below lines give a plot of the cumulative

# sum of the significances of all the variables

# from which we can understand how many variables

are to be considered for the analysis.

pca = PCA().fit(N_data)
```

```
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance')
plt.grid()
plt.show()
# The graph of principal Component 1 versus principal
# component 2 is plotted and the samples are the original
# variables in the data set. This graph gives us an idea
# of how the different variables contribute the two
# principle components.
for i, (comp,var) in enumerate(zip(eigen_vec.T,eigen_vals)):
    plt.scatter(N_data.T.iloc[:,0],N_data.T.iloc[:,1],
    alpha = 0.3,label = "Samples")
    comp = comp*var
    plt.plot([0,comp[0]],[0,comp[1]],
    label = f"Principal Component {i+1}",
    linewidth = 3,color = f"C{i+1}")
    plt.gca().set(title="2-D Dataset with Principal Components",
    xlabel="Principal Component 1",ylabel="Principal Component 2")
plt.legend()
plt.show()
```

5.7 SYMMETRIC ORTHOGONALIZATION

```
# The symmetric orthogonalization is applied on the
# covariance matrix of the data set. First, the matrix
# is decomposed by eigen decomposition for this.
lam_s, l_s = np.linalg.eigh(cov)
# Now, we have two arrays { one is the eigenvalue column
# matrix and other is the eigen vector matrix. The eigen
# value matrix is converted to a square array by multiplying
# it with the identity function and then the square
# root of the inverse is calculated (section 4.6.1).
# Further the symmetrical orthonalized matrix is computed.
lam_s = lam_s * np.eye(len(lam_s))
np.linalg.inv(lam_s)
lam_sqrt_inv = np.sqrt(np.linalg.pinv(lam_s))
symm_orthog = np.dot(l_s, np.dot(lam_sqrt_inv, l_s.T))
# A few of the columns of the symmetrical orthogonalized
# matrix are plotted versus each other.
plt.scatter(symm_orthog[0],symm_orthog[1])
plt.xlabel("Parameter 1")
plt.ylabel("Parameter 2")
plt.title("Symmetric Orthogonalization")
```

5.8 CANONICAL ORTHOGONALIZATION

```
# We follow the similar steps as in the previous section

# to obtain the canonical matrix.

e_values,e_vectors = np.linalg.eigh(cov)

e_values = e_values*np.eye(len(e_values))

sqrt_inv = np.sqrt(np.linalg.pinv(e_values))

canonical_ortho = np.dot(cov,np.dot(e_vectors,sqrt_inv))

# The graph of one parameter versus other

# is plotted for this also.

plt.scatter(canonical_ortho[0],canonical_ortho[1])

plt.xlabel("Parameter 1")

plt.ylabel("Parameter 2")

plt.title("Canonical Orthogonalization")
```

Chapter 6

DISCUSSIONS AND INFERENCE

6.1 INTERNET INFRASTRUCTURE FROM 2016

The data set under analysis consists of the different parameters of internet infrastructure, which includes – the different type of connections, data consumption and time usage, the revenue and cost for the service providers and the users respectively. The data from 2016 is considered for the study as it could give an insight of any change in the trends. The different variables studied in this work are –

- Ethernet/LAN: Number of subscribers for the Ethernet or LAN service in each quarter from 2016.
- Fibre: Number of Optical Fiber Cable (OFC) connections or the number of internet users using OFC technology.
- LTE/FW_LTE: Number of LTE or 4G internet users in each quarter.

- Total ISP subscribers: Number of subscribers with a connection for internet from any service provider in the country.
- Wireless Internet: Number of internet users connected via wireless technologies.
- Wired Internet: Number of internet users connected via any wired technologies.
- Broadband: Number of internet users connected via the broadband spectrum.
- Monthly ARPU: The monthly Average Revenue Per User collected by a service provider
- Minutes of Internet Usage per month: This indicates the number of minutes spend on internet by a single user averaged for three months of a quarter.
- Internet Telephony in minutes (in millions): This gives the total number of minutes spent in internet telephoning by all the users averaged to three months of a quarter.
- Wireless Data usage (Gb) per month: This gives the data consumption per user with the highest among the three months of a quarter is considered.
- Cost to subscriber per Gb: The amount spent by each user in availing one Gb of data in each quarter.

Initially, all the parameters were plotted against the time axis to visualize the trend. Along with that an extrapolation of the same from Q1 – 2020 to Q4 – 2021 is done to study the change in the trend due to COVID – 19 emergencies.

6.1.1 Extrapolation of different parameters

The extrapolations for linear trends are only considered for better results. The plots are as shown below.

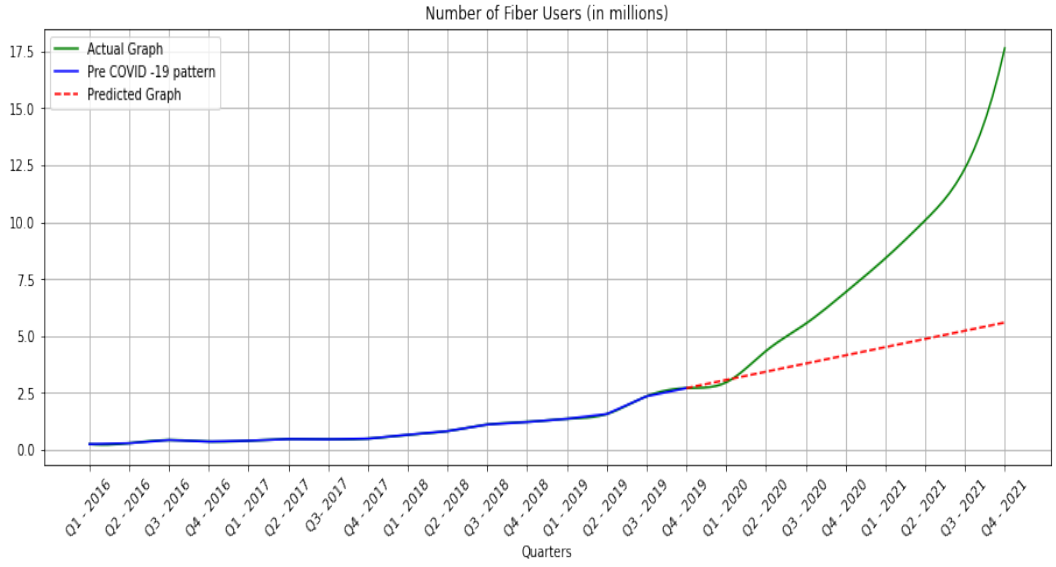

Figure 6.1: Number of Fiber Users

Number of Fiber users (Fig 6.1): This graph is an exceptional one because of the change in its trend during the COVID – 19. The OFC technology was in its nurturing stage before the pandemic and it provided a perfect window for the exponential increase in the number of subscriptions. The Government of India's BBNL or Bharat-Net project was on its way and it was boosted by the necessity. The establishments were sped and it was not at all a task to get enough

subscribers to manage it. Along with this, other private service providers also invested in the OFC technology thereby, becoming the major source for internet connection by the end of Q4 – 2021. Also, it is quite surprising to observe that the expected trend (Q1 – 2020 to Q4 – 2021), which is an extrapolation of the trend until Q4 – 2019, is highly deviated from the actual graph. This is a good indicator that the pandemic has affected the internet infrastructure in a positive manner. It sped the pace of the growth of the most reliable and advanced internet technology of the current time, i.e., OFC.

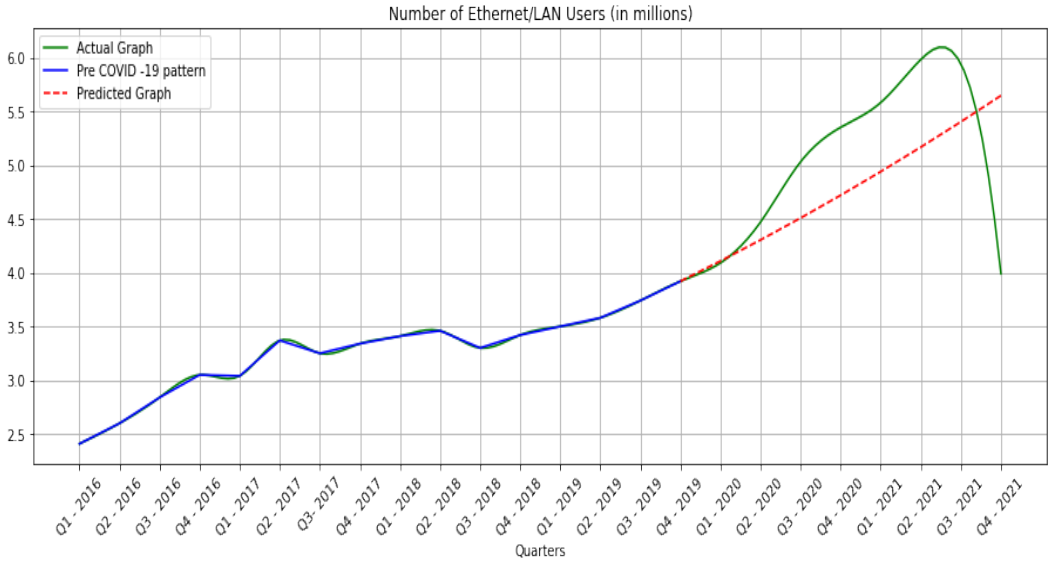

Figure 6.2: Number of Ethernet/LAN users

Number of Ethernet/LAN users (Fig 6.2): Although the graph is not perfectly linear, we can consider the Q1 – 2016 to Q4 – 2019 to be almost linear. However, the trend observed later on, during the COVID – 19 crisis is completely deviating from the usual trend. There is a significant hike in the number of subscribers during the lockdown from Q2 – 2020 to Q2 – 2021, however, afterwards the

number of subscribers tend to decrease exponentially. This may be due to the fact that the work-from-home and online classes declined and the conventional methods were restored eventually. Also, the fact that better technologies like OFC and LTE would have replaced the Ethernet and LAN connections for better service.

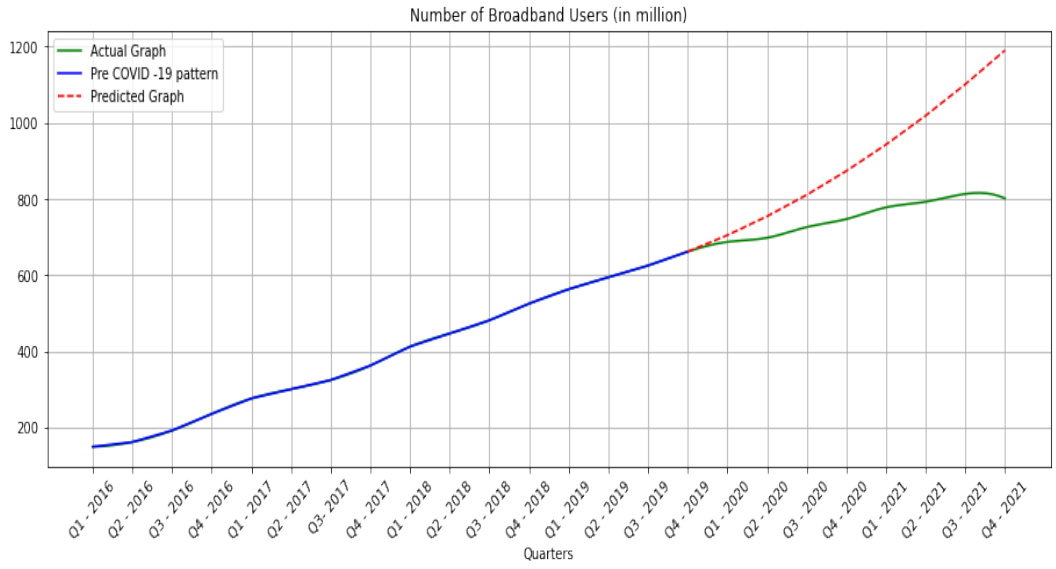

Figure 6.3: Number of Broadband Users

Number of broadband users(Fig 6.3): We can observe from the above graph that the number of broadband subscriptions, although increased, did not increase as per the expectation. The actual trend seems to be flattening from the Q2 – 2021. Thus, we get an unexpected saturation, when the demand of internet was highest due to the lockdown.

Number of ISP subscribers (Fig.6.4): In this graph we can observe a saturation in the quarters of 2021. Although the expected pattern showed an exponential growth, the actual trend does not. Here, we have to consider the total number of

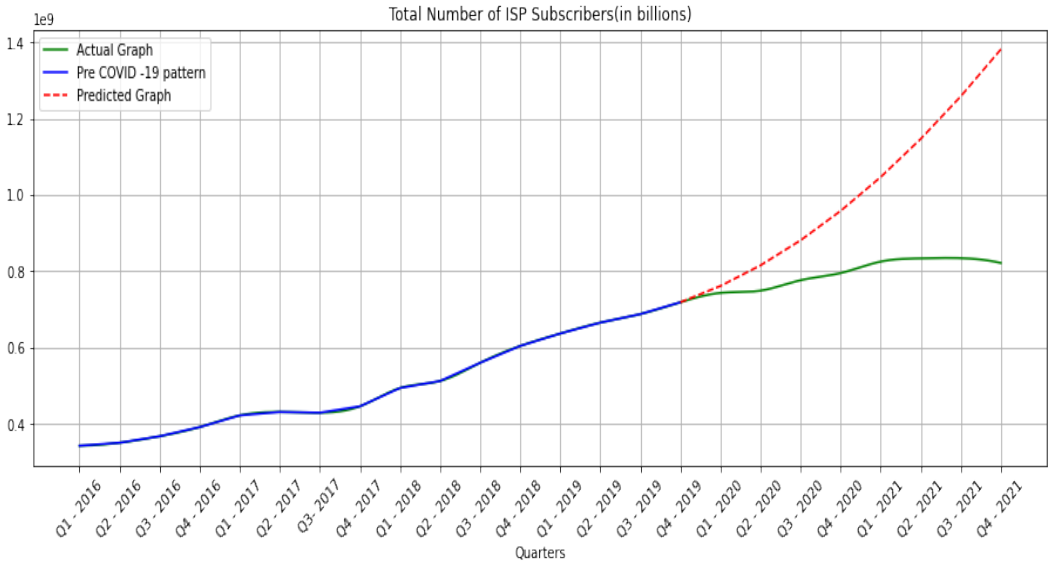

Figure 6.4: Number of ISP subscribers

people in India is almost 1.4 billion as of 2021 and the number of subscribers to any of the internet service providers is almost 822 million. Given that the population consists of all age groups, the saturation of the trend is not unexpected.

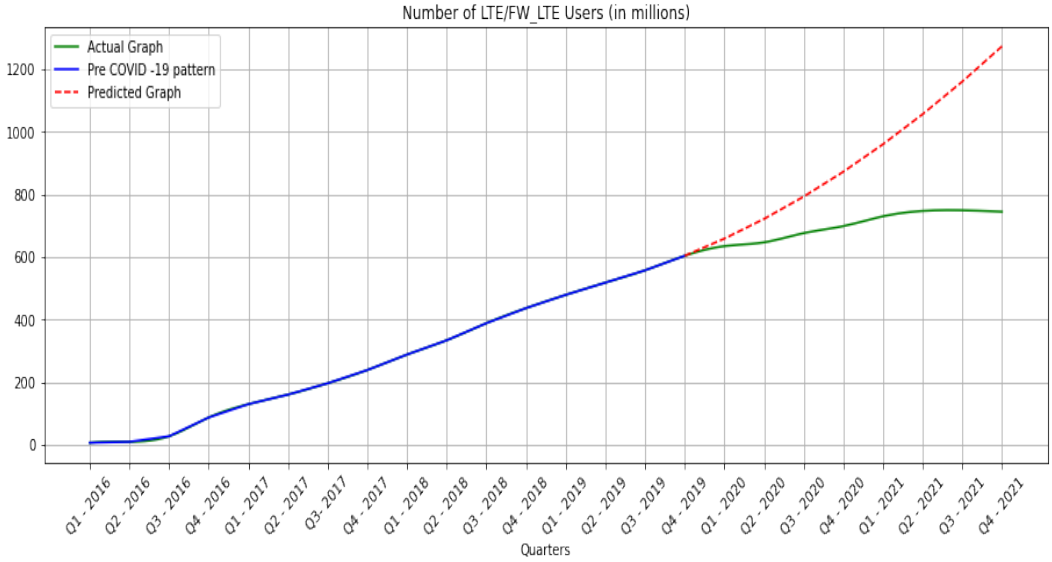

Figure 6.5: Number of LTE Users

Number of LTE/FW_LTE users (Fig. 6.5): This parameter also shows a deviation from the usual trend. The number of LTE users are saturated at a point in time. The maximum number of LTE users recorded was in Q3 - 2021, i.e., 748.63 million which is almost half the population.

In the three figures Fig.6.3,Fig.6.4 and Fig.6.5, we can observe that the number of subscribers have come to a saturation around 800 million. If we consider the population of India, which is around 1.4 billion, we can safely assume that this denotes that the adult population capable (almost 900 million) of using the technology have access to these technologies. i.e., there would be no further exponential increase that can be expected from these parameters.

6.1.2 Correlations between parameters

The correlation between all the pairs of the variables were calculated. A heat map of the covariance correlation is given in Fig.6.6. We can see that the diagonal elements are all one due to the fact that variance to itself is one. Other than that, we can observe that most of the variables have a high positive correlation with each other.

On closer observation, we can infer that the cost of the internet has a negative correlation with the other parameters. This denotes that despite the increase in demand, the internet tends to be less costlier than it used to be. This is a great advantage that the Indian internet users have. The price of internet facilities have come down such that it is affordable for the majority of the households in the country.

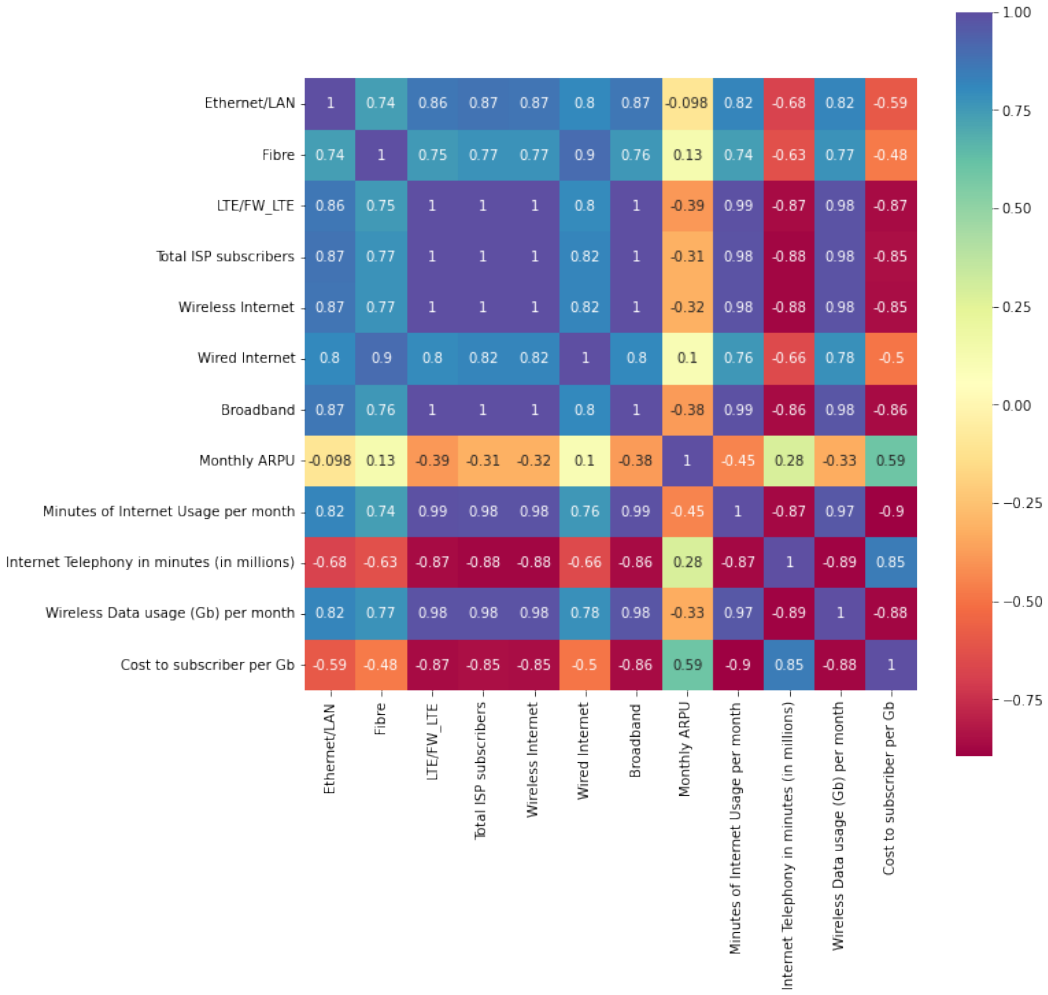

Figure 6.6: Correlation Heat map

Moreover, we can see that the internet telephony which used to be prominent telecommunication method especially for international calls have dropped or is highly negatively correlated to the other improvement parameters. This can be due to the better voice calling and video calling applications which are cheaper and more reliable replacing them.

Further, the Pearson and Spearman Correlation coefficients are calculated for each pair of variables and the most relevant ones with high correlation and certainty is given in Fig.6.7.

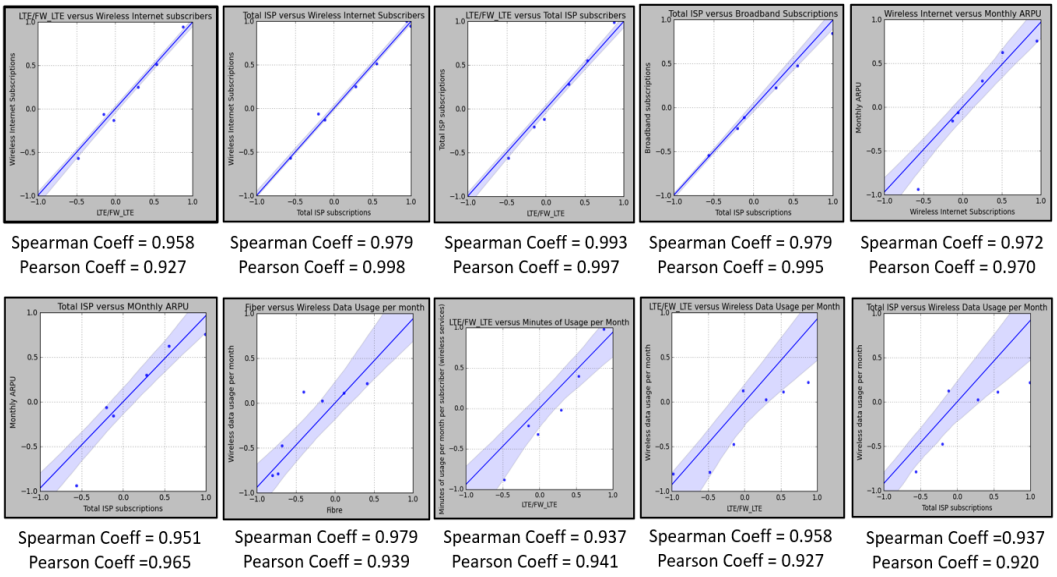

Figure 6.7: Significant Correlation Coefficients of some variables

We can observe that the Pearson coefficients and Spearman coefficients are not equal for any of the correlations. This is due to the fact that the variables exhibit different amount of linear and non-linear relationships; Pearson could only capture the linear effect and the non-linear effects are not regarded. Correlations with higher Pearson coefficient denote that those variables have a higher linear association and those with a higher Spearman coefficient denote a trace of non-linearity in their association.

6.1.3 Singular Value Decomposition (SVD)

The data consists of 12 variables with data from 24 quarters. i.e., it is a (24×12) matrix. The raw data set has been normalized by the mean and standard deviation of each variable so that all the variables have equal weightage when computing the singular values and the unitary matrices. The normalization is

done using the formula -

$$x_{norm} = \frac{x - x_{mean}}{\sigma_x} \tag{6.1}$$

Once the new data set of the normalized values have been created, the singular value decomposition is done on the data set. From section 4.4, we know that SVD gives three matrices among which Σ is the singular matrix which gives the magnitude of the relevance of the variables in the data set.

Moreover, before computing SVD on this non-square matrix, the covariance matrix is computed so that this same results could be extended to the computation of the Principal Components of the data set. As a result, we know have a (24×24) square matrix on which SVD is done.

The Singular matrix Σ obtained for the data set after multiplying with an identity matrix of dimensions (24×24) is -

$$\begin{pmatrix} 15.4708077 & 0 & \cdots & 0 & 0 \\ 0 & 1.95582436 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 2.63145728e-17 & 0 \\ 0 & 0 & 0 & 0 & 8.65514969e-19 \end{pmatrix}$$

The most prominent variables are those corresponding first two singular values that are contributing to the trend in the data set than the others. The percentage of relevance of these variables is computed used a simple formula given by -

$$Ratio = \frac{singular\_values}{\sum singular\_values} \tag{6.2}$$

We find the ratio between each singular value and the sum of all the singular values to get the percentage of relevance of each variable. For the first two variables, the ratio of relevance is thus given by -

$$\begin{pmatrix} 0.88776808 & 0.11223192 \end{pmatrix}$$

Thus, we can say that the first variable has a relevance of about 88.77% and the second has a relevance of about 11.22%, which together sum up to 99.99%. Therefore, we can say that the singular value decomposition successfully extracts the most relevant features of the data set so that it could be further used for dimensionality reduction.

6.1.4 Principal Component Analysis (PCA)

As obtained by the Singular Value Decomposition, we now have the two variables which contribute to the data set. These variables are a hybrid of the existing variables defined by the most prominent features of the data set.

Apart from SVD, we can obtain this using the normal eigen value decomposition also. We have already computed the SVD on a square matrix, i.e., the covariance matrix. Now, eigen value decomposition can also be done on this covariance matrix which will give the eigen values instead of the singular values. In this work, the principal components are obtained using the eigen value decomposition. The two relevant variables had the relevance ratio given by -

$$\begin{pmatrix} 0.81307025 & 0.1027886 \end{pmatrix}$$

Now, we have obtained the eigen values and the corresponding eigen vectors for the two principal components of the data set, it is used to plot the original variables of the data set (Fig.6.8).

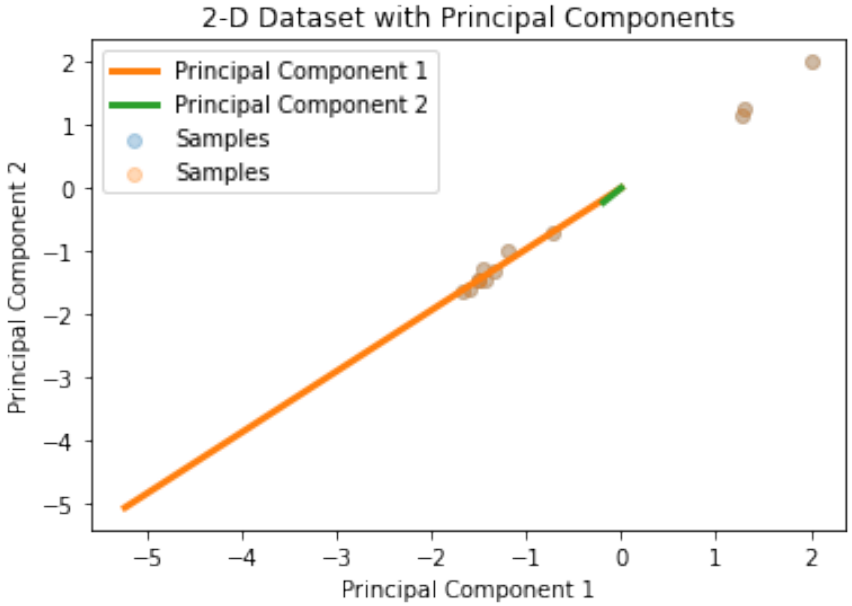

Figure 6.8: Principal Components of the Data set

We can observe that most of the variables in the data set align to the first principal component. This shows how powerful the newly defined variable is in defining the whole data.

Now, we can compare how much effect does each variable has on the data set using a cumulative ratio plot as given in Fig.6.9.

We can observe that the first two principal components itself give more 95% of the variance in the data set.

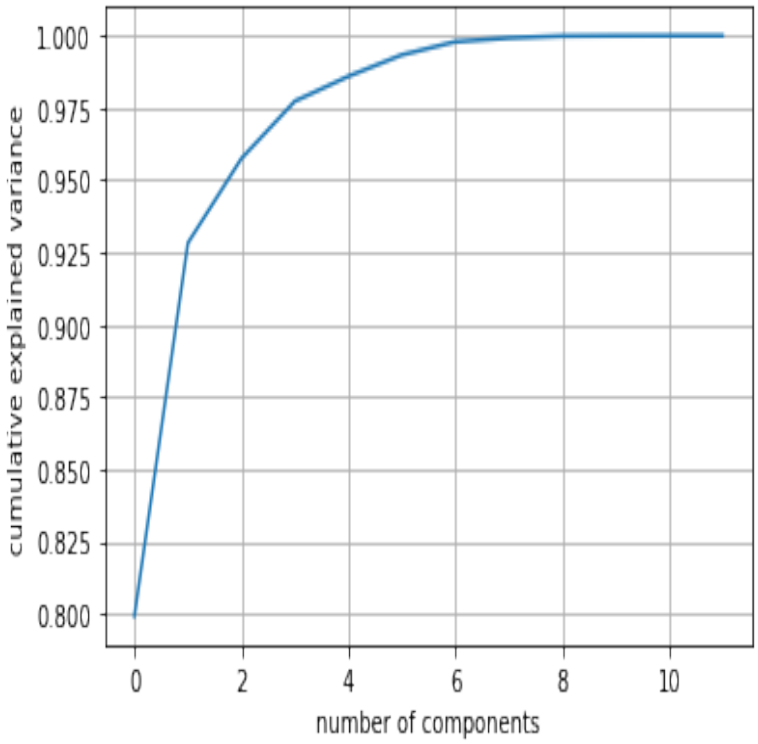

Figure 6.9: Cumulative relevance of Principal components

6.1.5 Symmetric and Canonical Orthogonalizations

The Symmetrical and canonical orthogonalizations of the data set is done to infer any lopsided variables that are ought to be present in the data set.Upon orthogonalization of the (24×12) data set, we obtain a (24×24) matrix.

The first two columns of both the orthogonalized matrices are plotted against each other as in figures - Fig.6.10, Fig.6.11, Fig.6.12.

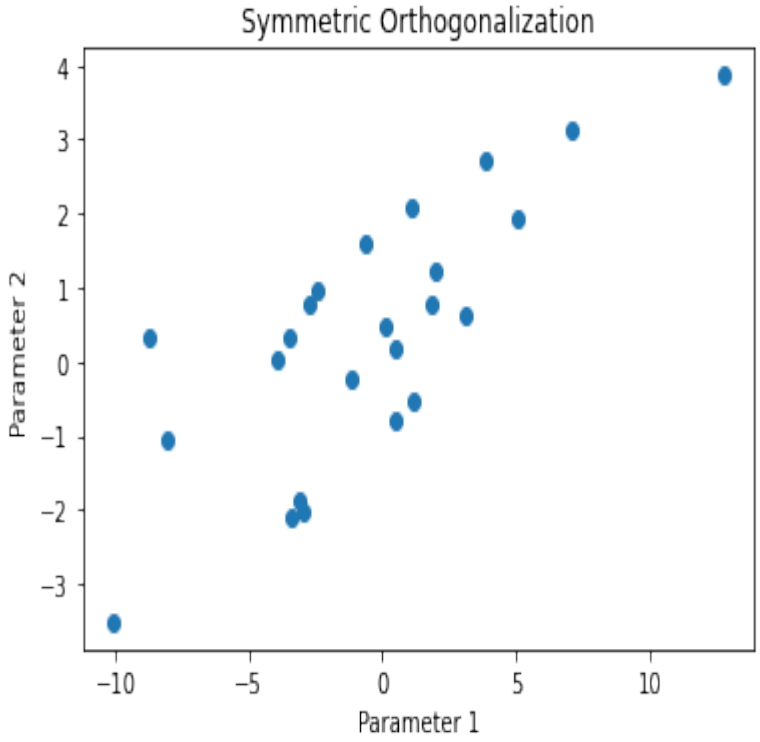

Figure 6.10: Parameter1 vs Parameter2 of Symmetric Orthogonalization matrix

6.2 COVID-19, INTERNET AND PEOPLE

The change in lifestyle that accompanied the lockdown was completely dependent on the internet. The hours of usage increased dramatically, given that everything was done online. The below graph gives an account of how the people who participated in the survey were using internet during the lockdown.

The leisure time excludes the internet usage that is in anyway related to the work or classes. The total internet usage time includes the leisure time also.

We can observe that the trend of usage (Fig.6.13) is the same for both the variables, however, the number of moderate users is higher for total usage as compared to leisure usage. This is balanced by having a larger number of extreme

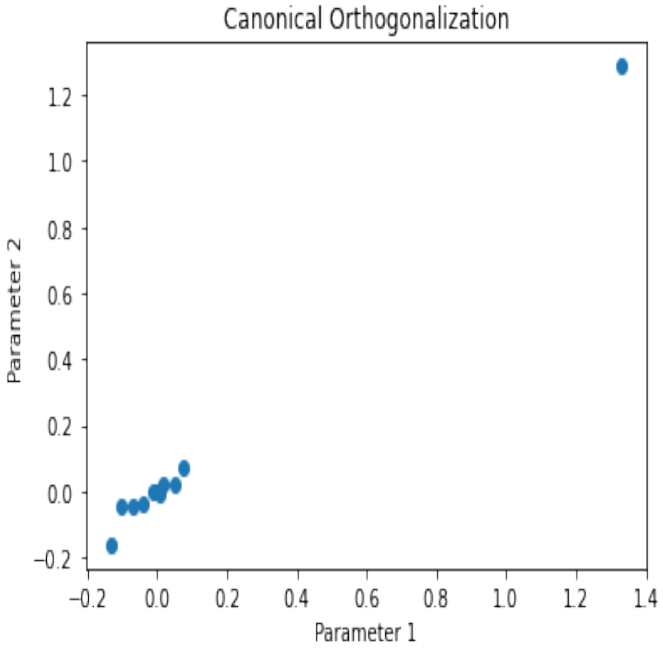

Figure 6.11: Parameter1 vs Parameter2 of Canonical Orthogonalization matrix

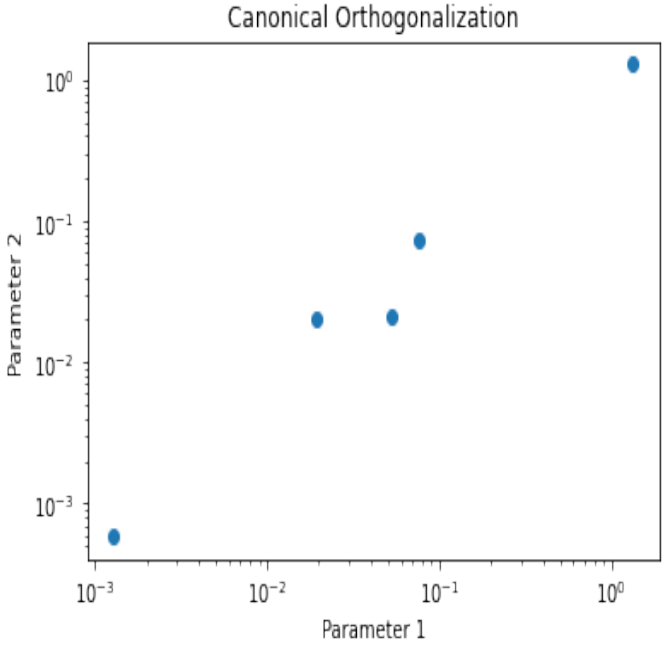

Figure 6.12: Parameter1 vs Parameter2 of Canonical Orthogonalization matrix (On log scale)

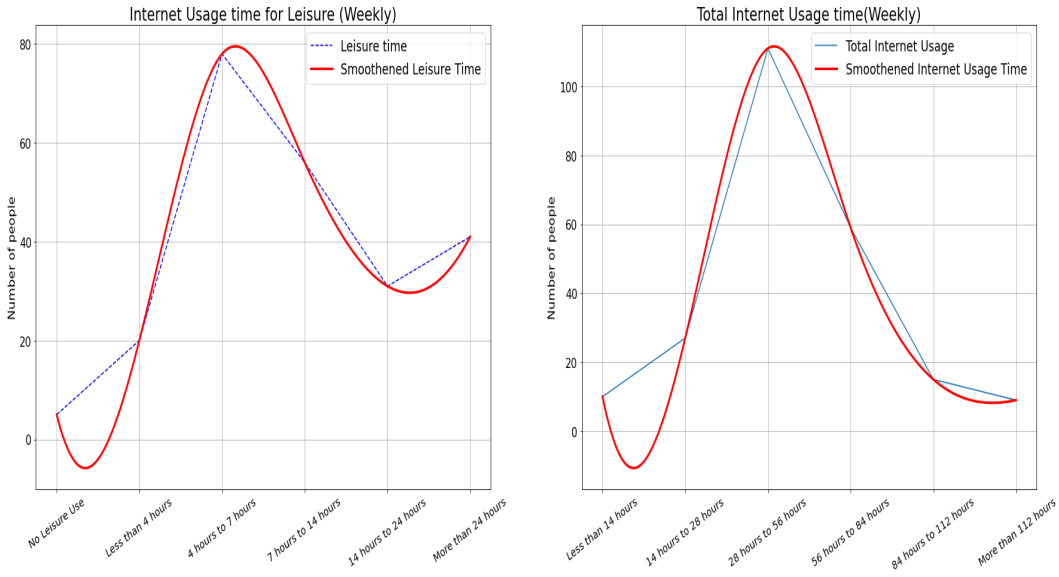

Figure 6.13: Internet Usage trend: Total usage time and Leisure usage time

users in the leisure usage. We can infer that more people used a good time on internet for leisure activities.

Apart from the usage trend, one of the striking observation from the survey was the correlation of the quality of workplace and the quality of work of the subject.

We can infer from Fig.6.14 that the number of people who reported to have a particular standard of work and the number of people who claimed that their workplaces had a corresponding quality are almost equal for all the five categories. This is a clear indicator that the quality of service provided by the workplace to the employees affected the work quality of the employees which is the same in physical mode of work also.

As an extension to this, the correlation of the two parameters has been plotted

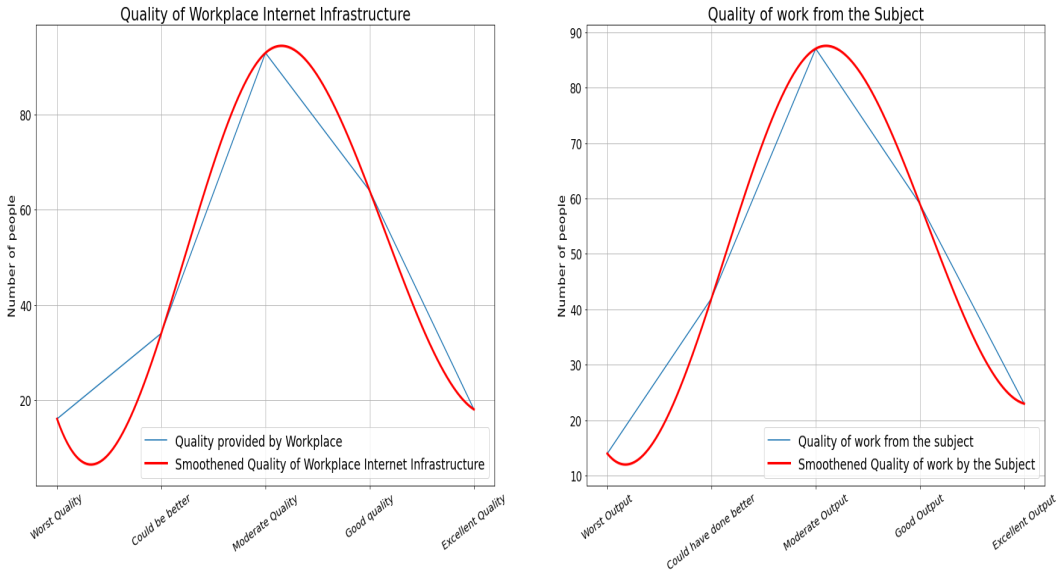

Figure 6.14: Quality of workplace and Quality of work

(Fig 6.15).

Figure 6.15: Correlation between Quality of work and workplace

It is clear that our approximation on the correlation between these two were correct. The Correlation coefficients are given by -

Pearson coefficient = 0.987

Spearman Coefficient = 0.999

6.2.1 Mobile Data Download Speed and Mobile Data Consumption

An interesting analysis was published by OpenSignal [20] in their website which showed how the mobile data speed and the rate of data consumption varied (Fig.6.16).

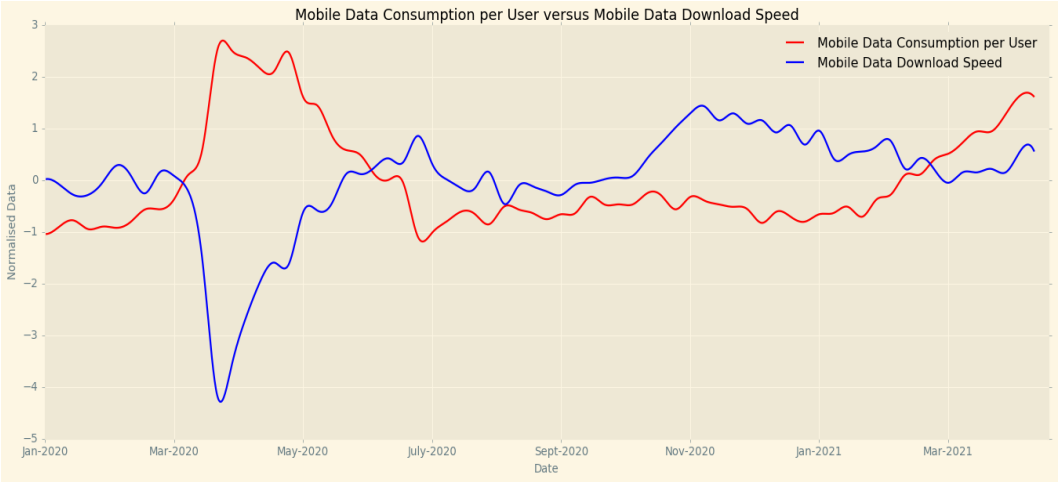

Figure 6.16: Mobile Data download speed versus Mobile Data consumption

The graph has three different regions which can be defined as the first wave phase, first relief phase and the second wave phase. During the first wave or the first lockdown (April 2020), there was an unexpected demand for the mobile data, i.e., the data consumption rate hiked. However, then the ISPs were not equipped to satisfy such a drastic demand and as a result the mobile data download speed had a nose-dive.

However, when we move to the Second wave (Mar 2021), we can see that there was a similar increase in the internet demand, but no noticeable change was observed for the data speed.

This difference in pattern within a year denotes that there was a significant improvement in the internet infrastructure in our country.

Chapter 7

CONCLUSION

All the analysis and study in this work has given a positive effect on the internet infrastructure during COVID-19. The then existing technologies like LTE, broadband and other wireless technologies had an increase initially which ended in a saturation over time. The comparatively newer technologies like OFC had an exponential growth due to the demand which is an indicator that the lockdown aided internet infrastructural development. Besides, we have inferred that the older and less efficient technologies like Ethernet was outdated soon to give space for the latest ones, which in turn has improved the digital standards. Moreover, the 5G implementation gathered momentum much faster than it would have.

Bibliography

[1] https://en.wikipedia.org/wiki/Internet_in_India *Wikipedia contributors. "Internet in India." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 10 Jun. 2022. Web. 25 Jun. 2022*

[2] https://www.trai.gov.in/release-publication/reports/telecom-subscriptions-reports

[3] https://ourworldindata.org/covid-cases?country= IND#country-by-country-data-on-confirmed-cases

[4] https://economictimes.indiatimes.com/tech/internet/the-internet-turns-25-in-india-a-timeline/the-2010s/slideshow/77589523.cms

[5] https://www.livemint.com/Opinion/gzWbpGZVD83W3iq3uOLD7O/Evolving-Internet-in-India.html

[6] https://www.digitalindia.gov.in/

[7] https://bbnl.nic.in/

[8] https://us.hitrontech.com/learn/what-is-gpon-technology/

[9] https://indiainvestmentgrid.gov.in/national-infrastructure-pipeline

[10] https://www.qualcomm.com/5g/what-is-5g

[11] https://www.downtoearth.org.in/coverage/science-and-technology/all-about-mobile-spectrum-33106

[12] https://people.cs.clemson.edu/ dhouse/courses/405/notes/splines.pdf

[13] https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[14] https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[15] https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-science-388a54be95d

[16] https://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm

[17] https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[18] https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

[19] Naidu, Annavarapu Ramesh. "Centrality of Lowdin orthogonalizations." arXiv preprint arXiv:1105.3571 (2011).

[20] https://www.opensignal.com/2021/06/10/indian-mobile-experience-has-been-remarkably-resilient-during-the-second-covid-19-wave-despite